CORECODE: A Common Sense Annotated Dialogue Dataset with Benchmark Tasks for Chinese Large Language Models

Dan Shi[1], **Chaobin You**[1], **Jiantao Huang**[2], **Taihao Li**[2], **Deyi Xiong**[1]*

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Zhejiang Lab, Hangzhou, China
{shidan, chaobinyou, dyxiong}@tju.edu.cn, {jthuang, lith}@zhejianglab.com

Abstract

As an indispensable ingredient of intelligence, commonsense reasoning is crucial for large language models (LLMs) in real-world scenarios. In this paper, we propose CORECODE, a dataset that contains abundant commonsense knowledge manually annotated on dyadic dialogues, to evaluate the **co**mmonsense **re**asoning and commonsense **co**nflict **de**tection capabilities of Chinese LLMs. We categorize commonsense knowledge in everyday conversations into three dimensions: entity, event, and social interaction. For easy and consistent annotation, we standardize the form of commonsense knowledge annotation in open-domain dialogues as "domain: slot = value". A total of 9 domains and 37 slots are defined to capture diverse commonsense knowledge. With these pre-defined domains and slots, we collect 76,787 commonsense knowledge annotations from 19,700 dialogues through crowdsourcing. To evaluate and enhance the commonsense reasoning capability for LLMs on the curated dataset, we establish a series of dialogue-level reasoning and detection tasks, including commonsense knowledge filling, commonsense knowledge generation, commonsense conflict phrase detection, domain identification, slot identification, and event causal inference. A wide variety of existing open-source Chinese LLMs are evaluated with these tasks on our dataset. Experimental results demonstrate that these models are not competent to predict CORECODE's plentiful reasoning content, and even ChatGPT could only achieve 0.275 and 0.084 accuracy on the domain identification and slot identification tasks under the zero-shot setting. We release the data and codes of CORECODE at https://github.com/danshi777/CORECODE to promote commonsense reasoning evaluation and study of LLMs in the context of daily conversations.

Introduction

Commonsense reasoning is a crucial component of intelligence(Liu and Singh 2004; Cambria et al. 2011), which involves the ability to make logical deductions, infer implicit information and apply background knowledge to solve problems as well as understand the world. In recent years, exploring and improving the ability of NLP models for the acquisition and application of commonsense knowledge has been attracting growing interest, leading to extensive research in this field (Lv et al. 2020; Wang et al. 2020; Liu et al. 2022).

It is widely acknowledged that LLMs, trained on a huge amount of data, are able to obtain broad knowledge covering a wide range of domains (Rae et al. 2021; Hoffmann et al. 2022; Touvron et al. 2023; Du et al. 2022a; Guo et al. 2023), including commonsense knowledge (West et al. 2022; Bian et al. 2023; Bang et al. 2023). However, commonsense reasoning is still regarded as a major challenge for LLMs (Zhou et al. 2020; Bhargava and Ng 2022). Studies disclose that LLMs fall short in performing adequate commonsense reasoning (Wei et al. 2022). For example, ChatGPT[1] does not precisely know what the needed commonsense knowledge for answering a specific question is (e.g., questions in social and temporal domains) (Bian et al. 2023).

To mitigate this issue, we propose CORECODE (**Co**mmonsense **Re**asoning and **Co**nflict **De**tection in dialogues), a dataset that contains abundant commonsense knowledge manually annotated on Chinese dyadic dialogues, to assess how much commonsense knowledge the LLMs have gained and how well they can be improved in commonsense reasoning and conflict detection with the annotated knowledge in CORECODE.

Specifically, we focus on annotating fine-grained commonsense knowledge in multi-turn dyadic dialogues. The knowledge annotated in a dialogue is context-sensitive and grounded exclusively in that particular dialogue. Inspired by the annotation convention used in task-oriented dialogue, in which dialogue states are denoted in the form of "domain: slot = value", e.g. "hotel: price range = moderate" (Budzianowski et al. 2018; Zhu et al. 2020; Quan et al. 2020), we standardize the representation of commonsense knowledge in open-domain dialogues also in the form of "domain: slot = value". We categorize commonsense knowledge into three dimensions, namely entity, event, and social interaction, and then construct an ontology over these dimensions, which defines all possible domains for each dimension and all possible slots for each domain. Thanks to the guidance of this ontology, crowdsourcing annotators are able to conveniently annotate fine-grained commonsense knowledge in a consistent way.

Over the curated dataset, we develop six benchmark tasks:

* Corresponding author

[1]https://openai.com/blog/chatgpt

commonsense knowledge filling, commonsense knowledge generation, commonsense conflict phrase detection, domain identification, slot identification and event causal inference. These tasks, organized in different forms (e.g., multiple-choice, span extraction, and text generation), facilitate the evaluation and enhancement of commonsense reasoning in LLMs.

We conduct numerous experiments on CORECODE, attempting to explore two main research questions: (1) Can LLMs master and apply commonsense knowledge well enough to achieve good performance on these tasks? (2) How much further improvements can be obtained by LLMs if they are fine-tuned on CORECODE? Extensive experiments demonstrate that our benchmark tasks are challenging for existing Chinese LLMs, as all evaluated LLMs perform poorly on most tasks. We also show that although the performance of LLMs improves after being fine-tuned on CORECODE, they fail to obtain robust commonsense reasoning ability. When perturbations are introduced, the fine-tuning performance has significantly dropped.

Related Work

A variety of datasets and benchmarks focusing on different aspects of commonsense knowledge over textual inputs have been proposed, including science common sense datasets ARC (Clark et al. 2018) and QASC (Khot et al. 2020), temporal common sense dataset MC-TACO (Zhou et al. 2019), numerical common sense dataset NumerSense (Lin et al. 2020), physical common sense dataset PIQA (Bisk et al. 2020), social common sense dataset Social IQA (Sap et al. 2019b) and general common sense datasets CommonsenseQA (Talmor et al. 2019), OpenBookQA (Mihaylov et al. 2018), and WSC (Levesque, Davis, and Morgenstern 2012). These datasets only examine the model's knowledge and ability in a certain commonsense aspect in the form of multiple-choice questions.

Meanwhile, there have also been many studies devoted to annotating commonsense knowledge within dialogues. ATOMIC (Sap et al. 2019a; Hwang et al. 2021) is one such dataset that consists of a large set of inference types. However, ATOMIC is context-insensitive, as its commonsense reasoning operates on phrases taken out of context, disregarding whether an event is performed by the same individual. TIMEDIAL (Qin et al. 2021) focuses on the time reasoning ability of language models in dialogues, while CICERO (Ghosal et al. 2022) provides cause, subsequent events, prerequisites, motivations, and emotional reactions for utterances in dialogues, focusing on these five event-related reasoning types. Both datasets cover only a specific aspect of commonsense knowledge. CIDER (Ghosal et al. 2021) extracts knowledge in dialogues into knowledge triplets, which covers fewer commonsense knowledge types than us. For example, *subsequent event*, *subsequent emotional reaction*, *frequency* are beyond the scope of CIDER.

To the best of our knowledge, CORECODE is the first large-scale Chinese dialogue-oriented commonsense knowledge annotation dataset involving comprehensive commonsense knowledge in three dimensions: entity, event, and social interaction, covering a large number of perspectives such as attributes, time, space, and causality. Yet another feature that must be emphasized is that within CORECODE, we manually provide phrases corresponding to the phrases in an original dialogue, which are against common sense in that context. This aims to probe the model's capacity to detect and locate such phrases that are inconsistent with the context in terms of commonsense reasoning.

Dataset Creation

The raw data of CORECODE is derived from NaturalConv (Wang et al. 2021) and DuLeMon (Xu et al. 2022) datasets, both of which contain multi-turn dialogues between two people. Dialogues in NaturalConv involve a variety of topics (including but not limited to sports, entertainment, and technology). We first take an automatic screening method to identify dialogues that are rich in commonsense knowledge, following Zhou et al. (2021).

Specifically, we first identify candidate concepts (nouns, verbs, adjectives) in each dialogue turn using part-of-speech tagging. We then query the ConceptNet using the identified concepts in each utterance to obtain a list of one-hop commonsense triples in the form of (e_1, r, e_2). Next, we examine if the entity e_2 from the triple appears in the concept set of the succeeding utterance. If there is a match, it indicates a potential commonsense link between the two utterances.

Unlike Zhou et al. (2021) who retain dialogues with only one commonsense triple match, we employ a stricter criterion by retaining dialogues where more than three commonsense triple matches are detected. This ensures that the kept dialogues possess a substantial amount of commonsense reasoning. The statistics of the screening results on NaturalConv and DuLeMon are shown in our arXiv version[2].

Moreover, to differentiate between the two sides of the conversation, we employ the notation "A: " or "B: " preceding each utterance to denote the respective speaker.

Data Annotation

Over the selected dialogues, we perform commonsense knowledge annotation. To guarantee the consistency of annotations across multiple crowd-sourced workers, we adopt a standardized annotation procedure.

We categorize commonsense knowledge in everyday conversations into three dimensions: *entity*, *event*, and *social interaction*. Crowd-sourced workers first need to identify specific instances under these three dimensions from dialogues. Then, with the assistance of linguists, we divide each of these three dimensions into multiple domains to which their commonsense knowledge belongs, and define different slots for each domain, forming a two-level hierarchical taxonomy. Such design is guided by three fundamental principles: *coverage*, *exclusivity*, and *easiness*. The *coverage* rule ensures that the commonsense knowledge system encompasses nearly all conceivable types of commonsense knowledge in dialogues. *Exclusivity* mandates that each commonsense knowledge type remains distinct, devoid of any overlap with other types. Lastly, the *easiness* principle indicates that the commonsense knowledge system is

[2]https://arxiv.org/abs/2312.12853

straightforward for annotators to employ. With this convention, crowd-sourced workers are instructed to annotate the identified instances with commonsense knowledge in the form of "domain: slot = value". In addition to such annotations, they are also required to provide phrases that, in terms of common sense, conflict with the original textual context. Below, we describe each step in detail.

Entity, Event, Social Interaction Recognition. The first step of the annotation process is to identify specific instances of entity, event, and social interaction that exist in dialogues, according to the following definitions.

- **Entities** refer to objectively existing and distinguishable physical objects in the real world, either representing a general category of people or things, such as "*cats*", "*movies*", or referring to specific individuals or objects, such as "*Yao Ming*", "*Wolf Warrior*", etc.
- **Events** are typically text spans in the form of "subject + predicate" or "subject + predicate + object". They are fine-grained semantic units that describe the state of entities and their actions (Zhou et al. 2022). For example, "*He looks very excited*" describes the state of the subject, and "*He broke his toy*" illustrates an action where the subject interacts with the object.
- **Social interactions** refer to the set of rules and guidelines that constrain people's behavior when interacting with others. They encompass a collection of social norms and customs that people are expected to adhere to (Bian et al. 2023). For instance, "*It is customary to knock on the door before entering someone else's room*".

Annotation of Involved Commonsense Knowledge. Under each of the three dimensions, we define domains and slots. For entities, we divide the relevant commonsense knowledge into three corresponding domains: attribute, comparison, and space. These domains capture specific properties of the object itself, relationships between the object and other objects, and relationships between the object and the spatial environment in which it is located, respectively. Under each domain, there are further divisions into different slots. For example, under the attribute domain, there are slots "Is", "Is A", "Has", "Is Made Of", and so on. For events, relevant commonsense knowledge includes the prerequisite, cause, and consequence of an event, as well as the temporal and spatial factors associated with the event. For social interactions, we focus on the social norms that humans follow. Instead of subdividing into multiple domains, we divide seven slots under the social norms domain. There are 9 domains and 37 slots included in the three dimensions in total. The full inventory of all domains and slots can be found in our arXiv version.

The second step of the annotation process is to label each entity, event or social interaction instance with its commonsense knowledge in the form of "domain: slot = value". The annotated "value" need not necessarily be an exact span extracted from the original dialogue, but can be a grammatically correct and semantically fluent clause summarized from the dialogue, ensuring that the event and its "domain: slot = value" in isolation is informationally complete and logically consistent. It has been emphasized to annotators that for the "event cause" slot in the "cause" domain and the "subsequent event" slot in the "consequence" domain, annotations should take the form of an event, i.e., either in the structure of "subject + predicate" or "subject + predicate + object". In addition, the annotators need to indicate which phrases or clauses in the original dialogue led to the identification of this commonsense knowledge, so as to provide a basis for the next step.

Rewriting of Commonsense Conflict Phrases. Finally, for each set of phrases from the original dialogue indicated in the previous step, annotators are required to choose one phrase and provide it with the following two commonsense conflict phrases:

(1) **Commonsense Conflict Phrase 1**: This phrase should be obtained by conforming to the minimal modification principle, i.e., modifying only one or two words in the original phrase. There should be a commonsense conflict or error after using this phrase to replace the original phrase in the dialogue.

(2) **Commonsense Conflict Phrase 2**: This phrase should be created by modifying as many words as possible in the original phrase in compliance with the maximum modification principle. When constructing this phrase, annotators can include words that appear in the dialogue to maintain consistency with the dialogue's context. However, it is crucial to ensure as much as possible that the meaning of this phrase differs from the Commonsense Conflict Phrase 1.

The purpose of this annotation step is to explore whether LLMs are able to detect the location of phrases that conflict with the dialogue context in terms of common sense. Therefore, annotators must ensure that after replacing the original phrase in the dialogue with the annotated conflict phrase, there should be only a commonsense error while the dialogue maintains grammatically correct and fluent.

To comprehensively evaluate the commonsense reasoning ability of LLMs, we propose two distinct annotated subsets with varying difficulty levels. During the annotation procedure on 9.7K dialogues, we represent the subject and object of events using the speaker indicators "A" or "B" from the dialogue and group these annotated instances as an EASY set. A HARD set is annotated on another 10K dialogues, where "x" is uniformly employed to denote the subject of all events, while "y" is used to represent the predicate of all events, regardless of the dialogue participant to whom the event pertains. Significant challenges in reasoning through events are provided in the HARD set, as LLMs are required to first deduce and locate the event initiator before reasoning.

Annotation Quality Control

In order to standardize the annotation form and control the quality of common sense annotations, we design and develop a knowledge acquisition platform where crowdsourced workers need to properly click on the appropriate buttons and fill in the corresponding values given the dialogue history.

We adopt a very strict quality control protocol to ensure

the quality of annotations. First, we train two reviewers with 200 dialogues. The annotation consistency of the two reviewers is high, with an average Cohen's Kappa (McHugh 2012) of 80.7% across the annotation tasks. We only hire annotators who have relevant experience in text annotation, e.g., those who have participated in annotation tasks such as Chinese multi-turn dialogue writing and correction, entity extraction or syntactic structure annotation in Chinese texts.

Second, 200 candidate workers participate in a pre-annotation stage. They adhere to the prescribed rules to annotate dialogues. The two reviewers will review annotations of these participants to distinguish whether the annotations meet the requirements. The process has an elimination rate of roughly 80%, with 43 labelers passing this stage.

Third, we proceed to the training phase. We divided the participants into groups of 5 people each. We train 1-2 quality inspectors within each group, who in turn are responsible for the instruction of the annotators. During this progression, quality inspectors evaluate the rule comprehension and error correction capabilities of the annotators. Those who do not meet the criteria are subjected to further training or eliminated from the process.

At last, 6 quality inspectors with an average Cohen's Kappa of 59.4%, as well as 15 annotators, proceed to the formal annotation stage. We take iterative verification and revision during this stage. Any data deemed unsatisfactory will be returned for revision until they are qualified.

Overall Statistics

The overall statistics of the annotated dataset are shown in Table 1. After annotating on 19.7K dialogues, we obtained 76,787 annotations, each comprising the original dialogue, an entity/event/social interaction instance, a commonsense knowledge represented by a domain-slot-value triplet, the involved phrase from the original dialogue, and two commonsense knowledge conflict phrases. The average number of turns and tokens per dialogue is 19.40 and 501.58, indicating that the annotated dialogues are lengthy and informative. The social interaction dimension's knowledge primarily serves to constrain behavior but is seldom mentioned in dialogues, resulting in limited annotated commonsense knowledge for this dimension. The annotations for entity, event, and social interaction dimensions constitute 58.42%, 41.54%, and 0.03% of the overall annotations, respectively.

	HARD	EASY	Total
# dialogues	10,000	9,700	19,700
Max. turns per dialogue	26	26	26
Min. turns per dialogue	14	16	15
Avg. turns per dialogue	18.69	20.10	19.40
Max. # tokens per dialogue	1,002	953	977.5
Min. # tokens per dialogue	194	231	212.5
Avg. # tokens per dialogue	464.18	538.98	501.58
Avg. # tokens per turn	24.83	26.81	25.82
# annotated instances	37,777	39,010	76,787
# annotated entities	21,320	23,541	44,861
# annotated events	16,439	15,461	31,900
# annotated social interactions	18	8	26
# domain-slot-value triplets	37,777	39,010	76,787
# commonsense conflict phrases	75,554	78,020	153,574

Table 1: Overall statistics of the CORECODE dataset.

Benchmark Tasks

We use our dataset as a testbed and define 6 tasks in different forms, attempting to evaluate dialogue-level commonsense reasoning capabilities of Chinese LLMs. For each task, we provide both its definition and associated prompt that is constructed to allow LLMs to complete the task in the continuation to the prompt.

Commonsense Knowledge Filling

Task definition. This task is to fill desirable commonsense knowledge into a masked dialogue where a commonsense phrase is replaced with [MASK]. In order to automatically assess the performance of the task, we formulate the task in the form of multiple-choice questions.

Prompt. The input prompt to LLMs for this task consists of the question, masked dialogue, answer choices, and suffix: question \n masked dialogue \n (a) $phrase_1$ (b) $phrase_2$ (c) $phrase_3$ \n "answer: the correct option is". The three phrases are the corresponding masked commonsense phrase and two manually composed commonsense conflict phrases. See our arXiv version for examples of all tasks.

Commonsense Knowledge Generation

Task definition. We frame this task as a generative task that takes the annotated commonsense knowledge values as ground truth and asks LLMs to generate the values according to the dialogue context.

Prompt. The input prompt is formatted as: dialogue \n question \n "answer:", where the question is formed by the entity/event/social interaction and its annotated slot through a predefined template and some explanatory text.

Commonsense Conflict Phrase Detection

Task definition. We define this task as a span extraction task. We replace the corresponding phrases in the original dialogue with the annotated commonsense conflict phrases, and ask LLMs to extract the commonsense conflict phrases.

Prompt. The prompt format is: dialogue with replaced commonsense conflict phrase \n question \n "answer:".

Domain Identification

Task definition. This task is also defined as a multiple-choice-question task. Take the entity dimension as an example, LLMs are required to select the domain to which the relationship between an entity and its annotated value belongs, based on the given dialogue context. Since the social interaction dimension includes a single domain, this task is performed on the entity and event dimensions.

Prompt. The prompt is formatted like: question \n entity or event \n annotated value \n dialogue \n (a) $domain_1$ (b) $domain_2 \cdots$ (x) $domain_n$ \n "answer: the correct domain is".

Slot Identification

Task definition. This task is similar to the Domain Identification task but involves selecting from more fine-grained slot options and is performed across all three dimensions.

Prompt. The input prompt is in the form of: question \n entity/event/social interaction \n annotated value \n dialogue \n (a) $slot_1$ (b) $slot_2$ $\cdots$ (x) $slot_n$ \n "answer: the correct option is".

Event Causal Inference

Causal inference is one of the crucial reasoning abilities of human intelligence, which involves establishing the correct cause-and-consequence relationships between events. These relationships are captured in the "cause: event cause" slot and the "consequence: subsequent event" slot of our taxonomy. We specially design three generative event causal inference tasks that utilize the annotated knowledge involved in these two slots.

- **Subtask 1: Event Cause Inference.** Given the dialogue and event, LLMs are required to generate the cause of the event.
- **Subtask 2: Subsequent Event Inference.** Given the dialogue and event, the consequence of the event is generated by LLMs.
- **Subtask 3: Clipped Subsequent Event Inference.** Given the event and the truncated dialogue where the context succeeding the event is discarded, we require LLMs to generate the consequence of the event.

Experiments

Evaluated LLMs

We evaluated a diverse list of Chinese LLMs that cover a variety of training processes and scales[3]: (1) LLMs only being pre-trained on large-scale training corpora, including GLM-10B (Du et al. 2022b) and BLOOM-7.1B (Scao et al. 2022), (2) LLMs being both pre-trained and instruction-tuned, including ChatGLM-6B[4], ChatGLM2-6B[5], MOSS-SFT-16B[6], Baichuan-7B[7], BLOOMZ-1.7B, BLOOMZ-7.1B, BLOOMZ-7.1B-MT (Muennighoff et al. 2022), and BELLE-7B, which is the SFT version based on BLOOMZ-7.1B-MT. We used two variants of BELLE fined-tuned on 200K and 2M instructions separately, i.e., BELLE-7B-0.2M[8] and BELLE-7B-2M[9]. We also evaluated two variants Chinese-Alpaca-Plus-7B and Chinese-Alpaca-Plus-13B of Chinese-Alpaca-Plus (Cui, Yang, and Yao 2023). We experimented on the recommended hyperparameter settings of all LLMs. We also evaluated ChatGPT (i.e., GPT-3.5-turbo) from OpenAI as a reference.

Furthermore, to explore the impact of in-context learning (ICL) on model performance, we also carried out experiments on ChatGLM-6B under the few-shot settings, including 1-shot, 3-shot and 5-shot settings.

Evaluation Metrics

For the commonsense knowledge filling, domain identification and slot identification tasks (we refer these three tasks to the selection tasks), we used the accuracy of selecting the correct answer as the evaluation metric. During inference, we have found that even if we explicitly state in the prompt that models should output only the answer option indicator (i.e. a, b, c, etc.), not all models follow this instruction. There is no uniformity in the form of answers generated by each model. Moreover, sometimes models output answers with rationales attached. In order to avoid the underestimation of the model performance due to the varying output formats, we adopted a series of filtering measures to find the correct answer in the output as much as possible. For example, in the case where the ground-truth is "(a) premise", the generated answers "a", "A", "(a)", "(A)", "a)", "A)", "(a)premise", "(a) premise", "premise" are all counted as correct.

For the span extraction task, i.e., the commonsense conflict phrase detection task, we used F1 and EM scores calculated by comparing model outputs to ground-truth answers.

For the two generation tasks, namely the commonsense knowledge generation task and the event causal inference task, we evaluated LLMs with F1 and EM scores together with reference based metrics: BLEU, METEOR, ROUGE and CIDEr.

Performance of LLMs without Fine-tuning

We report the performance of the selection tasks in Table 3, and the performance of the generation and span extraction tasks in Table 2. We can see that CoRECODE is a very challenging benchmark for all evaluated LLMs.

From Table 3, we observe that models which are instruction-tuned with SFT significantly outperform models being only pre-trained. The best-performing models across the three tasks are ChatGLM2-6B, BLOOMZ-7.1B, and Chinese-Alpaca-Plus-13B, respectively. Notably, on the slot identification task, Chinese-Alpaca-Plus-13B achieves an outstanding and unparalleled score.

On the commonsense knowledge generation task, BLOOMZ family achieves very high scores, as shown in Table 2. After checking the outputs of each model, we have found that models like ChatGLM and BELLE usually generate leading sentences or explanatory reasons in their responses, despite our prompt explicitly instructing them not to do so. In contrast, BLOOMZ-1.7B and BLOOMz-7.1B typically generate relatively short phrases as answers, which is consistent with the form of our annotations. They hence achieve higher scores than other evaluated LLMs.

To exclude the effect of answer form and answer length on the performance, we handed over the outputs of evaluated LLMs to ChatGPT for scoring, the average results of which are also reported in Table 2. We described the task

[3]All the experiments in the main paper were conducted on the HARD set. Experimental results on the EASY set are available in our arXiv version.

[4]https://github.com/THUDM/ChatGLM-6B

[5]https://github.com/THUDM/ChatGLM2-6B

[6]https://huggingface.co/fnlp/moss-moon-003-sft

[7]https://github.com/baichuan-inc/baichuan-7B

[8]https://huggingface.co/BelleGroup/BELLE-7B-0.2M

[9]https://huggingface.co/BelleGroup/BELLE-7B-2M

Model	Commonsense Knowledge Generation								CCPD	
	F1	EM	BLEU1	BLEU2	METEOR	ROUGE-L	CIDEr	ChatGPT Score	F1	EM
GLM-10B	0.023	0.000	0.000	0.000	0.032	0.000	0.001	3.190	0.011	0.000
BLOOM-7.1B	0.071	0.000	0.017	0.000	0.115	0.004	0.017	3.455	0.024	0.000
ChatGLM2-6B	0.160	0.004	0.001	0.000	0.145	0.001	0.002	3.940	0.029	0.001
BELLE-7B-0.2M	0.090	0.019	0.015	0.000	0.105	0.010	0.041	3.265	0.024	0.000
BELLE-7B-2M	0.111	0.008	0.004	0.000	0.140	0.003	0.010	3.555	0.007	0.000
BLOOMZ-1.7B	0.388	0.234	0.234	0.000	0.164	0.234	0.585	4.060	0.004	0.000
BLOOMZ-7.1B	0.438	0.284	0.282	0.000	0.199	0.283	0.707	3.980	0.041	0.003
BLOOMZ-7.1B-MT	0.435	0.300	0.300	0.000	0.184	0.300	0.750	4.030	0.047	0.010
MOSS-SFT-16B	0.199	0.071	0.066	0.000	0.147	0.049	0.174	3.780	0.038	0.001
Baichuan-7B	0.071	0.000	0.000	0.000	0.072	0.000	0.001	3.445	0.002	0.000
Chinese-Alpaca-Plus-7B	0.129	0.015	0.014	0.000	0.099	0.015	0.039	3.375	0.021	0.000
Chinese-Alpaca-Plus-13B	0.133	0.021	0.018	0.000	0.104	0.020	0.051	3.490	0.021	0.000
ChatGLM-6B	0.147	0.000	0.000	0.000	0.166	0.000	0.000	3.745	0.044	0.001
ChatGLM-6B 1-shot	0.202	0.061	0.035	0.000	0.154	0.048	0.120	3.770	0.038	0.002
ChatGLM-6B 3-shot	0.274	0.115	0.091	0.000	0.175	0.111	0.277	3.885	0.060	0.006
ChatGLM-6B 5-shot	0.215	0.097	0.095	0.000	0.147	0.096	0.240	3.685	0.052	0.007
ChatGPT	0.296	0.071	0.044	0.000	0.258	0.045	0.111	-	0.104	0.021

Table 2: Overall performance of evaluated LLMs on the commonsense knowledge generation and commonsense conflict phrase detection task. CCPD: Commonsense Conflict Phrase Detection.

Model	CKF	DI	SI
GLM-10B	0.157	0.060	0.051
BLOOM-7.1B	0.329	0.108	0.039
ChatGLM-6B	0.788	0.246	0.113
ChatGLM2-6B	0.818	0.286	0.153
BELLE-7B-0.2M	0.392	0.208	0.212
BELLE-7B-2M	0.599	0.169	0.109
BLOOMZ-1.7B	0.709	0.248	0.044
BLOOMZ-7.1B	0.758	0.444	0.165
BLOOMZ-7.1B-MT	0.695	0.341	0.168
MOSS-SFT-16B	0.445	0.353	0.110
Baichuan-7B	0.416	0.071	0.055
Chinese-Alpaca-Plus-7B	0.584	0.385	0.060
Chinese-Alpaca-Plus-13B	0.510	0.126	0.449
ChatGPT	0.896	0.275	0.084

Table 3: Overall performance of evaluated LLMs on the three selection tasks. CKF: Commonsense Knowledge Filling. DI: Domain Identification. SI: Slot Identification.

to ChatGPT and asked it to score the answers according to our pre-defined scoring criteria (see in our arXiv version). The average scores obtained by these models vary from 3 to 5. According to our criteria, this suggests that the answers generated by LLMs are more likely to be "answers that fit the context of the dialogue but are not a specific answer to the question" or "answers that are semantically inconsistent with the ground-truth answer but are also correct".

Table 2 also indicates improved model performance under the few-shot settings. However, the performance under the 5-shot setting is worse than that under the 3-shot setting. This might be due to the long length of our dialogues (as shown in Table 1, the average number of tokens per dialogue is 501). The excessive length of model inputs under the 5-shot setting might lead to a decline in performance.

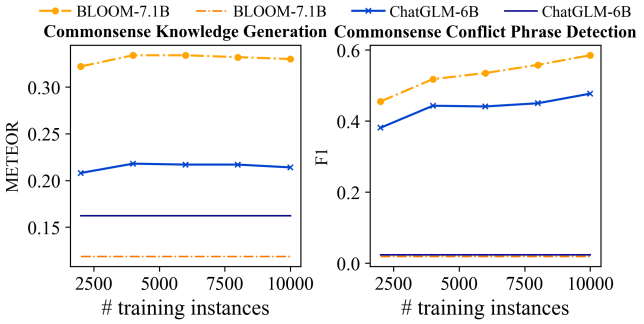

Figure 1: Performance of fine-tuned LLMs on the commonsense knowledge generation and commonsense conflict phrase detection task. The horizontal lines show the performance of LLMs without fine-tuning.

Performance of LLMs Being Fine-tuned

We further evaluated LLMs after they were fine-tuned on CORECODE. Specifically, we fine-tuned BLOOM-7.1B and ChatGLM-6B on 2K, 4K, 6K, 8K, and 10K examples respectively in the LoRA (Hu et al. 2022) manner, and tested these fine-tuned models on another 2K data.

Results on the commonsense knowledge generation and commonsense conflict phrase detection tasks are shown in Figure 1. Fine-tuning on different sizes of data results in large performance gains for both models. On the commonsense conflict phrase detection task, the F1 score rises as the size of training data increases. In contrast, on the commonsense knowledge generation task, the performance rises first and then falls as the number of training instances increases, indicating that approximately 4K training instances are sufficient for this task. Training with the same amount of training data for the same epochs on both tasks brings more performance gains for BLOOM-7.1B than for ChatGLM-

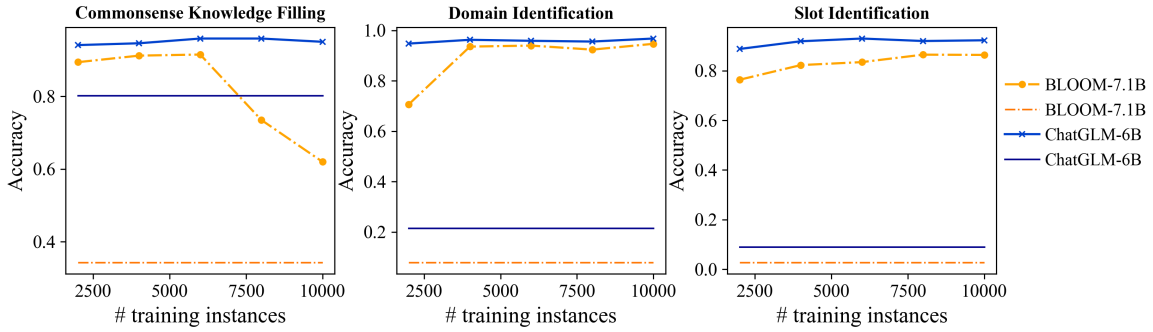

Figure 2: Results of fine-tuned LLMs on the three selection tasks. The horizontal lines show the performance of LLMs without fine-tuning.

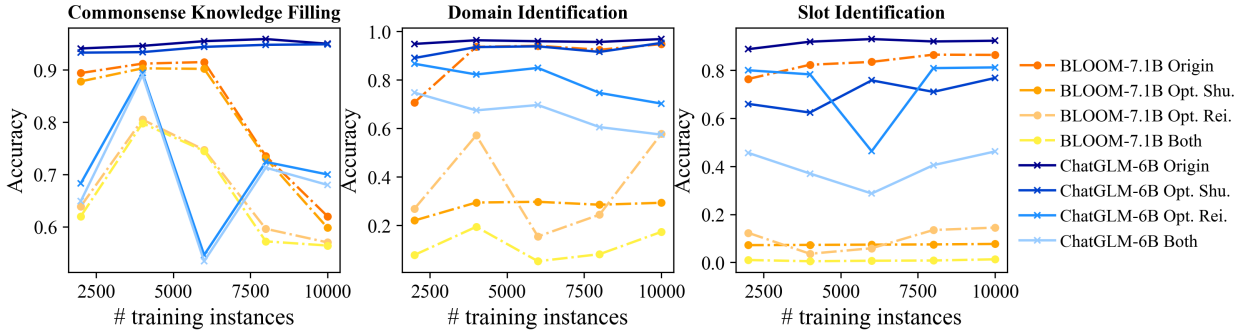

Figure 3: Results of fine-tuned LLMs on the perturbed test sets of the three selection tasks, by option re-indicating (Opt. Rei.), option shuffling (Opt. Shu.) and both.

6B. The reason could be that it is easier for BLOOM-7.1B without SFT to acquire such knowledge than ChatGLM-6B with SFT. For the three selection tasks, as shown in Figure 2, there is a positive correlation between model performance and training data size on most tasks. Both models obtain a substantial improvement after fine-tuning.

Robustness Analysis

Although fine-tuning on CORECODE significantly improves LLMs in commonsense reasoning, is the commonsense reasoning ability that LLMs obtained through fine-tuning robust? To investigate this question, we conducted three robustness tests on the three selection tasks: (1) option re-indicating, (2) option shuffling, and (3) both. For (1) option re-indicating, we change the option indicators from a, b, c to 1, 2, 3 in the process of forming the prompt. For (2) option shuffling, we shuffle the candidate options and then re-form the input prompt. For (3) both, we implement both option re-indicating and option shuffling.

Results in Figure 3 indicate decreased accuracy for both models. Generally, the two LLMs are especially sensitive to option re-indicating, demonstrating larger drops. However, they are more robust to option shuffling, maintaining relatively higher accuracy. The largest performance degradation occurs when both perturbations are executed.

Perturbation causes a dramatic drop to BLOOM-7.1B. In fine-tuning LLMs on CORECODE, we use option indicators, e.g., "b", as labels to be learned/predicted. ChatGLM-6B with SFT is better capable of understanding and following instructions than BLOOM-7.1B. It can align indicators to the corresponding answer options during training and combine them with task instructions to master the involved commonsense reasoning ability. BLOOM-7.1B, however, prefers to learn to answer by memorizing the corresponding input-label mappings and struggles to answer correctly after re-indicating and shuffling options. For instance, on the slot identification task, our training data has a large number of examples with the label "b". BLOOM-7.1B seems to learn such a shortcut incorrectly (i.e., mapping questions to label "b"). After shuffling answer options (the correct answer indicators are now mostly not "b"), the model still outputs plenty of "b", resulting in very low accuracies.

Conclusion

In this paper, we have presented CORECODE, a large-scale commonsense knowledge annotated dialogue dataset with over 76K annotations, and defined 6 benchmark tasks in the form of selection, extraction and generation, to assess the capability of LLMs in learning and applying commonsense knowledge. A diverse list of Chinese LLMs have been evaluated, which achieve poor performance on all tasks, demonstrating the difficulty and utility of the proposed dataset. We have further revealed the robustness issue of LLM commonsense knowledge acquisition via fine-tuning. We hope this work could be used to track and facilitate future advances in context-sensitive LLM commonsense reasoning.

Acknowledgments

The present research was supported by Zhejiang Lab (No. 2022KH0AB01). We would like to thank the anonymous reviewers for their insightful comments.

References

Bang, Y.; Cahyawijaya, S.; Lee, N.; Dai, W.; Su, D.; Wilie, B.; Lovenia, H.; Ji, Z.; Yu, T.; Chung, W.; et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Bhargava, P.; and Ng, V. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12317–12325.

Bian, N.; Han, X.; Sun, L.; Lin, H.; Lu, Y.; and He, B. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Bisk, Y.; Zellers, R.; Gao, J.; and Choi, Y. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7432–7439. ISBN 2374-3468. Issue: 05.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 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*, 5016–5026. Association for Computational Linguistics. EMNLP 2018.

Cambria, E.; Song, Y.; Wang, H.; and Hussain, A. 2011. Isanette: A Common and Common Sense Knowledge Base for Opinion Mining. *2011 IEEE 11th International Conference on Data Mining Workshops*, 315–322.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Cui, Y.; Yang, Z.; and Yao, X. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Du, N.; Huang, Y.; Dai, A. M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A. W.; and Firat, O. 2022a. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, 5547–5569. PMLR. ISBN 2640-3498.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022b. 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)*, 320–335. Dublin, Ireland: Association for Computational Linguistics.

Ghosal, D.; Hong, P.; Shen, S.; Majumder, N.; Mihalcea, R.; and Poria, S. 2021. CIDER: Commonsense Inference for Dialogue Explanation and Reasoning. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 301–313. Singapore and Online: Association for Computational Linguistics.

Ghosal, D.; Shen, S.; Majumder, N.; Mihalcea, R.; and Poria, S. 2022. CICERO: A Dataset for Contextualized Commonsense Inference in Dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5010–5028. Dublin, Ireland: Association for Computational Linguistics.

Guo, Z.; Jin, R.; Liu, C.; Huang, Y.; Shi, D.; Yu, L.; Liu, Y.; Li, J.; Xiong, B.; Xiong, D.; et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; and Clark, A. 2022. An empirical analysis of compute-optimal large language model training. 35: 30016–30030.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. (Comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6384–6392. ISBN 2374-3468. Issue: 7.

Khot, T.; Clark, P.; Guerquin, M.; Jansen, P.; and Sabharwal, A. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8082–8090. ISBN 2374-3468. Issue: 05.

Levesque, H.; Davis, E.; and Morgenstern, L. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Lin, B. Y.; Lee, S.; Khanna, R.; and Ren, X. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6862–6868. Online: Association for Computational Linguistics.

Liu, H.; and Singh, P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4): 211–226.

Liu, J.; Liu, A.; Lu, X.; Welleck, S.; West, P.; Le Bras, R.; Choi, Y.; and Hajishirzi, H. 2022. Generated Knowledge Prompting for Commonsense Reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3154–3169. Association for Computational Linguistics.

Lv, S.; Guo, D.; Xu, J.; Tang, D.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; and Hu, S. 2020. Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering. 34(5): 8449–8456.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2381–2391. Brussels, Belgium: Association for Computational Linguistics.

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T. L.; Bari, M. S.; Shen, S.; Yong, Z. X.; Schoelkopf, H.; Tang, X.; Radev, D.; Aji, A. F.; Almubarak, K.; Albanie, S.; Alyafeai, Z.; Webson, A.; Raff, E.; and Raffel, C. 2022. Crosslingual Generalization through Multitask Finetuning. *CoRR*, abs/2211.01786.

Qin, L.; Gupta, A.; Upadhyay, S.; He, L.; Choi, Y.; and Faruqui, M. 2021. TIMEDIAL: Temporal Commonsense Reasoning in Dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7066–7076. Online: Association for Computational Linguistics.

Quan, J.; Zhang, S.; Cao, Q.; Li, Z.; and Xiong, D. 2020. RiSAWOZ: A Large-Scale Multi-Domain Wizard-of-Oz Dataset with Rich Semantic Annotations for Task-Oriented Dialogue Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 930–940. Association for Computational Linguistics.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; and Young, S. 2021. Scaling language models: Methods, analysis & insights from training gopher.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3027–3035. ISBN 2374-3468. Issue: 01.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019b. Social IQa: Commonsense Reasoning about Social Interactions. 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)*, 4463–4473. Hong Kong, China: Association for Computational Linguistics.

Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; and Gallé, M. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. 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)*, 4149–4158. Minneapolis, Minnesota: Association for Computational Linguistics.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; and Azhar, F. 2023. Llama: Open and efficient foundation language models.

Wang, P.; Peng, N.; Ilievski, F.; Szekely, P.; and Ren, X. 2020. Connecting the Dots: A Knowledgeable Path Generator for Commonsense Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4129–4140. Association for Computational Linguistics.

Wang, X.; Li, C.; Zhao, J.; and Yu, D. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14006–14014.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. 35: 24824–24837.

West, P.; Bhagavatula, C.; Hessel, J.; Hwang, J.; Jiang, L.; Le Bras, R.; Lu, X.; Welleck, S.; and Choi, Y. 2022. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4602–4625. Seattle, United States: Association for Computational Linguistics.

Xu, X.; Gou, Z.; Wu, W.; Niu, Z.-Y.; Wu, H.; Wang, H.; and Wang, S. 2022. Long time no see! open-domain conversation with long-term persona memory. *arXiv preprint arXiv:2203.05797*.

Zhou, B.; Khashabi, D.; Ning, Q.; and Roth, D. 2019. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. 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)*, 3363–3369. Hong Kong, China: Association for Computational Linguistics.

Zhou, P.; Gopalakrishnan, K.; Hedayatnia, B.; Kim, S.; Pujara, J.; Ren, X.; Liu, Y.; and Hakkani-Tür, D. Z. 2021. Commonsense-Focused Dialogues for Response Generation: An Empirical Study. In *SIGDIAL Conferences*.

Zhou, X.; Zhang, Y.; Cui, L.; and Huang, D. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 9733–9740.

Zhou, Y.; Shen, T.; Geng, X.; Long, G.; and Jiang, D. 2022. ClarET: Pre-training a Correlation-Aware Context-To-Event Transformer for Event-Centric Generation and Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2559–2575. Dublin, Ireland: Association for Computational Linguistics.

Zhu, Q.; Huang, K.; Zhang, Z.; Zhu, X.; and Huang, M. 2020. CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset. *Transactions of the Association for Computational Linguistics*, 8: 281–295.