Empirical Analysis of Dialogue Relation Extraction with Large Language Models

Guozheng Li[1] , **Zijie Xu**[1] , **Ziyu Shang**[1] , **Jiajun Liu**[1] , **Ke Ji**[1] and **Yikai Guo**[2]
[1]School of Computer Science and Engineering, Southeast University
[2]Beijing Institute of Computer Technology and Application
{gzli, zijiexu, ziyus1999, jiajliu, keji}@seu.edu.cn

Abstract

Dialogue relation extraction (DRE) aims to extract relations between two arguments within a dialogue, which is more challenging than standard RE due to the higher person pronoun frequency and lower information density in dialogues. However, existing DRE methods still suffer from two serious issues: (1) hard to capture long and sparse multi-turn information, and (2) struggle to extract golden relations based on partial dialogues, which motivates us to discover more effective methods that can alleviate the above issues. We notice that the rise of large language models (LLMs) has sparked considerable interest in evaluating their performance across diverse tasks. To this end, we initially investigate the capabilities of different LLMs in DRE, considering both proprietary models and open-source models. Interestingly, we discover that LLMs significantly alleviate two issues in existing DRE methods. Generally, we have following findings: (1) scaling up model size substantially boosts the overall DRE performance and achieves exceptional results, tackling the difficulty of capturing long and sparse multi-turn information; (2) LLMs encounter with much smaller performance drop from entire dialogue setting to partial dialogue setting compared to existing methods; (3) LLMs deliver competitive or superior performances under both full-shot and few-shot settings compared to current state-of-the-art; (4) LLMs show modest performances on inverse relations but much stronger improvements on general relations, and they can handle dialogues of various lengths especially for longer sequences.

1 Introduction

Relation extraction (RE) aims to identify relationships between arguments in text, including sentences [Zhang *et al.*, 2017], documents [Yao *et al.*, 2019], or dialogues [Yu *et al.*, 2020]. However, as many relational facts span multiple sentences, sentence-level RE faces limitations in practice [Yao *et al.*, 2019]. Therefore, cross-sentence RE, which aims to identify relations between two entities not in the same sentence, is vital for building knowledge bases from extensive corpora.

S1: Hey, you guys!
S2: Hey!
S1: Guess what?
S3: What?
S1: **I got a gig!**
S4: Yay!!
S5: **See, that's why I could never be an actor. Because I can't say gig.**
S2: Yeah, I can't say croissant. Oh my God!
S6: What's the part?
S1: Well, it's not a part, no. **I'm teaching acting for soap operas down at the Learning Extension.**
S3: Come on! That's great.
S4: Wow!
S1: Yeah, yeah. It's like my chance to give something back to the acting community.

Argument pair	**Relation type**
(S1, actor)	*per:title*
(Learning Extension, S1)	*org:employees or members*
(S1, Learning Extension)	*per:employee or member of*

Figure 1: An example dialogue with its desired relations. S1-S6: anonymized speaker of each utterance.

As dialogues often exhibit cross-sentence relations, extracting relations from dialogues is necessary [Yu *et al.*, 2020].

To extract relations between two arguments that appear within a dialogue, dialogue relation extraction (DRE) models require reasoning over multiple sentences in a dialogue. To this end, a plethora of models have been proposed to address the challenges of DRE in recent studies including *sequence-based* methods [Yu *et al.*, 2020; Son *et al.*, 2022] and *graph-based* methods [Xue *et al.*, 2021; Lee and Choi, 2021; Liu *et al.*, 2023]. Sequenced-based methods utilize encoder-only pre-trained language models such as BERT [Devlin *et al.*, 2019] to encode utterances in the dialogue. Graph-based methods usually construct various heterogeneous graphs over dialogues to capture the relational information. However, neither paradigm of methods can deliver satisfactory DRE results, because these encoder-based methods are sensitive to the higher person pronoun frequency and lower information density in dialogues [Lee and Choi, 2021]. As shown in Figure 1, the sentence highlighted with red is a prerequisite for correct understanding of relationships in the following sen-

tences highlighted with green, but this sparse multi-turn information is difficult for models to capture. Moreover, current methods suffer grave performance declines when estimating how many turns it require to predict the relations between two arguments in partial dialogue instead of entire dialogue [Yu *et al.*, 2020]. For example, DRE models are expected to deduce the relation *per:title* between *S1* and *actor* with only first seven utterances in Figure 1 while current methods typically need to rely on entire dialogue. This motivates us to search for more effective methods that can alleviate the above issues.

Besides the above sequence-based and graph-based methods, *generation-based* methods are also occasionally used in RE, but only focus on sentence-level [Cabot and Navigli, 2021; Paolini *et al.*, 2021] and document-level [Giorgi *et al.*, 2022] RE. As the emergence of large language models (LLMs) such as GPT-3 [Brown *et al.*, 2020] significantly driving the progress of natural language processing (NLP), utilizing generation-based methods start to draw more attention [Wadhwa *et al.*, 2023; Zhang *et al.*, 2023; Ma *et al.*, 2023; Li *et al.*, 2023; Wan *et al.*, 2023] in RE community. To the best of our knowledge, however, there is currently no work specifically applying generation-based methods to DRE. To this end, we aim to investigate the capabilities of generation-based methods for DRE in this work. Specifically, we consider the evaluation of both powerful proprietary LLMs such as ChatGPT [OpenAI, 2022] and open-source LLMs such as LLaMA [Touvron *et al.*, 2023]. For proprietary LLMs, we design various prompt formats to extract relations from dialogues via few-shot prompting [Brown *et al.*, 2020] and zero-shot prompting [Kojima *et al.*, 2022]. For open-source models evaluation, we introduce **Landre** (**Lan**guage models driven **d**ialogue **r**elation **e**xtraction), a DRE framework driven by LLMs based on open-source foundation models, where it employs a simple prompt tuning method and a parameter efficient tuning technique LoRA [Hu *et al.*, 2022] for better RE task adaptation. We compare the performances of generation-based methods with previous sequence-based and graph-based methods, and conduct extensive experiments to provide valuable insights and guide further exploration.

Importantly, we empirically discover that generation-based methods significantly alleviate the issues of DRE mentioned above. On the one hand, for the difficulty of capturing long and sparse multi-turn information, we find that scaling up model size substantially boosts the overall DRE performance and achieves exceptional results. While the ChatGPT demonstrates promising performance with elaborated prompts, fine-tuned open-source models surpass prior state-of-the-art (SoTA) by a large margin, opening new frontiers for advancing the research of DRE. On the other hand, for the requirement of partial dialogue understanding, we surprisingly find that all generation-based methods encounter with much smaller performance drop from entire dialogue to partial dialogue setting, comparing to sequence-based and graph-based methods. In summary, our contributions are as follows:

- We present the initial evaluation of LLMs regarding to DRE. We design elaborated prompts for ChatGPT and introduce an LLM-driven DRE framework Landre for open-source foundation models. Evaluation on DRE benchmarks across various experimental settings reveals our significant performance improvements over previous sequence-based and graph-based methods.
- We propose to utilize generation-based methods in DRE, which significantly alleviate two main issues of DRE. Generation-based methods are able to capture important relation semantics from sparse multi-turn information and achieve relatively consistent performances under both partial and entire dialogue settings.
- By analyzing the performance and complexity of different generation-based methods, we emphasize their promising performance and obvious shortcomings compared to previous methods, providing valuable insights to advance future DRE research.

2 Dialogue Relation Extraction (DRE)

Given a dialogue $D = s_1 : t_1, s_2 : t_2, ..., s_m : t_m$ and an argument pair (a_1, a_2), where s_i and t_i denote the speaker ID and text of the i^{th} turn, respectively, and m is the total number of turns, we evaluate the performance of approaches in extracting relations between a_1 and a_2 that appear in D in the following two settings. (1) **Standard Setting:** Similar to standard RE tasks, the dialogue D is regarded as a document d, and we evaluate the extracted relations between a_1 and a_2 based on d. Specifically, we adopt F1, which is the harmonic mean of precision (P) and recall (R), for evaluation. (2) **Conversational Setting:** Instead of only considering the entire dialogue, we can regard the first $i \leq m$ turns of the dialogue as d. Given d containing the first i turns in a dialogue, relation type r associated with a_1 and a_2 is evaluable if a_1, a_2, and the trigger for r have all been mentioned in d. The definition is based on the assumption that we can roughly estimate how many turns we require to predict the relations between two arguments based on the positions of the arguments and triggers. Accordingly, we adopt a new metric $F1_c$, the harmonic mean of conversational precision (P_c) and recall (R_c), as a supplement to the standard F1 metric [Yu *et al.*, 2020].

3 Large Language Models for DRE

We leverage various LLMs to address DRE task. We consider both proprietary LLMs such as ChatGPT [OpenAI, 2022] and open-source LLMs such as LLaMA [Touvron *et al.*, 2023].

3.1 Leveraging ChatGPT for DRE

We access ChatGPT using official API service [1] by prompting. As prompting is a brittle process wherein small modifications to the prompt can cause large variations in the model predictions, we explore various prompt templates to investigate the potential of ChatGPT for DRE in two distinct extraction manner: direct extraction and indirect extraction.

Direct Extraction Direct extraction makes ChatGPT directly output the relation labels between an given argument pair based on the input dialogue and candidate relations. This vanilla prompting method is simple, direct, and efficient.

[1] https://platform.openai.com/

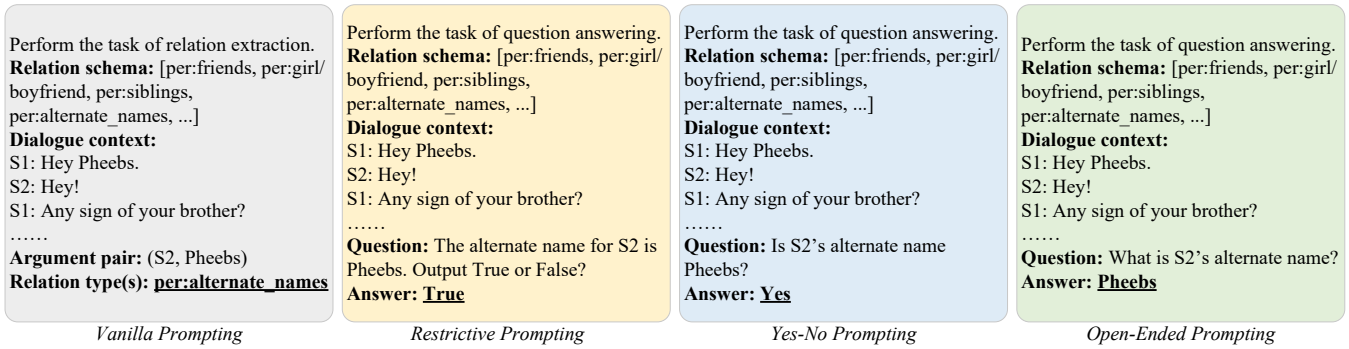

Figure 2: Formats of four prompting. The outputs of LLMs are highlighted with underline.

Indirect Extraction Indirect extraction decomposes the DRE task into two steps of asking then answering. Specifically, we sequentially ask ChatGPT if an argument pair expresses a certain relation among candidate relations based on the dialogue. We ground our analysis in three standard categories of prompts used in prior work [Brown *et al.*, 2020; Sanh *et al.*, 2022; Liu *et al.*, 2024; Arora *et al.*, 2023; Li *et al.*, 2024; Shang *et al.*, 2024] including restrictive prompts, yes-no prompts and open-ended prompts. Restrictive prompts (e.g., "S1 is an actor. Output True or False?") construct questions that restrict the model output particular tokens, while yes-no prompts (e.g., "Is S1 an actor?") and open-ended prompts (e.g., "What is the title of S1?") construct free-form questions, as shown in Figure 2. Typically, the indirect prompting method suffers from relatively high inference complexity as it needs to enumerate all possible triples to obtain questions and answers. Suppose we have n samples to be extracted and r candidate relations, the total complexity is $O(r \times n)$, which is time-consuming in practice. The purpose of experimenting with indirect extraction is to explore the DRE potential of LLMs as one step extraction regularly leads to suboptimal results [Li *et al.*, 2023].

3.2 Tuning Open-Source Models for DRE

Different with the aforementioned usage of ChatGPT, we fine-tune the open-source LLMs for better task adaptation. Specifically, we introduce Landre, a DRE framework driven by LLMs based on smaller, open-source foundation models such as LLaMA [Touvron *et al.*, 2023], where we investigate the performance of fine-tuning generative LLMs for DRE. We first outline the process of constructing input and output prompt formats. Then we utilize a parameter-efficient fine-tuning (PEFT) technique to train the foundation model with limited computational resources.

3.3 Prompt Tuning

As shown in Figure 3, for each sample (D, a_1, a_2) in the training set, we construct the corresponding input prompt and let the LLM directly generate the corresponding relation label(s) indicating the relationship(s) between arguments a_1 and a_2 expressing by the dialogue D. Specifically, we concatenate the given dialogue and argument pair separated by the recognizable delimiter "|" as the input prompt, and we also separate the relation labels between the given argument

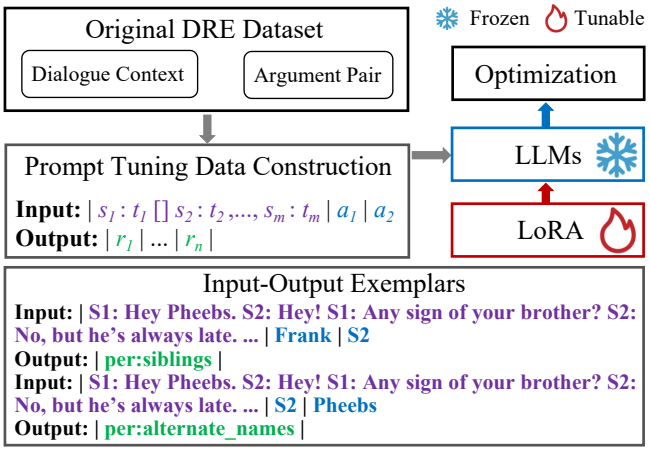

Figure 3: Paradigm of the Landre framework. In the first step, we construct the prompt tuning data from the original DRE dataset. Then we utilize the parameter-efficient fine-tuning technique LoRA to train the foundation model. The purple, blue and green texts in the above exemplars refer to dialogue context D, argument pair (a_1, a_2) and relation set R, respectively.

pair by the "|" symbol as the output prediction. The detailed prompt tuning settings are introduced as follows.

The input prompt is composed of three parts: (1) the dialogue context $D = s_1 : t_1, s_2 : t_2, ..., s_m : t_m$, (2) the argument a_1 and (3) the argument a_2. And the output is a relation set $R = \{r_1, ..., r_n\}$ where n denotes the number of relations between argument pair (a_1, a_2):

$$| s_1 : t_1 [] s_2 : t_2 \ldots s_m : t_m | a_1 | a_2 \rightarrow | r_1 | \ldots | r_n | \quad (1)$$

Note that in most cases, $n = 1$. For convenience, we concatenate all the speaker IDs and texts of multi-turn with the white space token [], to indicate the start of each utterance. We empirically find that although this prompting strategy is frustratingly simple but shows its effectiveness.

3.4 Parameter Efficient Tuning

In this part, we describe how to fine-tune the foundation model using a parameter efficient approach. Landre takes the dialogue context and argument pair from the dataset as inputs and retrieves the corresponding relation label(s) as output:

$$R = \text{Decoder}(D, a_1, a_2) \quad (2)$$

DialogRE	Train	Dev	Test
# Conversations	1,073	358	357
Average dialogue length	229.5	224.1	214.2
# Argument pairs	5,963	1,928	1,858
Average # of turns	13.1	13.1	12.4
Average # of speakers	3.3	3.2	3.3

Table 1: Statistics of DialogRE.

where Decoder indicates that the foundation model (e.g., LLaMA [Touvron *et al.*, 2023]) uses the Transformer-decoder framework. To enhance the efficiency of the fine-tuning process and reduce memory requirements, we utilize Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022], which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. Finally, the learning objective of the generation process is to minimize the negative loglikelihood of R given the dialogue context D and argument pair (a_1, a_2):

$$L = -\sum_{i}^{N}\sum_{j}^{|D_i|} \log p(R^j \mid D_i, a_1^j, a_2^j) \tag{3}$$

where N and $|D_i|$ denote the total number of dialogues in training set and argument pairs in dialogue D_i.

4 Experiments

4.1 Experimental Settings

Datasets and Metrics We conduct experiments on the two versions (V1 and V2 [2]) of DialogRE dataset [Yu *et al.*, 2020], the first human-annotated DRE dataset, originating from the complete transcripts of the series *Friends*. V1 is the original English version and V2 is the updated English version with a few annotation errors fixed. DialogRE has 36 relation types, 1,788 dialogues, and 8,119 argument pairs. The statistics of DialogRE are shown in Table 1. We calculate both the F1 and F1_c [Yu *et al.*, 2020] scores as the evaluation metrics. F1_c is an evaluation metric to supplement the standard F1, computing by taking in the part of dialogue as input, instead of only considering the entire dialogue. We experiment Landre in both full-shot and few-shot settings [3]. In the few-shot setting (V2), 8, 16, and 32-shot experiments are conducted by using three different randomly sampled data [Son *et al.*, 2022]. For the experiments on ChatGPT, to keep OpenAI API costs under control, we randomly select 100 dialogues from DialogRE valid and test set, respectively. For direct extraction, the evaluation is consistent with typical evaluation as the extracted relation labels are generated by ChatGPT. The indirect extraction could generate extra tokens that irrelavent with our targets and requires answer cleansing. Specifically, after the ChatGPT generation, we pick up only the part of the answer text that first satisfies answer format [Kojima *et al.*, 2022].

[2]https://dataset.org/dialogre/

[3]https://github.com/1iguozheng/Landre

Baseline Models For a comprehensive performance evaluation, we compare the ChatGPT and Landre with both sequence-based and graph-based methods on the DialogRE dataset. Sequence-based methods usually utilize pre-trained language models to encode utterances in the dialogue, including BERT [Devlin *et al.*, 2019], BERT_s [Yu *et al.*, 2020], RoBERTa_s [Lee and Choi, 2021] and GRASP [Son *et al.*, 2022]. BERT_s is speaker-aware BERT. GRASP captures relational semantic clues of a given dialogue with the prompt marker strategy and the relational clue detection task. Graph-based methods include GDPNet [Xue *et al.*, 2021], TUCORE [Lee and Choi, 2021] and HiDialog [Liu *et al.*, 2023]. GDPNet develops a multiview latent graph to better capture the key features. TUCORE constructs a turn-level graph to capture the relational information. HiDialog leverages a heterogeneous graph module to polish the learned embeddings.

Experiment Details For the ChatGPT, we consider both few-shot prompting [Brown *et al.*, 2020] and zero-shot prompting [Kojima *et al.*, 2022] for evaluation. For the zero-shot prompting, we utilize proposed four types of prompt formats. For the few-shot prompting, we only experiment with direct extraction (i.e. vanilla prompting), considering 1-shot, 3-shot and 5-shot settings because of the maximum length limitation of ChatGPT. And the few-shot demonstrations are randomly sampled from a uniform distribution over all training dialogue examples. Due to ICL is known to have high variance [Zhao *et al.*, 2021], we compute the average performance of three times accessed results. We use the GPT-3.5 series model "gpt-3.5-turbo" for experiments. For open-source LLMs, we experiment the Landre framework with generative models including GPT-2 [Radford *et al.*, 2019], BART [Lewis *et al.*, 2020], T5 [Raffel *et al.*, 2020], BLOOM [Scao *et al.*, 2022] and LLaMA [Touvron *et al.*, 2023]. Specifically, we use the GPT-2-117M, BART-large-400M, BLOOM-560M, T5-large-770M and LLaMA-7B as the foundation models. We utilize LoRA [Hu *et al.*, 2022] to tune all these models for simplicity and efficiency. We set the rank r of the LoRA parameters to 8 and the merging ratio α to 32. The model is optimized with AdamW [Loshchilov and Hutter, 2019] using learning rate 1e-4 with a linear warm up [Goyal *et al.*, 2017] for the first 6% steps followed by a linear decay to 0. We train Landre for 5 epochs with batch size 4, and all experiments are conducted on a single Geforce GTX 3090 GPU.

4.2 Main Results of ChatGPT

The main results on DialogRE using ChatGPT as relation extractors are shown in Table 2. We have the following findings:

Indirect extraction tends to show much better performance than direct extraction for ChatGPT. Obviously, the vanilla prompting, which makes ChatGPT directly output the relation labels between an given argument pair, achieves the worst F1 scores in both few and zero-shot prompting manners compared to other three zero-shot prompting formats. Decomposing extraction process into multiple steps can improve the zero-shot performance by more than 15% in absolute F1 scores. Different with existing findings on prompting formats [Arora *et al.*, 2023], open-ended prompting tends to perform worse than restrictive and yes-no prompting in DRE.

Type	Method	Based-model (# Parameters)	**Dev (V1)** F1	**Dev (V1)** $F1_c$	**Test (V1)** F1	**Test (V1)** $F1_c$	**Dev (V2)** F1	**Dev (V2)** $F1_c$	**Test (V2)** F1	**Test (V2)** $F1_c$
Sequence-based	BERT [Devlin *et al.*, 2019]	$\text{BERT}_{\text{base}}$ (110M)	60.6	55.4	58.5	53.2	59.4	54.7	57.9	53.1
	BERT_s [Yu *et al.*, 2020]	$\text{BERT}_{\text{base}}$ (110M)	63.0	57.3	61.2	55.4	62.2	57.0	59.5	54.2
	RoBERTa_s [Lee and Choi, 2021]	$\text{RoBERTa}_{\text{large}}$ (355M)	-	-	-	-	72.6	65.1	71.3	63.7
	GRASP [Son *et al.*, 2022]	$\text{RoBERTa}_{\text{large}}$ (355M)	-	-	75.1	66.7	-	-	75.5	67.8
Graph-based	GDPNet [Xue *et al.*, 2021]	$\text{BERT}_{\text{base}}$ (110M)	67.1	61.5	64.9	60.1	61.8	58.5	60.2	57.3
	TUCORE [Lee and Choi, 2021]	$\text{RoBERTa}_{\text{large}}$ (355M)	-	-	-	-	74.3	67.0	73.1	65.9
	HiDialog [Liu *et al.*, 2023]	$\text{RoBERTa}_{\text{large}}$ (355M)	<u>76.9</u>	<u>69.8</u>	<u>77.7</u>	<u>69.0</u>	<u>76.4</u>	<u>68.5</u>	77.1	68.2
Few-shot prompting	Vanilla Prompting (1-shot)	GPT-3.5 (200B)	59.8	57.8	61.0	58.6	59.2	57.5	60.6	58.4
	Vanilla Prompting (3-shot)	GPT-3.5 (200B)	61.9	58.8	62.3	59.6	61.7	58.2	62.0	58.9
	Vanilla Prompting (5-shot)	GPT-3.5 (200B)	62.2	58.9	62.9	60.4	61.2	58.3	62.6	59.7
Zero-shot prompting	Vanilla Prompting (0-shot)	GPT-3.5 (200B)	57.8	55.4	59.5	57.0	57.4	55.7	58.9	56.3
	Restrictive Prompting (0-shot)	GPT-3.5 (200B)	74.6	72.4	74.2	70.3	75.3	72.0	74.9	70.8
	Yes-No Prompting (0-shot)	GPT-3.5 (200B)	75.2	72.4	73.5	71.7	75.7	72.6	74.2	71.2
	Open-Ended Prompting (0-shot)	GPT-3.5 (200B)	67.4	64.8	67.0	64.2	68.4	65.1	67.3	64.9
Generation-based	GPT-2 [Radford *et al.*, 2019]	GPT-2 (117M)	55.9	54.0	56.8	54.3	54.6	52.5	55.3	53.7
	BART [Lewis *et al.*, 2020]	$\text{BART}_{\text{large}}$ (400M)	61.3	59.8	61.8	60.2	60.7	59.5	61.0	59.8
	BLOOM [Scao *et al.*, 2022]	BLOOM (560M)	69.1	67.8	69.8	68.2	68.3	67.0	69.1	67.4
	T5 [Raffel *et al.*, 2020]	T5_{large} (770M)	72.4	70.4	72.8	70.5	70.9	69.0	71.6	69.8
	LLaMA [Touvron *et al.*, 2023]	LLaMA (7B)	**78.2**	**77.1**	**79.4**	**78.5**	**77.8**	**76.6**	**79.0**	**77.9**

Table 2: Performance of ChatGPT and Landre on DialogRE in the full-shot setting, averaged over three runs. Best results are **bold** and our re-implemented results are marked with <u>underline</u>.

In addtion, for few-shot prompting results, the improvements regarding to zero-shot prompting are small, where 5-shot results seem to have limited advantages compared to 3-shot.

ChatGPT surpasses several strong sequence-based and graph-based methods. Even with the zero-shot prompting, ChatGPT with elaborated prompting formats delivers better results than fully supervised RoBERTa_s (+2.9%∼7.5%) and TUCORE-GCN (+1.1%∼5.6%). Although ChatGPT is slightly worse than previous SoTA HiDialog (-0.7%∼3.5%) in standard setting, we should note that our limited budget restricted our study to a small set of prompt styles. It is possible that having a larger prompt design search space could narrow the performance gap between ChatGPT and HiDialog.

Performance gap between standard and conversational settings for ChatGPT is less than previous methods. Previous methods suffer grave declines when encountering conversational setting compared to standard setting (e.g., HiDialog drops around 7.1%∼8.9%). Interestingly, ChatGPT significantly narrows this performance gap (less than 5%), indicating ChatGPT can effectively identify the relations between an argument pair while tolerating the variant length of input dialogues. In contrast, encoder-only methods are sensitive to the length of input dialogues and hard to recover good representations for classification, because conversational setting is actually make the data distribution change in testing stage.

ChatGPT still has limitations. In summary, ChatGPT exhibits comparable performance in DRE compared to the previous SoTA methods, which highlights the ability of current LLMs to capture and comprehend complex linguistic patterns and dependencies within multi-turn dialogues. However, ChatGPT still faces two notable limitations. First, achieving good DRE performance requires elaborated prompting methods, resulting in higher computation cost and time consumption while ChatGPT is subject to request limitations. Besides, responses from ChatGPT frequently contain substantial explanatory content, where it may not align perfectly with expected answer format and difficult to cleanse and evaluate. Second, ChatGPT is not open-source, restricting its modification and customization, raising concerns about privacy protection and feasibility of local deployment. Therefore, we propose to use open-source LLMs in DRE, which alleviates two issues in DRE but also avoids the limitations of ChatGPT.

4.3 Main Results of Open-Source Models

Full-shot Setting We show the performance of Landre with different foundation models on the DialogRE dataset in Table 2 compared with other baselines. It is shown that Landre achieves competitive or superior performance compared to all of the previous methods, including the current SoTA methods GRASP and HiDialog. On the one hand, scaling up the foundation models in Landre brings significant performance improvements. From 117M to 7B, generation-based methods demonstrate over 20% absolute F1 gains in both standard and conversational settings. Note that the significant improvements are also observed in encoder-based methods (from 110M to 355M), but as the current encoder scale is much smaller than the decoder, the generation-based methods become more promising in future DRE. On the other hand, Landre shows its efficiency in conversational settings by narrowing the performance gap between F1 and $F1_c$ (e.g., Landre with LLaMA only drops around 0.9%∼1.2% compared to 7.1%∼8.9% in HiDialog), thereby indicating that generation-based methods effectively overcome the low information density of dialogues. Similar with ChatGPT, generation-based

Method	Based-model (# Parameters)	Shot K=8	K=16	K=32
RoBERTa	$\text{RoBERTa}_{\text{large}}$ (355M)	29.8	40.8	49.7
TUCORE	$\text{RoBERTa}_{\text{large}}$ (355M)	24.6	40.0	53.8
$\text{GRASP}_{\text{base}}$	$\text{RoBERTa}_{\text{base}}$ (125M)	**45.4**	52.0	56.0
$\text{GRASP}_{\text{large}}$	$\text{RoBERTa}_{\text{large}}$ (355M)	36.0	55.3	62.6
GPT-2	GPT-2 (117M)	38.5	44.2	54.0
BART	$\text{BART}_{\text{large}}$ (400M)	36.4	**56.9**	58.8
BLOOM	BLOOM (560M)	30.3	53.6	60.2
T5	T5_{large} (770M)	29.6	52.0	61.9
LLaMA	LLaMA (7B)	26.7	48.2	**65.4**

Table 3: Low-resource DRE performance of F1 scores. We use K = 8, 16, 32 (# of examples per relation) for few-shot experiments.

Relation label	Triple	GPT-2	ER.	LLaMA	ER.
per:alternate_names	819	59	7.2	35	4.3
per:girl/boyfriend	306	160	52.3	83	27.1
per:positive_impression	287	208	72.5	81	28.2
per:friends	278	63	22.7	57	20.5
per:title	164	6	3.7	6	3.7
per:spouse	126	81	64.3	31	24.6
per:siblings	122	49	40.2	26	21.3
per:children	103	88	85.4	25	24.3
per:parents	103	60	58.3	37	35.9
per:negative_impression	102	102	100.0	51	50.0

Table 4: Error rates of Top-10 relations on the DialogRE dev and test set. **Triple** denotes the total number of triples of each relation in the dev and test set. **GPT-2** and **LLaMA** denote the number of wrong predictions and **ER.** denotes the corresponding error rate.

methods tend to perform robust encountering various length of input dialogues, and they are able to extract the golden relation based on partial dialogues with enough arguments and relation information. These results imply that guiding the model to pay attention to relation labels with a prompt tuning approach can be more effective than adding additional features and layers in encoder-based methods.

Few-shot Setting As presented in Table 3, Landre exhibits robust performance in few-shot settings. For example, Landre with GPT-2 surpasses TUCORE regardless of the number of shots. Except for the 8-shot setting, Landre presents new SoTA performance compared to previous methods. However, GRASP with smaller encoder still delivers better 8-shot performance compared to others. Moreover, although Landre with LLaMA is much better than other baselines in the 32-shot setting, the limited performance of LLaMA in the 8-shot and 16-shot setting can be attributed to an insufficient number of examples for tuning large-scale parameters.

5 Analysis

In this section, we analyze various aspects of Landre to provide a deeper understanding of generation-based methods performing in DRE task. Specifically, we analyze the results of GPT-2 and LLaMA on the DialogRE (V2) dev and test set.

5.1 Error Analysis

We calculate the error rate of each relation to investigate the fine-grained performance of Landre with GPT-2 and LLaMA. To mitigate the deviation caused by imbalanced relation types, we analyze the 10 most frequently appearing relations whose corresponding sample numbers are above 100. The overall results are summarized in Table 4, where scaling up model size consistently reduces the error rates of all relations. For most interpersonal relations that frequently appear such as per:girl/boyfriend, per:spouse and per:children, LLaMA delivers prominent performance improvements compared to GPT-2, showing increasing the model size can better capture intra-turn and inter-turn information, distinguishing the relationship semantic differences in the dialogues.

For specific relations such as per:alternate_names and per:title, GPT-2 and LLaMA achieve similar precision with low error rates. First, it is not surprising that the relation per:alternate_names is easy for models because it is the most frequently appearing relation and typically can be identified within a single-turn information (e.g., S1 calls S2's name then S2 responds). For relation per:title, interestingly, two models have the same number of wrong predictions. We compare the error cases of two models and find 5 out of 6 samples overlapping. The 5 samples are mistakenly annotated as per:title between PER and NAME in the first place while the subject and object types of per:title should be PER and STRING [Yu *et al.*, 2020]. Because most of argument pair types belong to PER and NAME, per:title is relatively easy to distinguish. In other words, generative methods can achieve near perfect performance on relation per:title because of the information from argument mentions, most of which is type information.

Another interesting finding is that GPT-2 delivers 100% error rate on relation per:negative_impression. We thus examine the proportion of its predicted results with respect to 102 samples annotated with per:negative_impression. We discover that 66.6% of samples are predicted with relations per:girl/boyfriend, per:positive_impression and per:friend for GPT-2, where these relations may share same argument pairs with per:negative_impression and tend to confuse the models because friends and girl/boyfriends often express impressions of each other. In addition, some specific relations such as per:acquaintance (not shown in Table 4) are challenging to Landre, where the error rates of per:acquaintance are 100% and 97.8% for GPT-2 and LLaMA, respectively. The main reasons include difficulty in defining such relation (i.e., it is difficult to distinguish between acquaintances and friends) and low trigger ratio (around 22%) [Yu *et al.*, 2020].

5.2 Qualitative Analysis

Inverse relations. We group the test set of DialogRE according to the relation types into three subsets: (I) asymmetric, when a relation type differs from its inversion (e.g. per:children and per:parents); (II) symmetric, when a relation type is the same as its inversion (e.g. per:friends); (III) other, when a relation type does not have inversion (e.g. per:title). We compare the performance of Landre with baselines and report the results in Table 5. We observe that GPT-2 struggles to capture inverse relations compared to BERT. Moreover, Landre with LLaMA shows comparable performance

Method	I	II	III
BERT [Devlin *et al.*, 2019]	42.5	60.7	65.6
GDPNet [Xue *et al.*, 2021]	47.4	59.8	68.1
RoBERTa$_s$ [Lee and Choi, 2021]	57.4	69.3	79.6
TUCORE [Lee and Choi, 2021]	62.3	**71.3**	79.9
HiDialog [Liu *et al.*, 2023]	**76.6**	70.5	80.9
GPT-2 [Radford *et al.*, 2019]	39.0	43.8	72.2
LLaMA [Touvron *et al.*, 2023]	75.3	69.3	**88.2**

Table 5: All methods performance on DialogRE. We break down the performance into three groups (I) asymmetric inverse relations, (II) symmetric inverse relations, and (III) others.

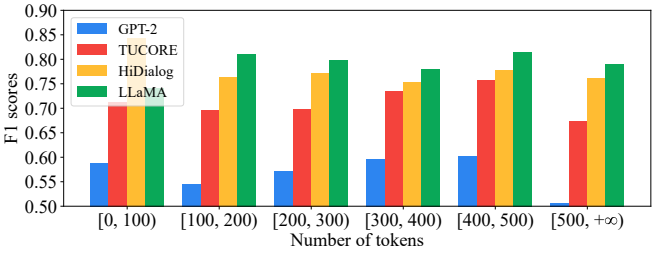

Figure 4: Analysis of robustness of Landre tackling increasing utterance length compared to baseline TUCORE and HiDialog.

on both asymmetrical and symmetrical relations compared to TUCORE and HiDialog, while it delivers exceptional results on other relation types. And this is also why GPT-2 and LLaMA have low error rates on specific relations such as per:alternate_names which is the most frequently appearing relation and contributes the majority of overall performance.

Utterance lengths. Compared to encoder-based methods, generation-based methods enable to handle dialogues of various lengths. We further divide the samples in the test set into six groups according to their lengths and report the F1 score for each group achieved by Landre and previous SoTA. As shown in Figure 4, LLaMA outperforms HiDialog in all groups except the group with less than 100 tokens, maintaining decent performance for long sequences.

5.3 Applicability Analysis

To evaluate the robustness and generalization of generation-based methods, we evaluate Landre with LLaMA on MELD [Poria *et al.*, 2019] and EmoryNLP [Zahiri and Choi, 2018] datasets, which are designed for emotion recognition in conversations (ERC). Following previous works [Lee and Choi, 2021; Son *et al.*, 2022], the ERC task is converted into DRE and the weighted-F1 is calculated to evaluate the MELD and EmoryNLP datasets. As presented in Table 6, the performance of Landre surpasses that of GRASP and HiDialog in both MELD and EmoryNLP, demonstrating significant promise in dialogue understanding.

Method	MELD	EmoryNLP
RoBERTa [Liu *et al.*, 2019]	62.0	37.3
TUCORE [Lee and Choi, 2021]	65.4	39.2
GRASP [Son *et al.*, 2022]	65.6	40.0
HiDialog [Liu *et al.*, 2023]	65.7	38.1
LLaMA [Touvron *et al.*, 2023]	**66.2**	**40.7**

Table 6: Experimental results of LLaMA on MELD and EmoryNLP.

6 Related Work

Large Language Models. Besides the "pre-train and fine-tune" paradigm [Devlin *et al.*, 2019], pre-trained LLMs possess characteristics that are advantageous for few-shot [Brown *et al.*, 2020] and zero-shot [Kojima *et al.*, 2022] learning, whereby appropriate prompts are used to effectively guide the model towards generating desired task outputs, thus beginning an era of "pre-train and prompt". However, prompting without updating parameters generally brings sub-optimal results and the efficient utilization of LLMs remains a significant challenge. Recent advancements in parameter-efficient fine-tuning (PEFT) techniques have effectively alleviated this problem, such as LoRA [Hu *et al.*, 2022] and Prefix Tuning [Liu *et al.*, 2022]. In this study, we initially investigate the capabilities of fine-tuned open source (e.g. LLaMA [Touvron *et al.*, 2023]) and non fine-tuned closed source (e.g. ChatGPT [OpenAI, 2022]) LLMs for DRE.

Dialogue Relation Extraction. Most previous works on RE focused on sentence-level RE [Li *et al.*, 2022; Wang *et al.*, 2023b; Wang *et al.*, 2023a], but recently cross-sentence RE has been studied more because many relational facts are expressed in multiple sentences in practice [Yao *et al.*, 2019]. Dialogue relation extraction (DRE) requires reasoning over multiple sentences in a dialogue to predict the relations between two arguments [Yu *et al.*, 2020]. Current popular DRE methods can be divided into sequence-based methods [Yu *et al.*, 2020; Son *et al.*, 2022] and graph-based methods [Xue *et al.*, 2021; Lee and Choi, 2021; Liu *et al.*, 2023]. The former utilize encoder-only language models such as BERT [Devlin *et al.*, 2019] and RoBERTa [Liu *et al.*, 2019] to encode utterances in the dialogue. The latter construct various heterogeneous graphs over dialogues to capture the relational information. And both two kind of methods use the softmax classifiers for relation classification. In this study, we aim to investigate the capabilities of generative methods for DRE. Due to the rapid development of decoder-based LLMs, we anticipate that Landre could serve as a compelling baseline or plug-in module for future work in the community.

7 Conclusion

In this study, we conduct an initial examination of generative LLMs in DRE, showcasing its excellence over previous sequence-based and graph-based methods. To solve the limitations of ChatGPT, we present Landre, an LLM-driven DRE framework based on smaller, open-source foundation models. By simply prompting tuning with LoRA, Landre delivers better performance than ChatGPT and achieves new state-of-the-art performance on DRE benchmarks. Comprehensive evaluations across different experimental settings showcase the unique characteristics of generation-based methods, opening new frontiers for advancing the research of DRE.

Acknowledgments

We thank the reviewers for their insightful comments. This work was supported by National Science Foundation of China (Grant Nos.62376057) and the Start-up Research Fund of Southeast University (RF1028623234). All opinions are of the authors and do not reflect the view of sponsors.

References

[Arora *et al.*, 2023] S Arora, A Narayan, M Chen, L Orr, N Guha, K Bhatia, I Chami, F Sala, and C Ré. Ask me anything: A simple strategy for prompting language models. In *ICLR*, 2023.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[Cabot and Navigli, 2021] Pere-Lluís Huguet Cabot and Roberto Navigli. Rebel: Relation extraction by end-to-end language generation. In *Findings of EMNLP*, 2021.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[Giorgi *et al.*, 2022] John Giorgi, Gary D Bader, and Bo Wang. A sequence-to-sequence approach for document-level relation extraction. In *BioNLP*, 2022.

[Goyal *et al.*, 2017] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

[Lee and Choi, 2021] Bongseok Lee and Yong Suk Choi. Graph based network with contextualized representations of turns in dialogue. In *EMNLP*, 2021.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

[Li *et al.*, 2022] Guozheng Li, Xu Chen, Peng Wang, Jiafeng Xie, and Qiqing Luo. Fastre: Towards fast relation extraction with convolutional encoder and improved cascade binary tagging framework. In *IJCAI*, 2022.

[Li *et al.*, 2023] Guozheng Li, Peng Wang, and Wenjun Ke. Revisiting large language models as zero-shot relation extractors. In *Findings of EMNLP*, 2023.

[Li *et al.*, 2024] Guozheng Li, Wenjun Ke, Peng Wang, Zijie Xu, Ke Ji, Jiajun Liu, Ziyu Shang, and Qiqing Luo. Unlocking instructive in-context learning with tabular prompting for relational triple extraction. *arXiv preprint arXiv:2402.13741*, 2024.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2022] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *ACL*, 2022.

[Liu *et al.*, 2023] Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. Hierarchical dialogue understanding with special tokens and turn-level attention. In *ICLR*, 2023.

[Liu *et al.*, 2024] Jiajun Liu, Wenjun Ke, Peng Wang, Ziyu Shang, Jinhua Gao, Guozheng Li, Ke Ji, and Yanhe Liu. Towards continual knowledge graph embedding via incremental distillation. In *AAAI*, 2024.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[Ma *et al.*, 2023] Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of EMNLP*, 2023.

[OpenAI, 2022] OpenAI. Introducing chatgpt, 2022.

[Paolini *et al.*, 2021] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In *ICLR*, 2021.

[Poria *et al.*, 2019] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, 2019.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(1):5485–5551, 2020.

[Sanh *et al.*, 2022] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun

Raja, et al. Multitask prompted training enables zero-shot task generalization. In *ICLR*, 2022.

[Scao *et al.*, 2022] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[Shang *et al.*, 2024] Ziyu Shang, Wenjun Ke, Nana Xiu, Peng Wang, Jiajun Liu, Yanhui Li, Zhizhao Luo, and Ke Ji. Ontofact: Unveiling fantastic fact-skeleton of llms via ontology-driven reinforcement learning. In *AAAI*, 2024.

[Son *et al.*, 2022] Junyoung Son, Jinsung Kim, Jungwoo Lim, and Heuiseok Lim. GRASP: Guiding model with RelAtional semantics using prompt for dialogue relation extraction. In *COLING*, 2022.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wadhwa *et al.*, 2023] Somin Wadhwa, Silvio Amir, and Byron C Wallace. Revisiting relation extraction in the era of large language models. In *ACL*, 2023.

[Wan *et al.*, 2023] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. Gpt-re: In-context learning for relation extraction using large language models. In *EMNLP*, 2023.

[Wang *et al.*, 2023a] Peng Wang, Tong Shao, Ke Ji, Guozheng Li, and Wenjun Ke. fmlre: a low-resource relation extraction model based on feature mapping similarity calculation. In *AAAI*, 2023.

[Wang *et al.*, 2023b] Peng Wang, Jiafeng Xie, Xiye Chen, Guozheng Li, and Wei Li. Pascore: a chinese overlapping relation extraction model based on global pointer annotation strategy. In *IJCAI*, 2023.

[Xue *et al.*, 2021] Fuzhao Xue, Aixin Sun, Hao Zhang, and Eng Siong Chng. Gdpnet: Refining latent multi-view graph for relation extraction. In *AAAI*, 2021.

[Yao *et al.*, 2019] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In *ACL*, 2019.

[Yu *et al.*, 2020] Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. Dialogue-based relation extraction. In *ACL*, 2020.

[Zahiri and Choi, 2018] Sayyed M Zahiri and Jinho D Choi. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI*, 2018.

[Zhang *et al.*, 2017] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *EMNLP*, 2017.

[Zhang *et al.*, 2023] Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of ACL*, 2023.

[Zhao *et al.*, 2021] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, 2021.