# XDAI: A Tuning-free Framework for Exploiting Pre-trained Language Models in Knowledge Grounded Dialogue Generation

Jifan Yu*
yujf21@mails.tsinghua.edu.cn
Tsinghua Univerisity
Beijing, China

Xiaohan Zhang*
zhang-xh19@mails.tsinghua.edu.cn
Tsinghua Univerisity & Zhipu.AI
Beijing, China

Yifan Xu*
xuyf@bupt.edu.cn
Tsinghua Univerisity
Beijing, China

Xuanyu Lei
leixy20@mails.tsinghua.edu.cn
Tsinghua Univerisity
Beijing, China

Xinyu Guan
xinyu.guan@biendata.com
Biendata
Beijing, China

Jing Zhang
xinyu.guan@biendata.com
School of Information
Renmin University of China
Beijing, China

Lei Hou
houlei@tsinghua.edu.cn
Department of Computer Science and Technology, BNRist & KIRC, Institute for Artificial Intelligence
Tsinghua Univerisity
Beijing, China

Juanzi Li†
lijuanzi@tsinghua.edu.cn
Department of Computer Science and Technology, BNRist & KIRC, Institute for Artificial Intelligence
Tsinghua Univerisity
Beijing, China

Jie Tang
jietang@tsinghua.edu.cn
Department of Computer Science and Technology, BNRist & KIRC, Institute for Artificial Intelligence
Tsinghua Univerisity
Beijing, China

## ABSTRACT

Large-scale pre-trained language models (PLMs) have shown promising advances on various downstream tasks, among which dialogue is one of the most concerned. However, there remain challenges for individual developers to create a knowledge-grounded dialogue system upon such big models because of the expensive cost of collecting the knowledge resources for supporting the system as well as tuning these large models for the task. To tackle these obstacles, we propose XDAI, a knowledge-grounded dialogue system that is equipped with the prompt-aware tuning-free PLM exploitation and supported by the ready-to-use open-domain external knowledge resources plus the easy-to-change domain-specific mechanism. With XDAI, the developers can leverage the PLMs without any fine-tuning cost to quickly create the open-domain dialogue systems as well as easily customize their own domain-specific systems. Extensive experiments including human evaluation, Turing test, and online evaluation have demonstrated the competitive performance of XDAI compared with the state-of-the-art general PLMs and specific PLMs for dialogue. XDAI pilots studies on the exploitation of PLMs and made intriguing findings which could be inspiring for the future research on other PLM-based applications.

*Equal Contribution.
†Corresponding author.


*KDD'22, August 14-18, 2022, Washington DC, USA*

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539135

Developers and related researchers can get access to our repository at https://github.com/THUDM/XDAI, which presents a series of APIs, incremental toolkits and chatbot service of XDAI platform.

## CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Natural language generation**; • **Software and its engineering** → **Development frameworks and environments**.

## KEYWORDS

Pre-trained Model Exploitation, Dialogue System

**ACM Reference Format:**
Jifan Yu, Xiaohan Zhang, Yifan Xu, Xuanyu Lei, Xinyu Guan, Jing Zhang, Lei Hou, Juanzi Li, and Jie Tang. 2022. XDAI: A Tuning-free Framework for Exploiting Pre-trained Language Models in Knowledge Grounded Dialogue Generation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3534678.3539135

## 1 INTRODUCTION

*Foundation model is the subject of a paradigm shift*, claimed by a group of experienced scientists recently [2], suggesting that various AI-driven systems in the future tend to directly build upon or heavily integrate large-scale pre-trained language models (PLMs) such as GPT [3], BERT [6], and T5 [22], because such big models have already demonstrated incredible advances on a large number of natural language processing (NLP) tasks such as question answering [12], machine reading comprehension [27], textual entailment [15], causal reasoning [14], etc. Among these applications

*Equal Contribution.
†Corresponding author.

of PLMs, dialogue, which targets to enable the real communications between human beings and the machine, can extensively benefit people and society, thus has drawn wide attention.

The recent models built for dialogue have shown great advances when injected with the relevant knowledge to the concerned dialogue. Such study, also known as *Knowledge-grounded Dialogue Generation*, rapidly attracts significant attention from academia and spawns the invention of multiple datasets [7, 17, 29] and powerful algorithms [1, 20, 32]. Despite the abundant research achievements, it is still far from convenient for developers to build their own dialogue systems due to the following main challenges:

- **Obstacles to High-quality Data Curation**. Unlike the well-prepared experimental environment in research studies, to build a real dialogue system, we need to collect the external knowledge resources for supporting the system and craft the domain-specific dialogue corpus for training the models from scratch, causing the difficulty of directly applying most advanced models. In addition, the external knowledge resources are usually presented in heterogeneous forms such as knowledge bases, textual documents, and structural tables, which significantly raises the design difficulty of the dialogue system.
- **Trade-off between Effectiveness and Efficiency**. Intuitively, conducting the more training, tuning, or constraining, the better the dialogue system performs. However, such operations rely heavily on adequate technical experience and high-quality data, as well as consume large computational resources and could even decelerate the inference efficiency of the model (e.g., inverse prompt [37]). Thus, making a trade-off between effectiveness and efficiency is crucial for building a real dialogue system.

**Presented Work.** We release a tuning-free framework for e**X**ploiting large-scale pre-trained language models in building knowledge-grounded **D**ialogue **AI** systems (**XDAI**), which supports a thorough workflow including **offline knowledge resource curation** and **online dialogue generation**. XDAI is designed to offer the services for developers with its unique features: (1) *Quick Start*: We provide an open-domain knowledge-grounded dialogue service with sufficient ready-to-use open-domain knowledge resources. Developers can easily deploy a dialogue system with this basic service. (2) *Efficient Inference*: XDAI designs a novel prompt pattern with knowledge injection, which optimizes the generated dialogues from PLMs without further training or tuning. (3) *Customized Deployment*: To build a domain-specific dialogue system, XDAI provides easy-to-change plugins to enable automatically searching and updating the external knowledge only from a few domain-specific seeds. (4) *Incremental Modification*: XDAI provides a series of toolkits for incremental developing, encouraging developers to refine and customize their personalized components.

**Impact & Beneficial Groups.** For the applications of PLMs, XDAI calls for the explorations of tuning-free model inference and indicates that such paradigm still has potential [37]. For the applications of knowledge-grounded dialogue generation, the practical results of XDAI unfold several interesting findings like discrete prompts are more likely to produce high-quality dialogues, and general language models show advantages in cross-domain dialogue generation.

XDAI can have a more positive impact on the industry. We expect that our framework and toolkits can provide more opportunities for the (1) developers with limited resources, (2) practitioners from other fields, and (3) newcomers to the machine learning domain to easily and quickly use PLMs to accomplish their creative ideas.

In the following sections, we first introduce the problem definition and the background techniques for building XDAI in Section 2 and then present the technical implementation with several illustration examples in Section 3. Finally, we conduct extensive online experiments to evaluate the effectiveness and efficiency performance of XDAI and discuss the insights for employing PLMs in real-world dialogue products in Section 4.

## 2 PRELIMINARIES

This section first formulates the problem and then explore the background techniques for building the dialogue system.

### 2.1 Problem Formulation

*Definition 2.1.* **Dialogue History** is a set of conversational utterances between two speakers, which can be formally denoted as $\mathcal{D}_t = \{U_1, S_1, ..., U_{t-1}, S_{t-1}, U_t\}$, where $U_i$ and $S_i$ are sentences made of word, belonging to the user and the dialogue system respectively. Specially, $U_t$ from the user is also called the $t$-th round *Query*. A piece of dialogue history usually involves a subject, named as the dialogue topic.

*Definition 2.2.* **External Knowledge Pool** contains multiple pieces of information associated with the dialogue topics in the system, which is denoted as $\mathcal{K} = \{k_i\}_{i=1}^{|\mathcal{K}|}$, where $k_i$ is a piece of knowledge information. These knowledge pieces could be collected and organized from the external heterogeneous web sources [10] such as the knowledge bases [32], Wikipedia pages [7], persona descriptions [29], and emotional contexts [18].

Problem 1. ***Knowledge-grounded Dialogue Generation task:*** *Given the dialogue history $\mathcal{D}_t$, the target of task is to generate a response $S_t$ for the $t$-th round query $U_t$ with the help of the external knowledge pool $\mathcal{K}$.*

### 2.2 Background Techniques

**Pre-trained Model Exploitation**. Since PLMs have significantly improved the performance of various NLP tasks [22], researchers have been exploring the PLMs-based algorithms as the groundwork for the next generation of artificial intelligence, where how to effectively and efficiently unfold the capabilities of PLMs becomes a crucial challenge [2]. Pre-training plus fine-tuning [24] that further trains the PLMs to adapt to the downstream tasks, is one of the most widely-adopted paradigms. However, with the rapid growth of the model scale, these large-scale PLMs require a large amount of high-quality corpus as well as expensive computational resources for the single fine-tuning, making them difficult to use in practice. To address this, the emergent techniques such as prompt tuning [18, 33] and adapter [25] have been invented to fine-tune only a few instead of the whole parameters, expecting to achieve the comparable performance with the fully fine-tuned mode. In addition, considering the feasibility of the PLMs, researchers also explore the approaches such as prompt searching or controlled generation [37] to guarantee the quality of the generation results.

**Knowledge-grounded Dialogue Generation**. The task aims to generate more real dialogues with the help of various external knowledge [10, 11, 35, 36] such as knowledge graphs, persona descriptions, and emotional strategies, which becomes a rising research topic. In addition to the advanced algorithms such as self-training and adapter [1, 20, 32] that perform excellently in existing dialogue benchmarks [7, 17, 29], some researchers surprisingly find that the general PLMs such as BERT [6] and GPT [3]), and the specific PLMs for dialogue generation such as PLATO-XL [1] and EVA [34], can achieve competitive performance with more efficient solutions [32]. These observations inspire us to explore the prompt engineering in PLMs for dialogue generation, which can potentially facilitate the development of PLMs in dialogue systems.

# 3 XDAI

In this section, we first introduce the overview framework of the proposed XDAI and then present the detailed workflow of building and applying XDAI for a real dialogue system.

## 3.1 Overview Framework

Figure 1 shows the overview framework of XDAI, which consists of two subsystems: offline knowledge curation system and online dialogue generation system. The two systems run independently, but their data constantly interact to achieve optimal dialogue results. Developers can either choose the basic service module for building open-domain dialogue systems or flexibly leverage the offered functions to create domain-specific dialogues system. The workflow of the two subsystems of XDAI is explained as below.

(1) *Offline Knowledge Curation System*: This subsystem is responsible for integrating, storing, and providing data, including historical logs generated by online dialogue systems, as well as heterogeneous resources from search engines and external knowledge bases. It consists of several web information acquisition tools, data processing methods, and databases. This subsystem also offers optional development components like the concept expansion toolkit which can help produce new knowledge concepts from only a few seeds. These optional components serve the developers to create the domain-specific knowledge pool for building their personalized dialogue systems.

(2) *Online Dialogue Generation System*: This subsystem aims to generate knowledge-grounded responses based on the external knowledge pool and the current session's dialogue contexts. For this purpose, we design a novel prompt manufacturing mechanism (as shown in Figure 2), which skillfully combine the background knowledge, the knowledge-related QA pairs, and the dialogue history associated with current query as PLM's input, to improve the dialogue generation quality without training.

In the next sections, we take several Chinese implementation examples to explain the technical details of XDAI. Specifically, we introduce the processing components of building XDAI's basic open-domain dialogue service (named as XDAI-open) and domain-specific dialogue service (named as XDAI-domain). The latter's implementations are related to the topics "Travel" and "Sports", each of which are prepared with a set of 20 topical keywords as the seed concepts $\mathcal{K}^s$ for collecting the external knowledge resources. We mainly employ the Chinese-versioned GLM with 10B parameters [9] as the base PLM, which is trained on 302GB raw Chinese data collected from multiple Chinese websites.

## 3.2 Offline System: Knowledge Curation

Knowledge Curation system provides the functions of data collection and resource integration for XDAI-open. It also offers an optional concept expansion function for discovering abundant domain-specific knowledge pieces from a few seeds for XDAI-domain.

*3.2.1 Data collection.* The function of data collection is for collecting the relevant knowledge resources to the seed concepts $\mathcal{K}^s$ from the whole external knowledge resources. For XDAI-domain, we select 20 concepts related to the concerned domain from the concepts in Xlore2 [13], one of the largest Chinese knowledge bases, as the seeds, but for XDAI-open, since no domain is specified, we employ all the concepts in Xlore2 as its seeds. Besides this semi-structured knowledge base, we also employ the unstructured snippets from the Bing search engine [1] to improve the knowledge coverage. Then based on the seed concepts, we collect the relevant knowledge resources from each kind of the sources.

For data collecting from the knowledge base, we leverage the official data acquisition toolkit released by Xlore2 to collect the knowledge resources relevant to the seed concepts. Since each seed concept can be linked to the knowledge base, we not only extract their structural (head entity, relation, tail entity) triples but also preserve the concepts' unstructured abstracts and descriptions.

For data collecting from the search engines, we search the relevant articles and immediate news by issuing the seed concepts as keywords. To cover more resources, apart from querying a single concept, we also select up to three concepts and combine them as a single query for searching. The semi-structured snippets in HTML format are then processed into plain text format with the corresponding concepts.

*3.2.2 Resource Integration.* The function of resource integration consists of data processing and storage, which is shared between XDAI-open and XDAI-domain. Since the external knowledge resources are in heterogeneous forms such as the knowledge triples, concept explanations, and snippet texts, we transform them into two unified forms, *Description-formatted* and *QA-formatted*, to serve as a general solution to the the dialogue system.

*Description-formatted* resources are mainly extracted from the concept's abstraction and description texts, which are acted as the background knowledge in the subsequent prompt manufacturing component.

*QA-formatted* resources are mainly extracted from the structured triples in the knowledge base and the snippet texts provided by the search engine. To create the QA-formatted resource for a triplet, we define several question templates to generate the question for the tail entity as the answer. While for a snippet text, we extract the sentence that contain the queried seed concepts as the answer and adopt a T5-based question generation tool [4] to generate the corresponding question. These QA-formatted resources are mainly used for knowledge injection in dialogue contexts.

[1] https://www.bing.com

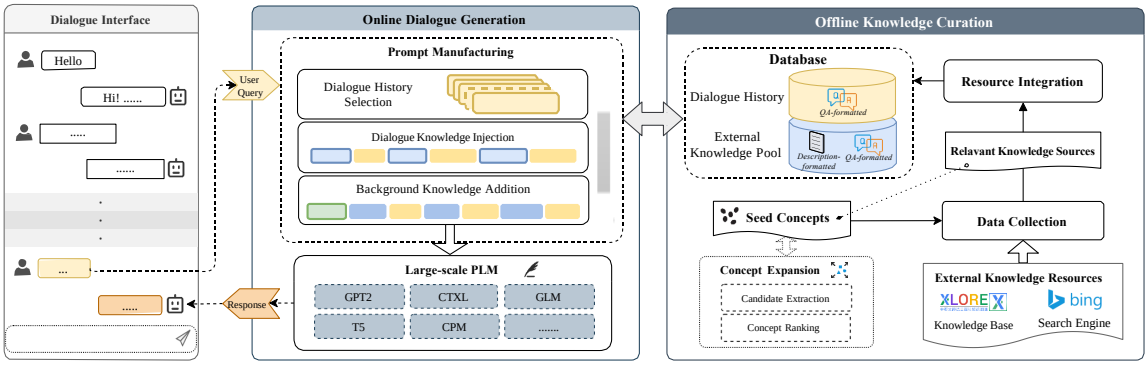


**Figure 1: The architecture of the XDAI. Two subsystems of offline knowledge curation and online dialogue generation work independently with data interaction. Developers can also collect more knowledge resources based on provided toolkits.**

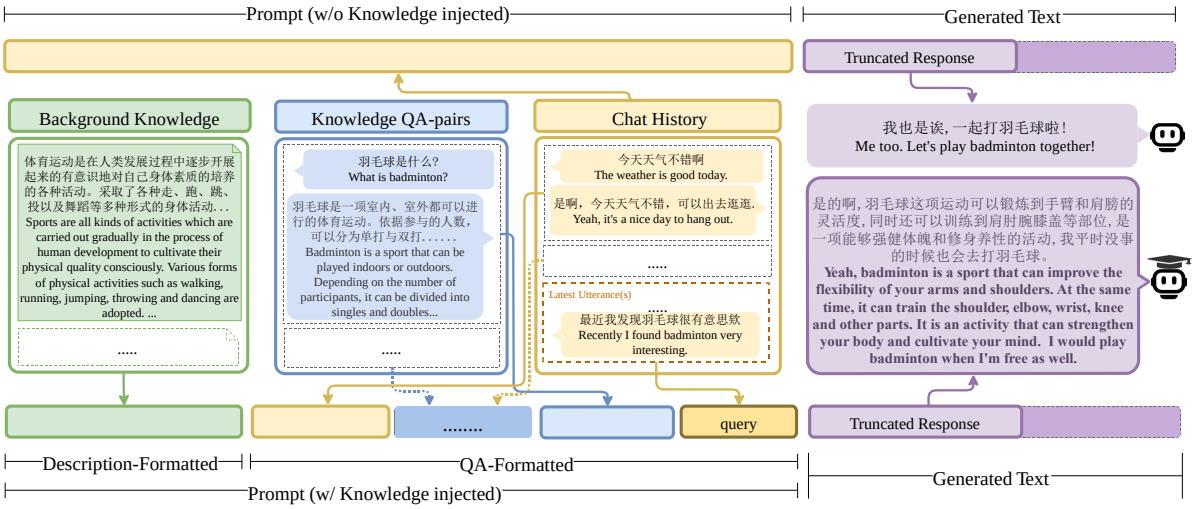


**Figure 2: An example of our prompt manufacturing result, which is about the topic of "badminton".**

All the types of knowledge resources are indexed by concepts, encoded by SentenceTransformer [23], and stored in the external knowledge pool.

*3.2.3 Concept Expansion (Optional).* Although the general knowledge base is sufficient to provide external knowledge in most scenarios, many domain-specific dialogue systems still require a specific collection of external knowledge resources, as the 20 seed concepts of XDAI-domain are clearly not enough. For extending the seed concepts, we provide the optional function of concept expansion, which performs **candidate extraction** for recalling the relevant concepts as much as possible and then **concept ranking** for selecting the most relevant ones. For the implementation, we simply extract all the noun phrases from the seeds' relevant knowledge resources as the concept candidates and then employ a cluster-based unsupervised method [28] to guarantee the extraction quality. Specifically, we first cluster all the candidate concepts into 15 clusters by their BERT embeddings with K-means and calculate the similarity $f(C_j, K^s)$ of each cluster $C_j$ to the seed set $K^s$, which is formulated as:

$$f(C_j, K^s) = \max_{s_k \in K^s} (\text{cosine}(\Sigma_{i=1}^{|C_j|} c_i / |C_j|, s_k)). \quad (1)$$

The intuition is a cluster is similar to the seed set if all the concepts in it is similar to any seed concept $s_k$. Then we preserve the top-2 clusters of candidates as the expanded concepts. Such expansion process can be further performed iteratively for more concepts. Finally, we respectively acquire 540 and 261 concepts for the "Travel" and "Sports" domains for further resource collection.

## 3.3 Online System: Dialogue Generation

Dialogue generation is the core component of both XDAI-open and XDAI-domain, which generates high-quality knowledge-grounded responses based on the offline created external knowledge pool and the dialogue history. The main idea is to exploit the PLM by manufacturing a novel knowledge-injected prompt, and then querying the tuning-free PLMs. Such prompt manufacturing is accomplished

in three stages: Dialogue History Selection, Dialogue Knowledge Injection, and Background Knowledge Addition. Figure 2 shows an example of the prompt manufacturing result about "badminton".

*3.3.1 Dialogue History Selection.* Since the input length of the PLM is limited, given the current query, we first search the most relevant historical utterances by calculating the cosine similarity between the query and each utterance pair $(U_i, S_i)$:

$$\alpha^{t-i} * \text{cosine}\left([U_i; S_i], U_t\right), \tag{2}$$

where $[U_i; S_i]$ indicates the concatenation of the user utterance $U_i$ and the system utterance $S_i$. We employ the SentenceTransformer [23] to encode both $U_i$ and $S_i$. To further ensure the coherence of the dialogue content, we assign a decay factor $\alpha^{t-i}$ with $\alpha \in (0, 1)$ to emphasize the most recent dialogue rounds.

*3.3.2 Dialogue Knowledge Injection.* After selecting dialogue contexts, we invoke the QA-formatted knowledge resources to enrich the dialogue, which is called dialogue knowledge injection. Specifically, for each utterance $U_i$ or $S_i$ in the selected dialogue history, we extract the concepts contained in it and retrieve their corresponding knowledge resources in the offline indexed external knowledge pool. The concrete implementation is resort to the fuzzyMatch function of Elastic Search [8][2]. Then the most similar QA-formatted knowledge to $[U_i; S_i]$ is selected as the prompt to be injected before its corresponding dialogue utterance.

*3.3.3 Background Knowledge Addition.* Apart from the knowledge injection to the historical dialogues, we also prefix these dialogues with *description-formatted* resources as background knowledge of the entire conversation. Specifically, we first select the most similar concept from the offline collected concepts to all the concepts occurred in the injected knowledge of the previous two stages, and then extract its corresponding description-formatted knowledge resources as the background knowledge.

Finally, the prompts consisting of the dialogue history, the dialogue knowledge, and the background knowledge are injected into the PLM for generating knowledge-grounded responses.

## 3.4 Availability

We provide a repository at https://github.com/THUDM/XDAI, with an easy-to-use XDAI toolkit, which aims to assist researchers in deploying their own chatbot by exploiting the power of PLMs. The basic toolkit is composed of the following modules, with which researchers can conveniently build an open-domain dialogue online service (XDAI-open):

(1) *SessionManager* manages the chat sessions, with or without the connection to a database for the storage of chat history logs. It stores and recalls personalized information in case the dialogue system is required to support multi-users.

(2) *ChatbotAgent* deals with the prompt manufacturing process and post-process of the raw text generated by the PLM. The PLM is called by the agent via API, so that it can be replaced by any alternative, which ensures decoupling and flexibility. Developers can implement their own prompt manufacturing by revising the functions in the agent.

[2]https://github.com/elastic/elasticsearch

**Table 1: Statistics of human evaluation protocol, where Avg-length is the average length of the generated utterance, and Label is the total number of collected scores of dialogues.**

| Task | Participants | Avg-length | Label |
|---|---|---|---|
| Open Dialogue | 20 | 10.63 | 75000 |
| Domain Dialogue | 10 | 5.61 | 6500 |

For deployment, we also provide API & database interface, with which developers can conveniently connect the system to an instant message app such as Wechat or submit their own domain seeds to collect domain-specific knowledge resources for building XDAI-domain. Beyond these pre-developed functions, XDAI welcomes developers to personalize and refine several existing methods, including but not limited to sensitive word checking, controlling the level of politeness, and QA-format converting.

# 4 EXPERIMENT

In this section, we design experiments to analyze the characteristics of XDAI. For method performance, we first conduct a series of human evaluations to estimate the performance of XDAI-open as well as our implemented XDAI-domain in specific domains. We then conduct an ablation study to analyze the role of each component and design a Turing test to double-check the naturalness of the generated dialogues. For system capabilities, we additionally present an analysis of the online testing of the XDAI-open version, including response time, user involvement, and pick interesting cases to discuss possibilities for future research.

## 4.1 Experimental Setting

**Human Evaluation Protocol.** As suggested in the previous empirical study [16], there is a significant gap between automatic metrics and human judgments in dialogue generation. We conduct a human evaluation on two conversation scenarios: open-domain dialogue generation (for XDAI-open) and domain-specific dialogue generation (for XDAI-domain). Therefore, we mainly employ manual annotation for evaluation in the experiments of conversation. For open-domain dialogue generation, we recruit 20 people, mostly university students, to generate and evaluate the quality of conversation. For each baseline, we generate chatting sessions and ensure that each session contains at least ten rounds of valid conversation. For domain-specific dialogue, we select "Travel" and "Sports" as two topics for building a dialogue system, invite ten domain experts and experienced volunteers, and select popular questions in Zhihu as initial topics to control the scope of conversations. Table 1 presents statistics of annotators and dialogue evaluation data.

**Evaluation Metrics.** We follow up previous dialogue generation efforts [1, 37] and employ metrics to evaluate 5 aspects of the dialogue quality. For utterance-level evaluation, we employ: (1) **Coherence** is to measure whether the response is consistent and relevant with the dialogue historical context. (2) **Informativeness** is to evaluate whether the generated response is informative or not. (3) **Inconsistency**↓ is a fine-grained metric for coherence evaluation, considering whether the response contradicts the given context. (4) **Hallucination**↓ is a fine-grained metric for informativeness

**Table 2: Human evaluation results of open-domain dialogue generation. As Inverse Prompt method is only designed for single-round dialogue, we cancel the evaluation of its conversation-level performance.**

| Category | Method | Open-Domain Dialogue | | | | |
|---|---|---|---|---|---|---|
| | | Coherence | Inconsisitency↓ | Informativeness | Hallucination↓ | Engagingness(C) |
| General Language Model | GLM | 1.267 | 0.071 | 1.163 | 0.180 | 0.767 |
| | Transfromer-XL | 1.158 | 0.119 | 1.096 | 0.193 | 0.750 |
| | CPM | 1.273 | 0.076 | 1.192 | 0.146 | 0.767 |
| Pretrained Dialogue Model | CDial-GPT | 0.913 | 0.158 | 0.746 | 0.270 | 0.543 |
| | DialoGPT | 0.857 | 0.175 | 0.772 | 0.293 | 0.556 |
| | EVA | 1.172 | 0.127 | 1.288 | 0.340 | 0.703 |
| | PLATO-XL | **1.534** | **0.048** | 1.297 | **0.072** | **1.256** |
| Controllable Generation | FSB | 0.784 | 0.100 | 0.833 | 0.298 | 0.383 |
| | Inverse Prompt | 1.408 | 0.238 | 1.492 | 0.225 | / |
| | XDAI | **1.512** | **0.054** | **1.658** | **0.162** | **1.125** |

evaluation, checking whether the response express any factual errors. For dialogue-level evaluation, we employ (5) **Engagingness** to evaluate whether the user would like to talk with the chatbot for a longer conversation.

Note that the Coherence, Informativeness and Engagingness scale is [0,1,2], whose higher score indicates a better performance. Moreover, the scale of Inconsistency and Hallucination is [0,1], whose lower score indicates a better performance.

**Baselines.** To evaluate the performance of our proposed model, we reproduce and deploy baseline methods of exploiting large models for comparison of three categories:

*General Language Models*: Since GPT-3 [3] achieved impressive performance on zero-shot dialogue, exploiting ultra-large language models becomes a basic solution. We deploy the following general language models for direct inference:

• Transformer-XL [5] is an improved self-attention network that introduces the notion of recurrence. We employ its 2.9B-parameter model that obtains the best performance of all Chinese versions.

• CPM-2 [31] is a GPT architecture language model pretraining on the Chinese Corpus, which provides 11 and 198B(MoE) parameter models. We employ its 11B version dense model.

• GLM [9] employs autoregressive blank-filling for jointly learning both the bidirectional and unidirectional attention mechanism, achieving a competitive performance of dialogue application.

*Pretrained Dialogue Models*: Besides general ultra-large language models, researchers also make efforts in building dialogue-oriented language models with specific training loss.

• CDial-GPT [26] is trained based on LCCC conversations, which is a Chinese GPT-architecture model with 95.5M parameters.

• ProphetNet-X [21] is a family of pretrained models, while we select the Chinese dialogue generation version that is trained on social media conversation of Douban with 379M parameters.

• DialoGPT [30] is a fine-tuned GPT-2 with Reddit comment data. We select the 345M (best performance) for comparison.

• EVA [34] is a 2.8B parameter Chinese dialogue generation model which is trained with the WudaoCorpura-Dialogue dataset that includes 1.4B conversation social media dialogue samples.

• PLATO-XL [1] is a recent 11B parameter model for dialogue generation, which is a competitive method both in research experiments and commercial scenarios.

*Controllable Generation*: Meanwhile, there are also efforts for exploiting language in dialogue generation at low cost, since the training or fully fine-tuning requires extensive data and computing resources. Some technical lines include learning continuous prompts and searching for optimal discrete prompts.

• FSB [32] is an implementation of parameter-efficient tuning on dialogue generation task, which learns continuous tokens as "prompts" for querying language model.

• Inverse Prompt [37] provides a self-optimization strategy that we can calculate the probability of results by inversely taking generated answers as input to infer the original question.

For comparison analysis, in open domain dialogue scenarios, we directly use their publicly available versions for dialogue generation; in domain-specific scenarios, we use our collected relevant corpus for fine-tuning or booting, but it is still difficult to guarantee the absence of bias, so these results are not intended as a comparison of strengths and weaknesses, but rather to reflect some trends. In subsequent experimental results, the optimal values are underlined and the highlights are in bold.

## 4.2 Result Analysis

*4.2.1 Results of Open-domain Dialogue.* We first analyze the human evaluation results for each type of model in the open domain dialogue scenario. The results of all 75,000 evaluations are summarized in Table 6. The results show the following trends:

First, XDAI surprisingly achieves competitive results without training or fine-tuning on the conversational corpus. No existing approach can outperform XDAI in all metrics exhaustively, suggesting that it is promising and potential to design discrete prompts to exploit general language models for dialogue tasks.

Second, compared to large-scale dialogue models such as Plato-XL, XDAI is better at *Informativeness*, which reflects the effectiveness of the knowledge injection mechanism. Moreover, general model-based approaches (GLM, Transformer-XL, CPM, and Inverse Prompt) generally perform well on this metric, confirming the advantages of general-purpose language models for knowledge tasks as expressed by empirical analysis.

Third, the further enhancement direction of XDAI should be hallucination. Although the overall dialogue is attractive (having

**Table 3: Human evaluation results of domain-specific dialogue generation, where Cohe., Inco.↓, Info., Hall↓., and Enga.(C), are the abbreviations corresponding to *Coherence, Inconsistency, Informativeness, Hallucination* and *Engagingness*.**

| Category | Method | Travel Domain | | | | | Sports Domain | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Cohe. | Inco.↓ | Info. | Hall.↓ | Enga.(C) | Cohe. | Inco.↓ | Info. | Hall.↓ | Enga.(C) |
| General Language Model | GLM | 1.590 | 0.028 | 1.457 | 0.057 | 1.057 | 1.099 | 0.099 | 1.001 | 0.253 | 0.624 |
| | Transfromer-XL | 1.293 | 0.240 | 1.413 | 0.293 | 1.120 | 0.983 | 0.192 | 0.961 | 0.292 | 0.701 |
| | CPM | 1.228 | 0.114 | 1.100 | 0.142 | 0.928 | 1.235 | 0.133 | 1.219 | 0.171 | 0.748 |
| Pretrained Dialogue Model | CDial-GPT | 1.384 | 0.078 | 0.984 | 0.107 | 0.538 | 0.611 | **0.082** | 0.447 | 0.164 | 0.270 |
| | DialoGPT | 1.164 | 0.095 | 1.027 | 0.191 | 0.630 | 0.983 | 0.216 | 1.033 | 0.403 | 0.667 |
| | EVA | 1.229 | 0.062 | 1.313 | 0.229 | 0.708 | 0.778 | 0.259 | 1.056 | 0.389 | 0.574 |
| | PLATO-XL | 1.625 | 0.069 | 1.458 | **0.041** | **1.460** | 1.447 | 0.236 | 1.237 | 0.158 | 0.833 |
| Controllable Generation | FSB | 0.944 | 0.083 | 1.056 | 0.250 | 0.472 | 0.745 | 0.160 | 0.849 | 0.292 | 0.490 |
| | Inverse Prompt | 1.250 | 0.208 | 1.292 | 0.208 | / | **1.684** | 0.157 | **1.790** | **0.105** | / |
| | XDAI | **1.660** | **0.020** | **1.770** | **0.060** | **1.430** | **1.378** | **0.179** | **1.410** | **0.215** | **0.936** |

the second-highest Engagingness), we suggest that some restrictions can be appropriately designed to prevent content deviations in practical usage.

*4.2.2 Results of Domain-specific Dialogue.* We further analyze the results in domain-specific scenarios to estimate the performance of XDAI with the offline knowledge exploration system. In general, the domain-specific performance of the individual models fluctuates significantly from the open-domain dialogue. However, XDAI maintains a competitive performance, especially leading in several metrics under the Travel topic. We conduct an analysis of these results in Table 3 and conclude with the following observations.

(1) Although the overall conversational engagingness of PLATO-XL remains high and the hallucinations are maintained at a superior level, its informativeness is severely lacking, and its coherence has dropped slightly. XDAI, by using the general language model as basis, still maintains an advantage in these dimensions.

(2) The performance of the models varies significantly in different domains, but the controllable generation-related methods are more stable due to more predefined constraints. Only dialogue models with a very large parameter scale can achieve relatively stable performance among the pre-trained methods. XDAI still maintains a good performance with its automatically explored knowledge.

(3) Methods that use few-shot learning or continuous prompts (e.g., FSB and Inverse Prompt) remain unimpressive, while the best performance is instead achieved by methods that use discrete prompts for large model queries (e.g., GLM and PLATO-XL) or methods that conduct high-quality search upon them (e.g., Inverse prompt). XDAI's performance in Domain remains relatively competitive.

*4.2.3 Ablation Study.* We perform an ablation study of XDAI-open by removing specific knowledge resources in prompt manufacturing and exploring its performance differences with two general language models as bases. The results are presented in Table 4.

(1) We compare Transformer-XL and GLM as the base models for open domain dialogues, respectively. The results show that GLM is more suitable for guiding dialogue generation with XDAI framework, with a slight leading in all metrics.

(2) Both two formats of knowledge are necessary. As evidenced by the decline in model performance after removing these two types of knowledge resources, the injection of background knowledge and QA-formatted knowledge is essential. Among them, the QA-format knowledge is more helpful in improving the overall conversation quality, which is proven by the significant decline of Coherence, Informativeness, and Engagingness.

**Table 4: Effects on XDAI performance when employing different prompt settings and base models, where Cohe., Inco.↓, Info., Hall.↓, and Enga.(C), are respectively the abbreviations of *Coherence, Inconsistency, Informativeness, Hallucination* and *Engagingness*. CTXL refers to Transformer-XL model.**

| Base | Version | Cohe. | Inco.↓ | Info. | Hall.↓ | Enga.(C) |
|---|---|---|---|---|---|---|
| GLM | Full | 1.512 | 0.162 | 1.658 | 0.054 | 1.125 |
| | w/o QA. | **-0.245** | +0.018 | **-0.495** | +0.016 | **-0.357** |
| | w/o Backg. | -0.209 | **+0.089** | -0.200 | **+0.061** | -0.221 |
| CTXL | Full | 1.469 | 0.260 | 1.509 | 0.061 | 1.045 |
| | w/o QA. | **-0.311** | -0.067 | **-0.412** | **+0.058** | **-0.296** |
| | w/o Backg. | -0.010 | +0.005 | -0.127 | +0.023 | -0.254 |

(3) The background description-format knowledge surprisingly alleviates Inconsistency and Hallucination. Although QA-formatted knowledge brings more information, it may cause more inconsistency in the whole dialogue (Under CTXL setting, the inconsistency even slightly reduces after removing it). Moreover, the background knowledge can effectively control the main topic of the conversation, preventing the instability of the dialogue.

## 4.3 Turing Test

Besides direct human evaluations, we further design Turing test [19, 37] experiments to double-check the actual performance of each method, which is regarded as one of the ultimate goals in AI.

**Setup.** We conduct the Turing test experiments on our specially developed game based on the Wudao Developer Platform[3]. The Turing test game is set up to simultaneously present the annotator with two conversations. One is a real conversation, and the other is a machine-generated response on the same topic. Finally, we collected 2200 scoring records of such binary selection from 20

[3]https://wudao.aminer.cn/turing-test/v2/

**Table 5: The results of Turing tests.**

| Category | Method | User Correct Rate |
|---|---|---|
| General Language Model | GLM | 69.09% |
| | Transfromer-XL | 71.97% |
| | CPM | 78.18% |
| Dialogue Language Model | CDial-GPT | 60.49% |
| | DialoGPT | 70.78% |
| | EVA | 64.77% |
| | PLATO-XL | 65.45% |
| Prompt Engineering | FSB | 63.64% |
| | Inverse Prompt | 70.78% |
| | XDAI | **58.71%** |

volunteer users. The real conversation topic came from Zhihu, one of the largest community Q&A sites in China.

**Result.** There are differences between the results of the Turing test and the previous manual evaluation, which express several interesting insights. Note that a lower ***Correct Rate*** of player selections indicates a better performance. Besides, some models such as CDial-GPT that achieve relatively poor evaluation performance in Table 6 and Table 3 are more likely to confuse the users in the Turing test. The dialogue pre-training method generally performed better, indicating that such fine-tuning with the dialogue corpus still enabled the dialogue to be more fluent and natural.

## 4.4 Online Evaluation

*4.4.1 User Involvement.* We deploy the open-domain XDAI in an online environment to collect feedback from users. Figure 4 presents the statistics of the 200 user activities from October 31st, 2021 to January 31st, 2022. We observe that the 82% users chatted with XDAI over an average of 10 rounds, while many users chatted nearly 100 rounds, with several people chatting an average of 150. In terms of user retention, more than half of all users activate at least ten sessions. The most surprising thing is that six users have become "friends" of XDAI and kept the conversation log with XDAI for over 80 days (92 days in total).

*4.4.2 Case Study.* Figure 3 shows an example of the influence of knowledge injection on the generation of results for large models. The dark-colored responses are the generated results with knowledge injection according to XDAI, while the light-colored responses are the results obtained directly from large-scale models. Both responses are consistent with the content of the dialogue and lead to a coherent context. However, there is a noticeable improvement in the overall quality of the dialogue by invoking knowledge into the query. Moreover, a very interesting phenomenon occurs frequently in our collected examples. The final responses generated by the model often include a wealth of content that is not included in the provided external knowledge (such as *How to learn to ski*), perhaps because this plugged-in knowledge allowed the model to recall some information it had previously learned during the pre-training phase. This phenomenon may be further explored in the future to help extract more knowledge content from the pre-trained models.

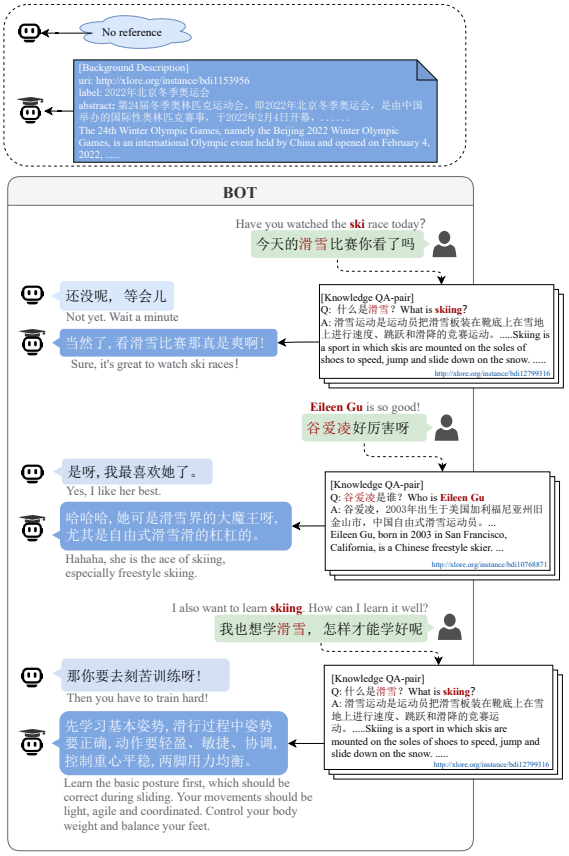


**Figure 3: A typical comparison case of generated responses of with and without knowledge injection.**

## 5 CONCLUSION

In this paper, we present a tuning-free framework, XDAI, for exploiting language models in knowledge-grounded dialogue generation. This framework offers a workflow including offline knowledge curation and online dialogue generation with prompt-based knowledge injection, which provides a convenient toolkit for developers to deploy a PLM-based dialogue system. We conduct extensive experiments including human evaluation, Turing test, and online evaluation to demonstrate the competitive performance of XDAI compared with the state-of-the-art general PLMs and the specific PLMs for dialogue. We hope our framework can provide an easy-to-use and cost-effective solution for individual developers in various domains to build diverse knowledge-grounded applications.

## Acknowledgement

This work is supported by the National Key Research and Development Program of China (2020AAA0106501), the NSFC for Distinguished Young Scholar (61825602), a grant from Beijing Academy of Artificial Intelligence (BAAI) and a grant from the Institute for Guo Qiang, Tsinghua University (2019GQB0003).

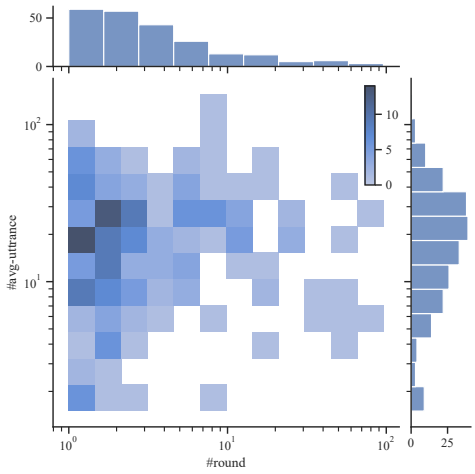


**Figure 4: User involvement of XDAI, where #round and #avg-utterance represent the number of sessions and average number of utterances per session for specific user.**

## REFERENCES


[1] Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, et al. 2021. Plato-xl: Exploring the large-scale pre-training of dialogue generation. *arXiv preprint arXiv:2109.09519* (2021).

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

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

[4] Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. 154–162.

[5] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2978–2988.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. 4171–4186.

[7] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*.

[8] Manda Sai Divya and Shiv Kumar Goyal. 2013. ElasticSearch: An advanced and quick search technique to handle voluminous data. *Compusoft* 2, 6 (2013), 171.

[9] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 320–335.

[10] Marjan Ghazvininejad, Chris Brockett, et al. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI conference*, Vol. 32.

[11] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations.. In *INTERSPEECH*. 1891–1895.

[12] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *EACL*. Association for Computational Linguistics, 874–880.

[13] Hailong Jin, Chengjiang Li, Jing Zhang, Lei Hou, Juanzi Li, and Peng Zhang. 2019. XLORE2: large-scale cross-lingual knowledge graph construction and application. *Data Intelligence* 1, 1 (2019), 77–98.

[14] Vivek Khetan, Roshni Ramnani, Mayuresh Anand, Subhashis Sengupta, and Andrew E Fano. 2022. Causal bert: Language models for causality detection between events expressed in text. In *Intelligent Computing*. Springer, 965–980.

[15] Tushar Khot, Ashish Sabharwal, et al. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of AAAI Conference*.

[16] Chia-Wei Liu, Ryan Lowe, et al. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2122–2132.

[17] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards Emotional Support Dialog Systems. 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)*. 3469–3483.

[18] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT Understands, Too. *arXiv preprint arXiv:2103.10385* (2021).

[19] Michael L Mauldin. 1994. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, Vol. 94. 16–21.

[20] Chuan Meng, Pengjie Ren, et al. 2021. Initiative-Aware Self-Supervised Learning for Knowledge-Grounded Conversations. In *SIGIR*. 522–532.

[21] Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, et al. 2021. ProphetNet-X: Large-Scale Pre-training Models for English, Chinese, Multi-lingual, Dialog, and Code Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. 232–239.

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[23] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. 3982–3992.

[24] Tianxiang Sun, Xiangyang Liu, Xipeng Qiu, and Xuanjing Huang. 2021. Paradigm shift in natural language processing. *arXiv preprint arXiv:2109.12575* (2021).

[25] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1405–1418.

[26] Yida Wang, Pei Ke, et al. 2020. A large-scale chinese short-text conversation dataset. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 91–103.

[27] An Yang, Quan Wang, Jing Liu, Kai Liu, Yajuan Lyu, Hua Wu, Qiaoqiao She, and Sujian Li. 2019. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension. In *ACL*. 2346–2357.

[28] Jifan Yu, Chenyu Wang, Gan Luo, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jie Tang. 2019. Course Concept Expansion in MOOCs with External Knowledge and Interactive Game. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4292–4302.

[29] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2204–2213.

[30] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, et al. 2020. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 270–278.

[31] Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, et al. 2022. Cpm-2: Large-scale cost-effective pre-trained language models. *AI Open* (2022).

[32] Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv preprint arXiv:2109.06513* (2021).

[33] Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual Probing Is [MASK]: Learning vs. Learning to Recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5017–5033.

[34] Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, et al. 2021. EVA: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547* (2021).

[35] Hao Zhou, Tom Young, et al. 2018. Commonsense knowledge aware conversation generation with graph attention.. In *IJCAI*. 4623–4629.

[36] Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. Kd-Conv: A Chinese Multi-domain Dialogue Dataset Towards Multi-turn Knowledge-driven Conversation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7098–7108.

[37] Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. 2021. Controllable generation from pre-trained language models via inverse prompting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2450–2460.

# 6 APPENDIX

## 6.1 Implementation Details

As mentioned in Section 3, the online service of XDAI employs the Chinese-versioned GLM with 10B parameters [9] as the base PLM, which is trained on 302GB of raw Chinese data collected from multiple Chinese websites. The online service is employed with a server with 8 Nvidia V100 GPUs, Intel CPU cores, and 376GB of Memory. The generation of responses is based on beam search with a bean size of 10. We employ MongoDB for building databases and utilize Elastic Search for indexing.

For our publicly available open-domain XDAI service, we finally collect over 5.4 GB of knowledge resources of 15,273,635 concepts.

## 6.2 Human Evaluation Details

For the open-domain and domain-specific experiments, the human evaluation is conducted on our developed online platform, as shown in Figure 5. To prevent annotation bias, the annotators are divided into groups of three-member size, where each group is assigned the same dialogues. The dialogues of all baselines are mixed, the source of which is hidden from the annotators.

Meanwhile, the scores given by any annotators are invisible to the others. All scoring items can be modified and withdrawn before submission. The evaluation does not necessarily need to be finished at once. People can log in, change their answers for the already-completed problems, or continue evaluation from their last evaluation position freely in one week. They only need to ensure that all evaluation questions have been answered before the deadline, with the scores being consistent.

For the Turing test, we selected robot responses from real conversations based on selected topics and invited several volunteers, mainly college students, to write corresponding human responses for each conversation. Annotators need to complete a binary option to pick human responses from human responses and machine-generated responses.

## 6.3 Deviation for Human Evaluation

Table 5, 6 display the deviation of the scorings for human annotators in open-domain and domain-specific evaluation. The deviation is calculated on a per-capita basis that we first average the scorings for each method for every annotator. We compute the deviation based on the average scores of all human annotators.

From the statistics, human evaluation can consistently demonstrate model effects. Although annotators agree more on the evaluation of open-domain dialogue rather than the domain-specific dialogue, such fluctuation is mainly likely to be caused by the limited acquisition of domain knowledge and the natural bias of the annotation samples.

## 6.4 Online Demo Deployment

Our service and repository are provided at https://github.com/THUDM/XDAI, where developers can also get a Q-R code for applying to be a test user and participate in the WeChat chatbot dialogue service. Users interested in the Turing test game can join the online game at https://wudao.aminer.cn/turing-test/v2/?nowGame=qa.

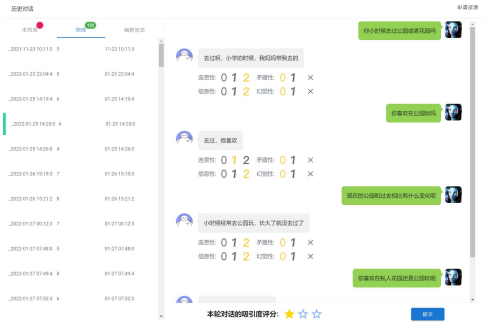

**Figure 5: A screenshot of our human evaluation platform, where annotators are grouped for double-blinded evaluation, i.e, not accessible to the source of the assigned conversations.**

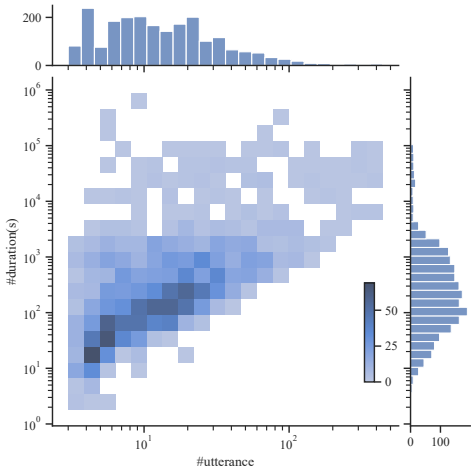


**Figure 6: User activity of our XDAI implementation example, where #utterance and #duration represent the number of utterances in one conversation and the duration of this conversation.**

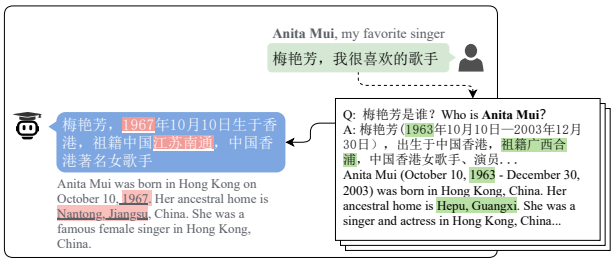


**Figure 7: Another case generated by XDAI w/o background knowledge, which indicates the Hallucination phenomenon of grounded dialogue generation.**

For each user who chats with XDAI on WeChat, XDAI automatically sends out a daily survey to collect feedback from them. We

**Table 6: 95% confidence intervals for human evaluation results of open-domain dialogue generation.**

| Category | Method | Open-Domain Dialogue | | | | |
|---|---|---|---|---|---|---|
| | | Coherence | Inconsisitency↓ | Informativeness | Hallucination↓ | Engagingness(C) |
| General Language Model | GLM | 1.267±0.092 | 0.071±0.046 | 1.163±0.088 | 0.180±0.038 | 0.767±0.081 |
| | Transfromer-XL | 1.158±0.052 | 0.119±0.025 | 1.096±0.056 | 0.193±0.021 | 0.750±0.043 |
| | CPM | 1.273±0.055 | 0.076±0.023 | 1.192±0.056 | 0.146±0.018 | 0.767±0.045 |
| Pretrained Dialogue Model | CDial-GPT | 0.913±0.091 | 0.158±0.046 | 0.746±0.088 | 0.270±0.038 | 0.543±0.081 |
| | DialoGPT | 0.857±0.079 | 0.175±0.043 | 0.772±0.080 | 0.293±0.036 | 0.556±0.071 |
| | EVA | 1.172±0.063 | 0.127±0.035 | 1.288±0.063 | 0.340±0.025 | 0.703±0.061 |
| | PLATO-XL | **1.534±0.063** | **0.048±0.025** | 1.297±0.073 | **0.072±0.027** | **1.256±0.115** |
| Controllable Generation | FSB | 0.784±0.085 | 0.100±0.047 | 0.833±0.089 | 0.298±0.031 | 0.383±0.065 |
| | Inverse Prompt | 1.408±0.081 | 0.238±0.050 | 1.492±0.079 | 0.225±0.049 | / |
| | XDAI | **1.512±0.034** | **0.054±0.017** | **1.658±0.029** | **0.162±0.011** | **1.125±0.037** |

**Table 7: 95% confidence intervals for human evaluation results of domain-specific dialogue generation.**

| Domain | Category | Methods | Coherence | Inconsistency↓ | Informativeness | Hallucination↓ | Engagingness(C) |
|---|---|---|---|---|---|---|---|
| Travel Domain | General Language Model | GLM | 1.590±0.120 | 0.028±0.032 | 1.457±0.147 | 0.057±0.044 | 1.057±0.144 |
| | | Transfromer-XL | 1.293±0.173 | 0.240±0.097 | 1.413±0.170 | 0.293±0.103 | 1.120±0.182 |
| | | CPM | 1.228±0.191 | 0.114±0.075 | 1.100±0.191 | 0.142±0.082 | 0.928±0.140 |
| | Pretrained Dialogue Model | CDial-GPT | 1.384±0.185 | 0.078±0.065 | 0.984±0.199 | 0.107±0.075 | 0.538±0.155 |
| | | DialoGPT | 1.164±0.187 | 0.095±0.068 | 1.027±0.187 | 0.191±0.090 | 0.630±0.173 |
| | | EVA | 1.229±0.227 | 0.062±0.069 | 1.313±0.241 | 0.229±0.120 | 0.708±0.240 |
| | | PLATO-XL | 1.625±0.141 | 0.069±0.058 | 1.458±0.176 | **0.041±0.046** | **1.460±0.113** |
| | Controllable Generation | FSB | 0.944±0.269 | 0.083±0.091 | 1.056±0.301 | 0.250±0.143 | 0.472±0.240 |
| | | Inverse Prompt | 1.250±0.089 | 0.208±0.050 | 1.292±0.075 | 0.208±0.050 | / |
| | | XDAI | **1.660±0.128** | **0.020±0.027** | **1.770±0.103** | **0.060±0.046** | **1.430±0.153** |
| Sports Domain | General Language Model | GLM | 1.099±0.110 | 0.099±0.039 | 1.001±0.114 | 0.253±0.057 | 0.624±0.095 |
| | | Transfromer-XL | 0.984±0.138 | 0.192±0.068 | 0.961±0.141 | 0.292±0.078 | 0.701±0.134 |
| | | CPM | 1.235±0.114 | 0.133±0.048 | 1.219±0.114 | 0.171±0.054 | 0.748±0.104 |
| | Pretrained Dialogue Model | CDial-GPT | 0.611±0.164 | **0.082±0.058** | 0.447±0.141 | 0.164±0.079 | 0.270±0.095 |
| | | DialoGPT | 0.983±0.235 | 0.216±0.105 | 1.033±0.223 | 0.402±0.125 | 0.666±0.240 |
| | | EVA | 0.777±0.211 | 0.259±0.117 | 1.055±0.228 | 0.388±0.131 | 0.574±0.191 |
| | | PLATO-XL | 1.447±0.135 | 0.236±0.105 | 1.236±0.183 | 0.157±0.090 | 0.833±0.168 |
| | Controllable Generation | FSB | 0.745±0.151 | 0.160±0.070 | 0.849±0.157 | 0.292±0.087 | 0.490±0.142 |
| | | Inverse Prompt | **1.684±0.070** | 0.157±0.045 | **1.789±0.064** | **0.105±0.038** | / |
| | | XDAI | **1.378±0.099** | **0.179±0.047** | **1.410±0.097** | **0.215±0.050** | **0.936±0.104** |

also welcome users to upload interesting conversation transcripts, which are used to be posted for sharing in the public community.

As introduced in the online evaluation, we collect the interaction data for 92 days with permission. Figure 6 represents some other statistics of these records. The average number of chat rounds for users is 27, and an astonishing 5% of conversations have more than 100 rounds. In terms of conversation duration, the average human response time is about 27 seconds, while the total duration of a chat is about 16 minutes. About 38% of the conversations had an average response time of fewer than 10 seconds, indicating that these users really seemed to view the conversation as a conversation with a human rather than a chat experiment with a bot.

Figure 7 shows another case generated by XDAI without background knowledge. This example about Anita Mui fails to generate factually correct content and even has a retelling error after the knowledge injection, which indicates that Hallucination remains a significant challenge to be addressed for Knowledge-grounded dialogue generation applications. We also expect developers to design components for XDAI to help address such potential risks.

## 6.5 Time Efficiency

We also analyzed its efficiency by comparing the average speed with baselines. XDAI maintains a pretty fast response speed regarding the response time because it only performs a single inference. Compared with its base model GLM, steps such as knowledge injection only slightly increase the latency and do not affect the usability; while Inverse Prompt, which previously performed well in the sports domain, has an obvious shortcoming in efficiency because it needs to generate multiple times and inversely search for optimal results, making it almost impossible to be used in practical applications, although it provides excellent results.