DialogueForge: LLM Simulation of Human-Chatbot Dialogue

Ruizhe Zhu*
ETH Zurich
Zurich, Switzerland

Hao Zhu*
ETH Zurich
Zurich, Switzerland

Yaxuan Li*
ETH Zurich
Zurich, Switzerland

Syang Zhou
Calvin Risk AG
Zurich, Switzerland

Shijing Cai
Calvin Risk AG
Zurich, Switzerland

Małgorzata Łazuka
Calvin Risk AG
Zurich, Switzerland

Elliott Ash
ETH AI Center, ETH Zurich
Zurich, Switzerland

Abstract

Collecting human-chatbot dialogues typically demands substantial manual effort and is time-consuming, which limits and poses challenges for research on conversational AI. In this work, we propose DialogueForge – a framework for generating AI-simulated conversations in human-chatbot style. To initialize each generated conversation, DialogueForge uses seed prompts extracted from real human-chatbot interactions. We test a variety of LLMs to simulate the human chatbot user, ranging from state-of-the-art proprietary models to small-scale open-source LLMs, and generate multi-turn dialogues tailored to specific tasks. In addition, we explore fine-tuning techniques to enhance the ability of smaller models to produce indistinguishable human-like dialogues. We evaluate the quality of the simulated conversations and compare different models using the UniEval and GTEval evaluation protocols. Our experiments show that large proprietary models (e.g., GPT-4o) generally outperform others in generating more realistic dialogues, while smaller open-source models (e.g., Llama, Mistral) offer promising performance with greater customization. We demonstrate that the performance of smaller models can be significantly improved by employing supervised fine-tuning techniques. Nevertheless, maintaining coherent and natural long-form human-like dialogues remains a common challenge across all models.

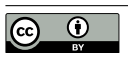
Keywords

Large Language Models, Human-Chatbot, Fine-Tuning, Dialogue Generation

*Equal contribution.

Author contributions using the CRediT framework [4, 10]:
Ruizhe Zhu: Dialogue Generation, Experiments, Codebase maintenance, Writing
Hao Zhu: Dialogue Evaluation, Visualization, Writing
Yaxuan Li: Datasets, Experiments, Writing
Syang Zhou, Shijing Cai, Małgorzata Łazuka: Conceptualization, Data Preprocessing, Supervision, Financial Support, Review
Elliott Ash: Supervision

Syang Zhou: sz@calvin-risk.com



Agentic & GenAI Evaluation Workshop KDD'25, Toronto, ON, Canada

1 Introduction

Large language models (LLMs) have rapidly emerged as a transformative force in artificial intelligence, which marks a significant milestone in the development of natural language processing (NLP), driving substantial progress across a wide range of NLP tasks and paving new paths for research. Their ability to generate coherent and human-like text has made them central to the development of advanced dialogue systems [9]. As these models become more capable, the demand for high-quality human-chatbot dialogue data, both for training and evaluation, continues to grow [29].

In order to advance the study and development of LLMs, large-scale human-chatbot dialogue data collection is indispensable. However, gathering and annotating these human-involved dialogues remains a major bottleneck. Traditional crowd-sourcing data collection methods, which rely on human participants and manual annotation, are time-consuming, costly, and often difficult to scale, which can pose significant challenges to the research process [28]. LLMs have recently emerged as effective solutions to this challenge. Their powerful architectures and capacity to learn from massive datasets allow them to produce contextually appropriate, informative, and human-like responses [28]. Several recent works leverage large language models directly for synthetic dialogue generation:

One-Go In-Context Prompting. A pre-trained LLM is prompted with a seed (topic description, knowledge-graph-derived summary, or few-shot examples) to produce an entire multi-turn conversation in a single pass. For example, PLACES [5] uses few-shot prompts combining human-written background info and sample turns to GPT-3 to generate full dialogues. SODA [19] employs GPT-3.5 to expand knowledge graph triplets into short narrative seeds and then into multi-turn chats. BotChat [12] leverages multiple different LLMs to generate multi-turn dialogues utterance-by-utterance from authentic human-bot chat seeds.

Fine-Tuning. The LLM is first fine-tuned on a small dialogue completion corpus, then used for generation, for instance, AUGESC [38] fine-tunes GPT-J on ESConv [25] dataset, then generates conversations from a description plus the first turn.

Turn-by-Turn Multi-Agent Simulation. Two or more LLM agents converse sequentially, often to model different personas or to mix skills: PERSONACHATGEN [23] runs two GPT-3 instances, each conditioned on distinct persona profiles, generating one turn at a time. BOTSTALK [20] engages multiple GPT-3 models that alternate turns, selecting from different skill-specific datasets turn by turn.

Task-Oriented In-Context LLM Simulators. Prompt-based LLM methods generate dialogues for task-completion settings without any fine-tuning. ICL-US [33] uses few-shot examples plus a user goal and history to generate user-agent turns via in-context learning. Dialogic [24] similarly prompts GPT-3 with ontology-extracted goals and in-context examples, then applies a critic step to enforce belief-state alignment.

While prior LLM-based conversation generation techniques tend to specialize in a single paradigm, whether one-go in-context prompting, fine-tuning, or turn-by-turn multi-agent simulation, we introduce **DialogueForge**, a novel framework that unifies these strategies. DialogueForge starts from authentic human-chatbot exchanges, using only the first user utterance as a prompt. It infers each dialogue's underlying task objective from the original conversation and then iteratively generates alternating "Inquirer" (human) and "Responder" (bot) turns with a chosen LLM (see Figure 1). This approach enables scalable synthesis of diverse, task-tailored multi-turn dialogues, drastically reducing cost and time required to collect human-chatbot style conversation data manually.

To assess LLMs' ability to produce human-like conversational flow, adhere to task goals, and sustain coherent dialogue across multiple turns, DialogueForge adopts two evaluation metrics proposed in [12]. Both metrics employ an LLM-as-a-judge approach: **UniEval** evaluates individual dialogue quality, while **GTEval** compares generated conversations against ground-truth conversations.

Finally, to further boost generation quality, DialogueForge applies supervised fine-tuning: base LLMs are trained on a curated corpus of high-quality human-chatbot exchanges that span varied tasks, linguistic styles, and interaction patterns. Through this fine-tuning process, the model learns to predict each next turn given full dialogue context, which reinforces realistic conversational behaviors and enhances the overall naturalness of synthetic dialogues.

In this work, we make three key contributions:

(1) **DialogueForge**, a unified synthetic conversation generation framework that seeds real human-chatbot exchanges with only the first user utterance, automatically infers each dialogue's task goal, and combines in-context LLM prompting, iterative multi-agent simulation, and supervised fine-tuning to produce richly varied, task-tailored multi-turn dialogues at scale.
(2) A comprehensive evaluation of both proprietary (e.g. GPT-4o) and open-source (e.g., Llama, Mistral) LLMs using two LLM-as-judge metrics: UniEval and GTEval, showing that (a) larger models generally outperform smaller ones, (b) dialogue quality degrades as conversation length increases, and (c) fine-tuning substantially boosts the performance of smaller models.
(3) An analysis of evaluation bias, where we adapt BotChat's judging protocols [20] to compare different judge LLMs (GPT-4o, Claude 3.7, Gemini 2.0 Flash), demonstrating the robustness of our metrics across judge choices.

All code, models, and datasets are publicly available on GitHub[1]. We believe that developing DialogueForge and releasing open-source code of the framework as well as conversation data and fine-tuned LLMs, we will enable and accelerate further research in the area of conversational AI, allowing researchers with limited resources to contribute to this field. Our approach addresses a key challenge faced in dialogue research: the expensive and time-consuming process of collecting human-chatbot interactions, by providing a unified framework that can generate diverse, task-oriented conversations at scale. The demonstrated performance of fine-tuned smaller models suggests that high-quality dialogue generation does not necessarily require large proprietary systems, which has important implications for research accessibility and the development of conversational AI in resource-limited settings.

The rest of the paper is organized as follows: Section 2 reviews related work on LLM-based dialogue generation and evaluation; Section 3 details the DialogueForge methodology, including seed prompt extraction, dialogue generation, and fine-tuning strategies; Section 4 introduces our evaluation setup (Section 4.1) and presents experimental results (Section 4.2); Section 5 discusses the implications, limitations, and potential enhancements of DialogueForge; and Section 6 concludes and outlines directions for future work.

2 Related Works

2.1 LLM Data Generation

Recent advances in synthetic dialogue generation have been driven by large-scale, high-quality datasets. SODA [19] distills socially contextualized dialogues from LLMs, producing more natural and consistent conversations. UltraChat [11] scales instruction-tuned dialogue generation with 1.5M synthetic conversations, enabling the strong UltraLM model. Baize [34] generates self-chat data using ChatGPT and introduces feedback-based self-distillation to fine-tune open-source models. Complementing synthetic data, WildChat [37] provides a real-world dataset of 1M ChatGPT-user interactions, enriched with demographic metadata. Lately, LLM Roleplay [30] demonstrated that persona-based conditioning improves engagement and specificity. These works collectively highlight trends in scaling, personalization, and leveraging both synthetic and real-world data to improve dialogue quality.

Recently, numerous studies have employed fine-tuning techniques spanning full fine-tuning, adapter-based PEFT, and progressive learning to improve the models' ability to generate high-quality data following specific instructions. For example, the Falcon series [1] fine-tunes open LLMs on massive, diverse text corpora to boost generation fluency and factual consistency. More recently, Orca [26] employs progressive fine-tuning on GPT-4 explanation traces to teach complex reasoning patterns, and WizardLM [35] fine-tunes large pre-trained models to follow intricate multi-step instructions with high fidelity.

2.2 Human-Chatbot Datasets

Several large-scale conversational datasets have been released to support both open-domain chit-chat and human-chatbot interaction research. PersonaChat [36] comprises crowd-authored daily chit-chat dialogues between paired personas. The OpenAssistant [21] dataset provides human-crafted assistant responses to a wide variety of prompts in English, as do Anthropic HH [14], Chatbot Arena [6], and LMSYS-Chat-1M [39], each collecting high-quality human-bot exchanges across diverse topics. Finally, MT Bench [3]

[1]Available at https://github.com/nerchio/Human_Chatbot-Generation

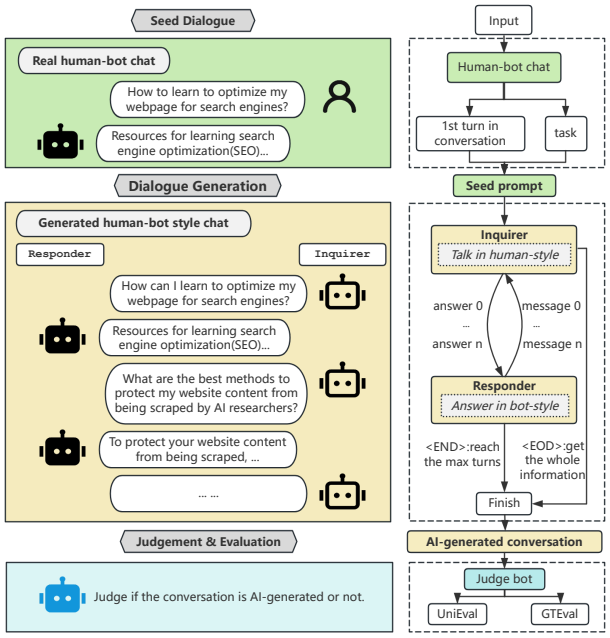


Figure 1: Flow of the dialogue generation and evaluation process of DialogueForge.

contributes fine-grained human judgments on model outputs to benchmark machine translation in conversational settings. Together, these resources underpin much of the recent progress in training and evaluating advanced dialogue systems.

2.3 Evaluation of LLMs

To better assess LLM competence, Hendrycks et al. [16] introduced the MMLU benchmark, which tests multitask language understanding across 57 diverse subjects. Addressing the challenge of multi-step reasoning, particularly in math, Cobbe et al. [7] proposed a verifier-based method to improve model performance on the GSM8K dataset. Meanwhile, GPTScore [13] is a customizable, model-based evaluation framework that leverages large pre-trained language models for multi-dimensional text assessment without requiring annotated data. Duan et al. [12] used three evaluation protocols to assess LLM's capability of having multi-turn dialogues. These efforts collectively contribute to both developing more capable models and building the tools necessary to assess their real-world performance more effectively.

3 Methodology

In this section, we present the general structure of DialogueForge, our dialogue generation methodology, and further implementation details. First, Section 3.1 presents a high-level overview of the dialogue generation and evaluation workflow used in DialogueForge. Then, Section 3.2 provides a detailed description of DialogueForge implementation, including the LLMs used internally to generate chat messages. Finally, Section 3.3 discusses our experimental methodology of fine-tuning smaller LLMs to enhance their performance in generating human-like conversations.

3.1 DialogueForge Architecture

The overall process of dialogue generation and evaluation implemented in DialogueForge has been adapted from BotChat [12] and is illustrated in Figure 1.

3.1.1 Seed prompt extraction. We begin by seeding the system with a real dialogue segment, extracted from an actual interaction between a human and a chatbot. This seed consists of an initial human query and the corresponding response from the chatbot. Additionally, the seed prompt extraction process involves parsing the original human-chatbot exchange to identify the central topic of the conversation, which is then embedded into generation prompts. This topic serves as a guiding constraint, helping to ensure that the subsequent generated dialogue remains coherent and does not deviate significantly from the theme of the original exchange.

However, it is important to note that DialogueForge does not attempt to faithfully recreate the original seed dialogue beyond utilizing the initial utterance as a generation prompt, nor does it explicitly evaluate task completion, toxicity or hallucination rates. The framework generates entirely novel conversational trajectories guided by the inferred topic. Our evaluation methodology focuses exclusively on conversational naturalness and the ability to produce human-indistinguishable dialogue, as quantified through UniEval and GTEval metrics. This design choice reflects our primary research objective of advancing synthetic dialogue generation quality rather than task-oriented performance assessment.

3.1.2 Dialogue generation. Following this initialization, we simulate an extended multi-turn dialogue by alternating between two large language models, each assuming a distinct conversational role. The first model, referred to as the **inquirer**, acts as the simulated human participant, while the other model, called the **responder**, plays the role of the chatbot. Importantly, we treat the inquirer and the responder models as two independent agents. They are instantiated with different prompts and can even be based on entirely different LLM architectures. This allows us to study agent interactions with a high degree of flexibility and realism.

Throughout the generation process, the full history of the conversation is passed between the two models at each turn, enabling context-aware responses. The generation continues iteratively, with each model producing one utterance per turn, until either (a) a predefined maximum number of turns is reached, or (b) the inquirer determines that the dialogue has naturally concluded. In our experimental setup, we set strict turn limits of 6 and 12 turns to enable controlled comparison across models and datasets. The inquirer model is therefore not explicitly programmed to recognize task completion, and the termination is mainly based on these predetermined constraints.

3.1.3 Judgement. The AI-generated conversation between the inquirer and the responder models generated in the dialogue generation step is then passed to the judge bot. The judge assesses the dialogue using UniEval and GTEval metrics (discussed in more detail in Section 4.1).

3.2 Implementation

The implementation of DialogueForge leverages LangGraph [22], a framework designed to streamline the construction of multi-agent communication pipelines. LangGraph facilitates the orchestration of message passing between agents and provides mechanisms for specifying various termination conditions, enabling more structured and manageable dialogue simulations. Below, we provide further details regarding the real-world dialogue datasets and large language models used in DialogueForge.

3.2.1 Seed dialogue datasets. Obtaining high-quality, large-scale public datasets containing real human-chatbot dialogues remains a non-trivial challenge due to the difficulty of large-scale data collection and privacy concerns. In this study, we leverage two publicly available datasets hosted on Hugging Face: OpenAssistant Conversations (OASST1) [21] with 413 conversations and Chatbot Arena Conversations [40] with 1,002 conversations. To ensure consistency and relevance for our experiments, we performed data cleaning and reformatting to extract coherent and contextually rich human-chatbot dialogue samples. These curated datasets serve as the basis for all subsequent evaluations. More details on data preprocessing can be found in our publicly available code repository[1].

To investigate how dialogue length affects model behavior and performance, we define two dialogue length constraints in our experimental design: a maximum of 6 turns and a maximum of 12 turns per conversation. This stratification enables us to analyze model robustness and consistency across varying interaction depths.

3.2.2 Inquirer LLMs. The primary focus of our research is to improve the performance of the inquirer agent (that is, the model that drives the conversation and simulates the human chatbot user). To that end, we consider and evaluate a variety of candidate models for this role, grouped into two broad categories:

- Large, high-capacity LLMs that demonstrate strong conversational fidelity and produce human-like dialogue, but are often resource-intensive and impractical for local or edge deployments:
  - GPT-4o [27]
  - GPT-4o mini [27]
  - DeepSeek-V3 [8]
  - Llama-3.3-70B [15]
  - Gemma-2-27B [32]
- Smaller, lightweight models that are more computationally efficient and better suited for deployment in resource-constrained environments, yet may underperform in generating natural and engaging conversations:
  - Llama-3.2-3B [15]
  - Llama-3.1-8B [15]
  - Mistral-7B [18]

To bridge the performance gap between large and small models, we apply supervised fine-tuning to the smaller models, described in detail in Section 3.3.

3.2.3 Responder LLM. For all experiments, we designate GPT-4o mini as the responder model due to its demonstrated stability and cost-effectiveness in diverse conversational scenarios [27]. The responder receives the complete conversation history at each generation step and produces contextually appropriate chatbot responses that are consistent with the established topic and conversational tone. This model selection ensures that any variations in dialogue quality observed in our experimental results can be attributed primarily to the inquirer model performance rather than inconsistencies in the responder model's behavior.

3.2.4 Judge bot. In the majority of experiments presented in Section 4, GPT-4o [27] serves as the default judge LLM unless otherwise specified. However, we acknowledge that the choice of the judge LLM may cause evaluation bias – for example, the judge LLM might favor dialogues generated using the same LLM as the judge bot itself. We have conducted experiments to verify this using two alternative LLMs acting as judge models: Claude 3.7 [2] and Gemini 2.0 Flash [31]. The results of these experiments are presented and discussed in Section 4.2.5.

3.3 Fine-tuning

To enhance the performance of smaller models, we apply supervised fine-tuning using Low-Rank Adaptation (LoRA) [17] and a carefully curated subset of our synthetic dialogue corpus. This fine-tuning process aims to transfer conversational behaviors and strategies learned from more capable models, thereby improving the ability of smaller models to emulate human-like inquiries while retaining their lightweight deployment advantages.

The fine-tuning datasets are carefully constructed by extracting and formatting samples from the same dialogue corpora as used for seed prompt extraction: OASST1 and Chatbot Arena (see Section 3.2.1). Specifically, we adopt a cross-dataset fine-tuning strategy: when evaluating a model on the OASST1 dataset, we fine-tune it using data sampled from the Chatbot Arena dataset and vice versa. This approach is designed to improve the generalization of smaller models across diverse dialogue styles and domains.

```
{
  "chat_template": "tokenizer",
  "distributed_backend": "ddp",
  "mixed_precision": "fp16",
  "optimizer": "adamw_torch",
  "peft": "true",
  "scheduler": "linear",
  "unsloth": "false",
  "batch_size": "2",
  "block_size": "1024",
  "epochs": "3",
  "gradient_accumulation": "4",
  "lr": "0.00001",
  "model_max_length": "16384",
  "target_modules": "all-linear",
  "merge_adapter": "true"
}
```

We use Hugging Face Autotrain to easily generate fine-tuning scripts and here we show the parameters we use. Further details on training configurations and evaluation procedures can be found in our publicly available code repository[1].

4 Evaluation

This section presents the methodology and results of our experimental analysis of various LLMs used in DialogueForge. First, in Section 4.1 we describe the evaluation metrics used to assess and compare the performance of various LLMs. Then, in Section 4.2 we present the experimental results obtained using various LLMs, including those fine-tuned in our study.

4.1 Evaluation metrics

We employ two evaluation metrics introduced in BotChat [12].

UniEval: In this evaluation, the generated conversation is independently assessed by the LLM judge, which determines whether the dialogue resembles a genuine human-chatbot interaction. For each conversation, the LLM judge performs the following steps:

(1) Determines whether the conversation appears to involve an AI participant (Yes/No).
(2) If the answer is "Yes," identifies the first utterance it recognizes as AI-generated and outputs its index.
(3) Provides a rationale for its decision.

Below is a sample UniEval result.

```
{
  "choice": "Yes",
  "index": "3",
  "reason": "The human utterance in the
      third chat is overly detailed and
      structured, resembling a list of
      questions that are too perfectly
      organized and comprehensive for a
      natural human conversation. It lacks
      the informal, sometimes inconsistent
      phrasing typical of human dialogue,
      and instead reads like a scripted set
      of questions designed to cover all
      aspects of the topic systematically."
}
```

GTEval: In this setting, the LLM judge is presented with a pair of dialogues, one of which is an authentic conversation between a human and a chatbot. The judge is tasked with identifying whether either dialogue appears to involve AI-generated content. It also provides a rationale for its decision.

GTEval plays a crucial role in the evaluation framework as it assesses whether the generated conversations are indistinguishable from real human-chatbot interactions.

A sample GTEval result is shown below:

```
{
  "choice": "Conversation 2",
  "reason": "In Conversation 2, the human
      utterances are unusually detailed and
      structured, with a level of fluency
      and coherence that is atypical for
      spontaneous human conversation. The
      human responses include multiple
      questions and reflections in a single
      turn, which is more characteristic of
      AI-generated text. In contrast,
      Conversation 1 has more typical human-
      like brevity and directness in the
      questions asked."
}
```

Both evaluation metrics are designed to assess authenticity and human-likeness of conversations, addressing the primary interest of our study. They do not encompass other dimensions of dialogue quality assessment, such as task completion rates, factual accuracy, or toxicity, which fall outside the scope of our evaluation framework.

4.2 Experimental results

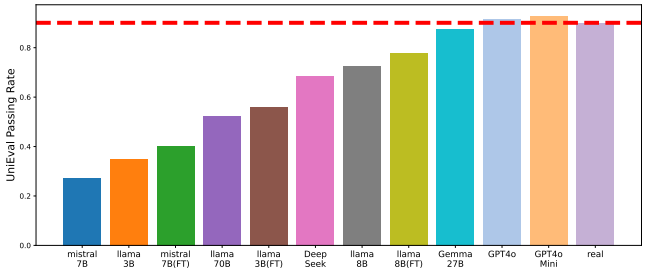
In this section, we present the results of all experiments conducted to evaluate DialogueForge. First, in Sections 4.2.1 and 4.2.2 we present the UniEval and GTEval scores, respectively, achieved by various inquirer LLMs considered in DialogueForge. In Section 4.2.3 we study the performance improvement achieved with smaller LLMs thanks to supervised fine-tuning. Then, in Section 4.2.4 we examine how model performance degrades as conversation length increases. Finally, in Section 4.2.5 we verify the impact that the choice of the judge LLM has on the final evaluation scores.

4.2.1 UniEval Results. Figure 2 presents the UniEval passing rates achieved by various inquirer LLMs considered in DialogueForge. Based on these results, we make the following observations:

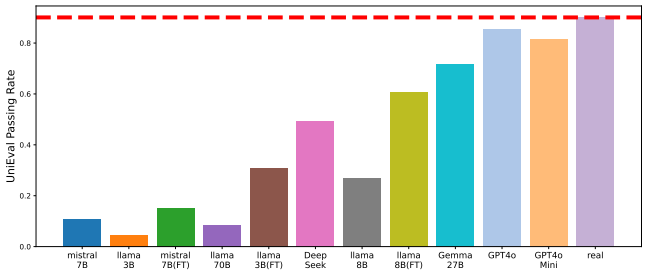
1. **Larger models generally perform better**. Among the evaluated LLMs, GPT-4o and GPT-4o mini achieve the highest passing rates. On the OASST1 6-turns dialogue dataset, both of them exceed 90%, even slightly outperforming real human-chatbot conversations. In contrast, smaller models perform significantly worse. For example, on the OASST1 dataset, Mistral-7B achieves a passing rate of only 27.36%, and Llama-3.2-3B reaches 35.11%. However, **some exceptions** exist. Notably, Llama-3.3-70B underperforms compared to Llama-3.1-8B, contrary to expectations based on model size. Conversely, Gemma-2-27B performs remarkably well – its passing rate on the OASST1 6-turn conversation dataset reaches 87.41%, closely approaching the performance of GPT-4o and GPT-4o mini.

2. **Fine-tuning improves performance**. All smaller models (e.g. Mistral-7B, Llama-3.2-3B, and Llama-3.1-8B) exhibit improved passing rates after fine-tuning. This demonstrates the effectiveness of task-specific adaptation, even for relatively limited model capacities. A more detailed comparison of fine-tuned versus base model performance is presented in Section 4.2.3.

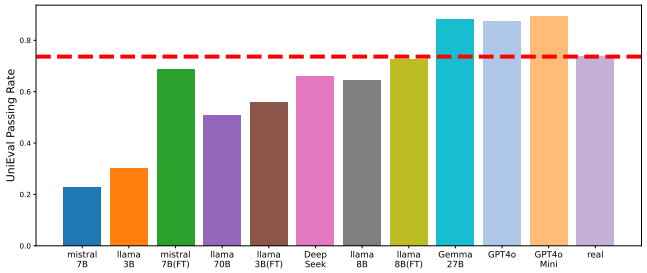
3. **LLMs' ability to simulate human-like dialogue declines with longer conversations**. As the number of utterances in a conversation increases, the passing rates tend to decrease across



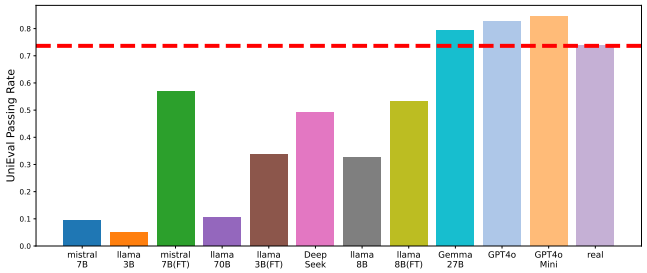
(a) OASST1 dataset: 6 turns



(b) OASST1 dataset: 12 turns



(c) Chatbot Arena dataset: 6 turns



(d) Chatbot Arena dataset: 12 turns

Figure 2: UniEval passing rates (No AI-Involved) of OASST1 dialogues. Models labelled with "FT" are fine-tuned. The red dashed line marks the passing rate of real human-chatbot conversations for reference.

all models. However, some LLMs exhibit a less pronounced performance drop compared to others, indicating better robustness in maintaining human-likeness over extended interactions. A more detailed analysis of this trend is provided in Section 4.2.4.

4.2.2 GTEval Results. Comparing the GTEval results (presented in Figure 3) with the UniEval results (Figure 2), we find that the two evaluation methods generally produce consistent rankings

	Base (%)	Finetuned (%)	Improvement (%)
Llama-3.2-3B	35.11|4.36	55.93|30.75	20.82|26.39
Llama-3.1-8B	72.64|26.88	77.97|60.77	5.33|33.89
Mistral-7B	27.36|10.90	40.19|15.25	12.83|4.35

(a) Improvement on the OASST1 dataset (UniEval Passing Rate, No AI-involved). Each cell: 6 and 12 max-turns.

	Base (%)	Finetuned (%)	Improvement (%)
Llama-3.2-3B	30.34|4.79	55.88|33.83	25.54|29.04
Llama-3.1-8B	64.57|32.44	72.55|53.29	7.98|20.85
Mistral-7B	22.65|9.48	68.56|56.79	45.91|47.31

(b) Improvement on the Chatbot Arena dataset (UniEval Passing Rate, No AI-involved). Each cell: 6 and 12 max-turns.

Table 1: UniEval Passing Rate improvements after fine-tuning on (a) OASST1 and (b) Chatbot Arena datasets.

and trends. However, several noteworthy differences and insights emerge, which merit further discussion:

1. **The GTEval indistinguishability rate is significantly lower than the UniEval passing rate**. This suggests that when a real human-chatbot conversation is provided as a reference point, the LLM judge becomes more discerning in identifying AI-generated dialogues.

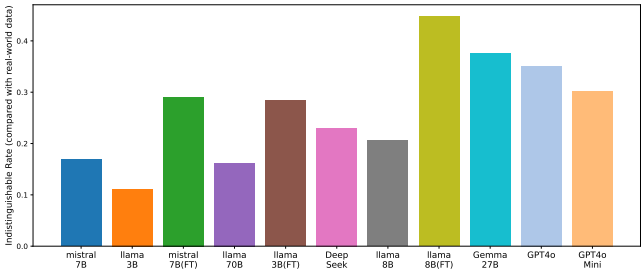
2. **The fine-tuned Llama-3.1-8B model exhibits surprisingly strong performance**. On the OASST1 dataset, for dialogues with 6 utterances, it even achieves a higher GTEval indistinguishability score than GPT-4o. This finding highlights the potential of smaller models to approximate human-like conversational behavior when fine-tuned effectively. It provides evidence that, with targeted fine-tuning, compact LLMs can generate high-quality, realistic dialogues with chatbot that are competitive with much larger models.

4.2.3 Fine-tuning effects. In this project, we fine-tuned three small-scale language models to assess whether fine-tuning enhances their ability to simulate human-like dialogue in conversations with chatbots. The goal is to evaluate whether such adaptations make smaller models viable for high-quality interaction tasks.

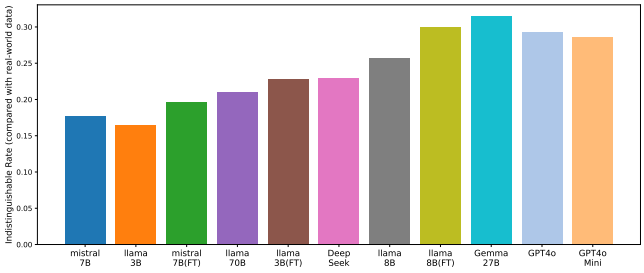
Performance improvements resulting from fine-tuning are presented in Tables 1 and 2. We use a cross-dataset evaluation strategy: models fine-tuned on the Chatbot Arena dataset are evaluated on OASST1, and vice versa.

The benefits of fine-tuning are evident across all three models. Notably, we observe greater performance improvements on the Chatbot Arena dataset (when models are fine-tuned using the OASST1 dataset) than on the OASST1 dataset (when fine-tuned using Chatbot Arena), despite the larger size of the Chatbot Arena dataset. We hypothesize that this discrepancy stems from differences in data quality: the human-chatbot dialogues in the Chatbot Arena dataset may be of lower quality, which could negatively impact the effectiveness of fine-tuning. This highlights the importance of not only dataset size but also the quality and consistency of the training data in model adaptation.

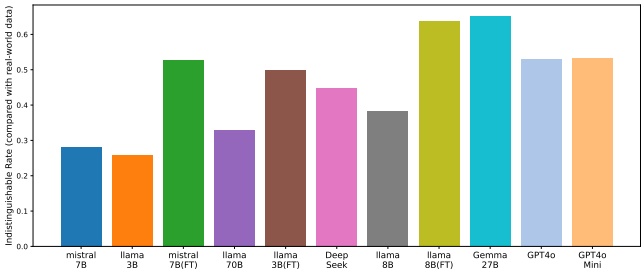
4.2.4 Performance degradation as the dialogue becomes longer. We analyze how the model performance changes as the length of the



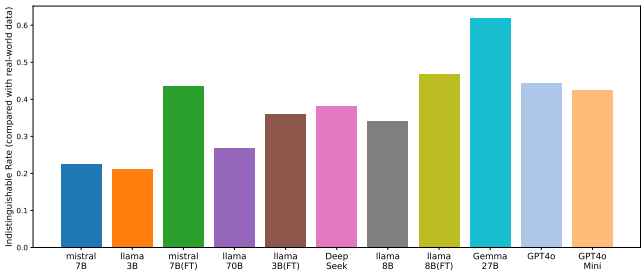
(a) OASST1 dataset: 6-turns



(b) OASST1 dataset: 12-turns



(c) Chatbot Arena dataset: 6-turns



(d) Chatbot Arena dataset: 12-turns

Figure 3: GTEval indistinguishability rates (with real data) of OASST1 dialogues. Models labelled with "FT" are fine-tuned.

conversation increases, focusing on UniEval passing rates and GTEval indistinguishability rates as the number of utterances grows from 6 to 12. The results are summarized in Table 3. Based on this data, we make the following observations:

1. **All models show decreased UniEval passing rates as dialogue length increases from 6 to 12 turns**. This trend suggests that maintaining human-likeness over longer conversations remains a common challenge for all evaluated LLMs. Among the models, GPT-4o, GPT-4o mini, Gemma-2-27B, and DeepSeek-V3 exhibit the highest robustness to increasing dialogue length. In contrast, Llama-3.3-70B shows the most significant decline, with a 43.83% reduction in passing rate – despite its larger size compared to smaller models like Llama-3.2-3B – highlighting that model scale alone does not guarantee better performance in long-form dialogue.

2. **GTEval indistinguishability rates exhibit much smaller changes with increasing dialogue length**. We believe this is because GTEval is already highly effective in detecting AI-generated conversations, even when the dialogues are relatively short.

This performance degradation can be attributed to several factors, including contextual drift, where models lose track of earlier conversation elements, and coherence decay, where responses become increasingly generic or disconnected from the established conversational thread. These phenomena are particularly present in smaller models, which are generally known to exhibit a limited capacity to maintain long-range dependencies across extended dialogue sequences.

	Base (%)	Finetuned (%)	Improvement (%)
Llama-3.2-3B	11.14|16.46	28.33|22.76	17.19|6.30
Llama-3.1-8B	20.58|25.67	44.79|30.02	24.21|4.35
Mistral-7B	16.95|17.68	29.06|19.61	12.11|1.93

(a) Improvement on the OASST1 dataset (GTEval indistinguishability rates). Each cell: 6 and 12 max-turns.

	Base (%)	Finetuned (%)	Improvement (%)
Llama-3.2-3B	25.85|21.16	49.90|36.03	24.05|14.87
Llama-3.1-8B	38.42|34.03	63.67|46.81	25.25|12.78
Mistral-7B	28.04|22.55	52.69|43.61	24.65|21.06

(b) Improvement on the Chatbot Arena dataset (GTEval indistinguishability rates). Each cell: 6 and 12 max-turns.

Table 2: GTEval indistinguishability rates (with real data) improvements after fine-tuning on (a) OASST1 and (b) Chatbot Arena datasets.

	UniEval (%)	GTEval (%)
Mistral-7B	-16.46	+0.73
Llama-3.2-3B	-30.75	+5.32
Mistral-7B (FT)	-24.94	-9.44
Llama-3.3-70B	-43.83	+4.84
Llama-3.2-3B (FT)	-25.18	-5.57
DeepSeek-V3	-19.13	0.0
Llama-3.1-8B	-45.76	+5.08
Llama-3.1-8B (FT)	-17.19	-14.77
Gemma-2-27B	-15.74	-6.05
GPT-4o	-6.05	-5.81
GPT-4o mini	-11.14	-1.69

Table 3: The UniEval passing rate and GTEval indistinguishability rate change (+: increase, -: reduction) as the number of utterances in the conversation grows from 6 to 12 (OASST1 dataset).

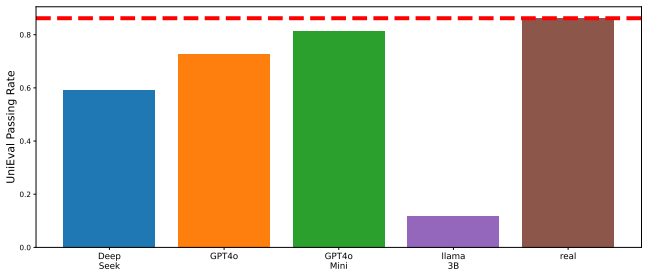


Figure 4: UniEval passing rate (no AI involved) of the OASST1 dataset (6-turns) with Claude 3.7 used as the LLM judge. The red dashed line indicates the passing rate of real human-chatbot conversations.

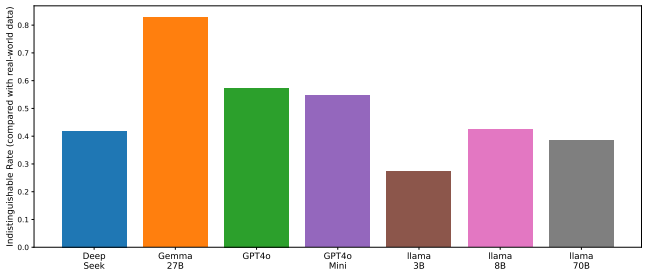


Figure 5: GTEval indistinguishability rate (compared with real data) of the OASST1 dataset (6-turns) with Gemini-2.0-Flash used as the LLM judge.

4.2.5 LLM judge Bias. In this project, GPT-4o is used as the default LLM judge. However, to investigate the possibility of model-specific bias, we conduct additional evaluations using alternative LLM judges: Claude 3.7 and Gemini 2.0 Flash. These models are used to assess dialogues generated by a subset of models. The corresponding results are presented in Figures 4 and 5.

Overall, the evaluation results from Claude 3.7 and Gemini 2.0 Flash are largely consistent with those obtained using GPT-4o. When Claude 3.7 is used as the judge LLM (Figure 4), GPT-4o demonstrates a high UniEval passing rate, indicating strong human-likeness. GPT-4o mini also performs well, with a passing rate comparable to that of real human-chatbot dialogues.

When using Gemini 2.0 Flash as the judge LLM (Figure 5), Gemma-2-27B achieves a surprisingly high GTEval indistinguishability rate. Nonetheless, both GPT-4o and GPT-4o mini continue to show strong performance, outperforming other large models such as DeepSeek-V3 and Llama-3.3-70B.

These results suggest that **GPT-4o does not exhibit a systematic bias in favor of its own outputs** and can be considered a reliable LLM judge for this evaluation framework.

5 Discussion

5.1 Failure Cases

In the process of dialogue generation, the inquirer LLM receives only the corresponding prompt as its instruction, resulting in a high degree of output freedom and introducing some unpredictability into the generation process. To prevent infinite turns of conversations, we imposed maximum turn limits for dialogues. The results of the experiments indicate that LLM-generated dialogues tend to have more turns than their real-world counterparts when given specific task summaries, and we have observed that many generated dialogues did not complete the assigned tasks within the given limits. Ideally, the generated dialogues should align closely with the predefined task summaries and follow a specified JSONL format. However, in practice, several unexpected failure cases emerged.

Topic deviation. Certain models, such as Llama-3.2-3B, occasionally generated dialogues entirely unrelated to the assigned task, indicating a significant deviation from the topic of the generation task and reflecting the instability of smaller models in dialogue generation.

Refusal to simulate human-like dialogue. We provide two prompts to the inquirer, which detail the generation task and instruct it to simulate conversations in human-style. Nevertheless, some models consistently declined to produce simulated human interactions, even under varied prompt settings, thereby halting the generation process.

Discrepancies in output quantity. Each prompt is derived from a real conversation from the source dataset, ensuring that the number of generated dialogues matches the number of dialogues in the dataset. However, some LLMs unpredictably produced fewer dialogues than expected, and this phenomenon appears to occur in a random and unexpected manner.

5.2 Limitations

Our framework enables the scalable simulation of human-chatbot dialogues, but several limitations persist. Firstly, while the tested LLMs are capable of generating fluent, grammatically human-like messages, they can fall short in producing goal-oriented dialogues and struggle to maintain long conversation. This applies to both large pre-trained LLMs and small LLMs improved by fine-tuning. As the number of turns increases, the models tend to exhibit contextual drift, loss of coherence by producing generic or irrelevant responses, and lack of task focus. Secondly, although evaluation with UniEval and GTEval provides useful insights, these metrics may not fully reflect the nuanced qualities of natural conversations, such as the target completion rate or emotional coherence. Additionally, our framework does not currently incorporate measures for toxicity or hallucinations, which may limit its applicability in scenarios requiring careful content moderation. Finally, synthetic dialogues, while fluent, may overfit the stylistic patterns of the models, reducing the diversity typically seen in real human interactions. These observations highlight the need for more nuanced evaluation methods and model strategies that go beyond AI-human-involved judgement.

6 Conclusion & Future Work

In this work, we present DialogueForge, a framework for generating human-like multi-turn dialogues using LLMs that simplifies the costly and time-intensive process of collecting human-chatbot dialogue data. Unlike existing approaches that focus on single generation paradigms, our framework combines in-context prompting,

iterative multi-agent simulation, and supervised fine-tuning to produce diverse, task-tailored multi-turn dialogues at scale. As part of our exploration, we employ a variety of both open-source (e.g. Llama, Mistral) and proprietary (e.g. GPT-4o) LLMs of various sizes. Our analysis reveals notable differences in human-likeness across models, with proprietary models generally exhibiting stronger coherence and realism, while open-source models offer promising performance with greater controllability and customization. Most significantly, we find that the performance gap between large and small models can be substantially reduced through fine-tuning, which improves the performance of smaller language models considerably.

The practical implications of DialogueForge extend beyond the technical contributions. By providing an accessible framework for synthetic dialogue generation, our work enables researchers with limited computational resources to contribute to conversational AI research, while publicly available code, models, and datasets lower barriers to entry for dialogue research. Our results demonstrate that fine-tuned smaller models can achieve competitive performance, challenging the assumption that only large proprietary models are viable for dialogue generation tasks. This finding has important implications for research reproducibility and the development of conversational AI systems in resource-constrained environments. Overall, we conclude that generating high-quality human-like conversation data using LLMs is possible and can provide satisfactory results depending on the choice of the LLM used to simulate the human inquirer. Despite certain limitations of the current results, such as struggles to maintain long conversations, we believe DialogueForge to be a strong contribution towards accessible, scalable dialogue data generation.

Moving forward, we plan to extend DialogueForge with more complex dialogue scenarios (e.g. persona conditioning) to better reflect real-world interactions and test model robustness under more diverse settings. To improve and directly assess models' ability to complete task-oriented dialogues, we also aim to move beyond AI-human-involved judgement by measuring actual goal achievement rates and functional coherence. Finally, we plan to thoroughly assess our evaluation metrics using human assessments, and employ hybrid evaluation that combines automatic metrics with lightweight human judgement. Such an approach can provide a more comprehensive and reliable assessment of dialogue quality, ultimately accelerating progress toward more natural and effective conversational AI systems.

References

[1] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models. arXiv:2311.16867 [cs.CL] https://arxiv.org/abs/2311.16867

[2] Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet/.

[3] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7421–7454. doi:10.18653/v1/2024.acl-long.401

[4] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. 2015. Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing* 28, 2 (2015), 2.

[5] Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting Language Models for Social Conversation Synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 844–868. doi:10.18653/v1/2023.findings-eacl.63

[6] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: an open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning* (Vienna, Austria) *(ICML'24)*. JMLR.org, Article 331, 30 pages.

[7] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168 [cs.LG] https://arxiv.org/abs/2110.14168

[8] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, (...), and Zizheng Pan. 2025. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL] https://arxiv.org/abs/2412.19437

[9] Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10602–10621. doi:10.18653/v1/2023.findings-emnlp.711

[10] Arnout Devos. 2024. *Few-shot Learning for Efficient and Effective Machine Learning Model Adaptation.* Ph. D. Dissertation. EPFL.

[11] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3029–3051. doi:10.18653/v1/2023.emnlp-main.183

[12] Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2023. BotChat: Evaluating LLMs' Capabilities of Having Multi-Turn Dialogues. arXiv:2310.13650 [cs.CL] https://arxiv.org/abs/2310.13650

[13] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as You Desire. arXiv:2302.04166 [cs.CL] https://arxiv.org/abs/2302.04166

[14] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858 [cs.CL] https://arxiv.org/abs/2209.07858

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, (...), and Zhiyu Ma. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY] https://arxiv.org/abs/2009.03300

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[18] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[19] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12930–12949. doi:10.18653/v1/2023.emnlp-main.799

[20] Minju Kim, Chaehyeong Kim, Yong Ho Song, Seung-won Hwang, and Jinyoung Yeo. 2022. BotsTalk: Machine-sourced Framework for Automatic Curation of

Large-scale Multi-skill Dialogue Datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5149–5170. doi:10.18653/v1/2022.emnlp-main.344

[21] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotirisr Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. OpenAssistant conversations - democratizing large language model alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2064, 13 pages.

[22] LangGraph. 2024. LangGraph Overview. https://langchain-ai.github.io/langgraph/.

[23] Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. PERSONACHATGEN: Generating Personalized Dialogues using GPT-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, Heuiseok Lim, Seungryong Kim, Yeonsoo Lee, Steve Lin, Paul Hongsuck Seo, Yumin Suh, Yoonna Jang, Jungwoo Lim, Yuna Hur, and Suhyune Son (Eds.). Association for Computational Linguistics, Gyeongju, Republic of Korea, 29–48. https://aclanthology.org/2022.ccgpk-1.4/

[24] Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022. Controllable Dialogue Simulation with In-context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4330–4347. doi:10.18653/v1/2022.findings-emnlp.318

[25] 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 3469–3483. doi:10.18653/v1/2021.acl-long.269

[26] Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive Learning from Complex Explanation Traces of GPT-4. arXiv:2306.02707 [cs.CL] https://arxiv.org/abs/2306.02707

[27] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, (...), and Yury Malkov. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276

[28] Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Data Augmentation for Conversational AI. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1234–1237. doi:10.1145/3589335.3641238

[29] Heydar Soudani, Roxana Petcu, Evangelos Kanoulas, and Faegheh Hasibi. 2024. A Survey on Recent Advances in Conversational Data Generation. arXiv:2405.13003 [cs.CL] https://arxiv.org/abs/2405.13003

[30] Hovhannes Tamoyan, Hendrik Schuff, and Iryna Gurevych. 2024. LLM Roleplay: Simulating Human-Chatbot Interaction. arXiv:2407.03974 [cs.CL] https://arxiv.org/abs/2407.03974

[31] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, (...), and Oriol Vinyals. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[32] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, (...), and Alek Andreev. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118 [cs.CL] https://arxiv.org/abs/2408.00118

[33] Bo-Hsiang Tseng, Yinpei Dai, Florian Kreyssig, and Bill Byrne. 2021. Transferable Dialogue Systems and User Simulators. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 152–166. doi:10.18653/v1/2021.acl-long.13

[34] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6268–6278. doi:10.18653/v1/2023.emnlp-main.385

[35] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2025. WizardLM: Empowering large pre-trained language models to follow complex instructions. arXiv:2304.12244 [cs.CL] https://arxiv.org/abs/2304.12244

[36] 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)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2204–2213. doi:10.18653/v1/P18-1205

[37] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. WildChat: 1M ChatGPT Interaction Logs in the Wild. arXiv:2405.01470 [cs.CL] https://arxiv.org/abs/2405.01470

[38] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. arXiv:2202.13047 [cs.CL] https://arxiv.org/abs/2202.13047

[39] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. arXiv:2309.11998 [cs.CL] https://arxiv.org/abs/2309.11998

[40] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]