Beyond Single-Turn: A Survey on Multi-Turn Interactions with Large Language Models

Yubo Li   Xiaobin Shen   Xinyu Yao[†]   Xueying Ding[†]   Yidi Miao[†]
Ramayya Krishnan   Rema Padman

Carnegie Mellon University
{yubol, xiaobins, xinyuyao, xding2, yidim, rk2x, rpadman}@andrew.cmu.edu

Abstract

Recent advancements in large language models (LLMs) have revolutionized their ability to handle single-turn tasks, yet real-world applications demand sophisticated multi-turn interactions. This survey provides a comprehensive review of recent advancements in evaluating and enhancing multi-turn interactions in LLMs. Focusing on task-specific scenarios—from instruction following in diverse domains such as math and coding to complex conversational engagements in roleplay, healthcare, education, and even adversarial jailbreak settings—we systematically examine the challenges of maintaining context, coherence, fairness, and responsiveness over prolonged dialogues. The paper organizes current benchmarks and datasets into coherent categories that reflect the evolving landscape of multi-turn dialogue evaluation. In addition, we review a range of enhancement methodologies under multi-turn settings, including model-centric strategies (contextual learning, supervised fine-tuning, reinforcement learning, and new architectures), external integration approaches (memory-augmented, retrieval-based methods, and knowledge graph), and agent-based techniques for collaborative interactions. Finally, we discuss open challenges and propose future directions for research to further advance the robustness and effectiveness of multi-turn interactions in LLMs. Related resources and papers are available at https://github.com/yubol-cmu/Awesome-Multi-Turn-LLMs.

[†]Equal contribution.

Contents

1 Introduction

The advent of Large Language Models (LLMs), exemplified by influential systems such as GPT series [1, 2, 3], PaLM [4], and LLaMA [5], has significantly reshaped numerous domains, from education and healthcare to customer service and software engineering. These powerful language models demonstrate remarkable proficiency in generating coherent and contextually relevant responses, achieving groundbreaking performances across various language understanding and generation benchmarks.

However, much of the early progress in both the evaluation and improvements of LLMs has been concentrated in single-turn interactions, where models are tested on isolated prompts without considering prior conversational context. While this has led to impressive performance on a range of benchmarks, it overlooks the broader potential of LLMs in multi-turn dialogue, which more accurately reflects real-world usage. In practice, real-world scenarios rarely consist of isolated queries; instead, meaningful and productive interactions usually occur through continuous, multi-turn exchanges. Effective communication between humans and artificial intelligence inherently demands an understanding of conversational history, nuanced interpretation of previous exchanges, iterative refinement of goals, and adaptive response strategies. Single-turn interaction thus presents a significant limitation, restricting the deployment and utilization of the robust capabilities that LLMs possess.

Recognizing this crucial limitation, substantial research attention has recently shifted toward multi-turn interactions, focusing on enhancing LLM capabilities to sustain context, maintain consistency, handle ambiguity, and dynamically respond across sequential conversational turns. Multi-turn interactions introduce additional layers of complexity, such as managing dialogue coherence, maintaining alignment with user intentions, and addressing issues like cumulative errors, hallucinations, and contextual drift.

This emerging field presents rich opportunities and considerable challenges, prompting a rapidly expanding body of research dedicated to optimizing, evaluating, and deploying LLMs within multi-turn conversational settings. Understanding these multi-turn dynamics and systematically addressing their inherent complexities is essential for unlocking the next stage in LLM evolution, thereby significantly expanding their applicability and effectiveness in real-world scenarios.

Scope To provide clarity and facilitate further advancements, this survey specifically addresses multi-turn interaction with LLMs, categorized by tasks along with corresponding evaluation standards, improvement techniques, and current challenges in this domain. We categorize multi-turn interactions according to two main tasks: instruction following tasks and conversational engagement tasks, with the latter encompassing various real-world domain applications. While we refer to several LLM agent-based approaches, we emphasize that LLM-based agents represent a distinct research direction that, though related, falls outside our primary scope. Similarly, despite the growing body of research in Multi-modal LLMs (MLLMs) - including notable benchmarks and frameworks such as ConvBench [6], MMDU [7], MMMT-IF [8], MMDialog [9], TheaterGen [10], and SVBench [11] - this survey deliberately excludes multimodal capabilities to maintain a focused analysis of text-based multi-turn interactions. Such boundary allows us to provide a comprehensive exploration and in-depth analysis of the landscape of multi-turn interactions specific to LLMs.

Existing Surveys Foundational surveys [12, 13] provide rigorous analyses of LLMs' ability to follow instructions under a single-turn setting. These works evaluate metrics like faithfulness, robustness, and generalization in static, one-off interactions. While they establish critical baselines for instruction alignment, their scope excludes the dynamic, context-dependent challenges inherent to multi-turn scenarios, such as coherence across turns, state tracking, and user intent adaptation.

Several dialogue system surveys are also particularly relevant: [14] present a overview of recent advances in multi-turn dialogue systems based on LLMs; healthcare-specific dialogue systems are examined by surveys [15, 16]. Additionally, [17] specifically explores multi-turn interaction capabilities of LLMs, with a strong focus on core model abilities, evaluation metrics, and algorithmic enhancements for multi-turn contexts.

Comparison with Previous Surveys   While existing instruction-following surveys predominantly address single-turn scenarios, and dialogue system surveys excel in targeted evaluations and performance considerations within their respective scopes, a clear gap emerges regarding a structured and holistic comparison under broader, multi-turn interaction settings.

Among current literature, Zhang et al. (2025)'s work [17] aligns most closely with our work. While it focuses primarily on LLM core multi-turn interaction capabilities, corresponding evaluation methods, and algorithmic improvements, their approach is fundamentally capability-oriented rather than task-oriented. In contrast, we recognize that multi-turn settings more closely resemble real-world applications, where performance emerges from the complex collaboration of multiple capabilities rather than isolated ones. Therefore, we deliberately categorize interactions by tasks instead of capabilities, offering a more practical taxonomy. Furthermore, Zhang et al. (2025) [17]'s survey lacks substantial discussion of multi-turn interactions in critical real-world scenarios, such as healthcare consultations and educational assistance, where multi-turn dynamics are essential. Their survey also presents limited analysis of improvement methodologies and future challenges in multi-turn contexts.

Main Contributions   To address these gaps and facilitate greater research efforts in multi-turn LLM interactions, this survey presents a structured and detailed analysis that explicitly considers practical scenarios and characteristics of multi-turn deployments. We categorize multi-turn interactions by task, exploring both mixed-topic instruction-following tasks (§2.1) and more complex, open-ended conversational engagement tasks (§2.2) across several key domains where LLMs have demonstrated substantial impacts.

Beyond categorization, we contribute by detailing improvement methodologies across three crucial dimensions: (1) Model-Centric Approaches, directly refining and adapting LLMs to effectively handle sequential dialogue dynamics through strategies like in-context learning, supervised fine-tuning, reinforcement learning, and innovative architectures (§3.1); (2) External Integration Approaches, enhancing LLM performance by leveraging external resources such as memory structures, retrieval mechanisms, and knowledge graphs to overcome contextual limitations and maintain factual consistency (§3.2); and (3) Agent-Based Approaches, representing a paradigm shift toward proactive, iterative agents that interact individually or collaboratively, managing complexity and improving reasoning capabilities in extended interactions (§3.3).

Additionally, we thoroughly discuss open challenges (§4), proposing a clear taxonomy that categorizes these challenges into five major areas: Context Understanding, Complex Reasoning, Adaptation & Learning, Evaluations, and Ethical & Safety Issues. Finally, we summarize key insights, reflect on overarching themes, and provide perspectives on future directions in a dedicated conclusion (§5).

To the best of our knowledge, this survey is the first to provide an extensive landscape of multi-turn LLM interactions, covering tasks, real-world applications, evaluation methods, improvement strategies, and critical open challenges.

Date[*]	Benchmark/Dataset	Category	Dataset Size		Data Curation		Evaluator				Evaluation Criteria
			Total #Dial.	Avg. #Turns/Dial.	Human-based	LLM-based	Rule-based	Human-as-a-judge[†]	LLM-as-a-judge	Agreement Check[‡]	
Mar 2022	MTPB [18]	Coding	115	/	no	no	yes	no	no	/	Pass rate, Perplexity
Dec 2022	ABC-Eval § [19]	Conversation	400	30.3	yes	no	no	yes	no	/	Consistency, Emotion, Understanding, Engagingness, Grammaticality, Informativeness, Quality, Proactivity, Relevance
Jun 2023	MathChat-Agent [20]	IF-math	/	15	no	no	yes	no	no	/	Accuracy
Jun 2023	MT-Bench [21]	IF-general	80	2	yes	no	no	no	yes	yes	Correctness, Helpfulness
Jun 2023	InterCode [22]	IF-coding	/	10	no	yes	yes	no	no	/	Execution Success
Sep 2023	MINT [23]	IF-reasoning	586	< 5	no	yes	yes	no	no	yes	Evaluation Quality, Success Rate
Oct 2023	MT-Bench++ [24]	IF-general	80	8	no	yes	no	no	yes	no	Helpfulness, Relevance, Accuracy, Depth, Creativity, Level of Detail
Oct 2023	BotChat [25]	Conversation	547	16	yes	yes	no	no	yes	/	Quality, Similarity to Human Dialogues
Jan 2024	EHRAgent [26]	IF-coding	/	/	yes	no	yes	no	no	/	Success Rate, Complete Rate
Jan 2024	MT-Eval [27]	IF-general	168	6.96	no	yes	yes	no	yes	no	Helpfulness, relevance, accuracy, depth, creativity, how well it conforms to constraints; accuracy of classification and recollection
Jan 2024	WebLINX [28]	IF-general	2337	43	yes	no	yes	no	no	/	Intent Match, Element Similarity Using IoU, Text Similarity Using F1, Turn-Level Score and Overall Score
Feb 2024	AQA-Bench [29]	IF-reasoning	/	15,30¶	/	/	yes	no	no	/	Goal Metric, Policy Metric (Both for Each Interactive Environment)
Feb 2024	MT-Bench-101 [30]	IF-general	1388	3.03	no	yes	no	no	yes	yes	Perceptivity (Context Memory, Context Understanding, Context Interference), Adaptability (Rephrasing, Reflection, Reasoning), Interactivity (Questioning)
May 2024	MathChat-Bench [31]	IF-math	5276	4	no	yes	no	no	yes	yes	Follow-Up QA, Error Correction, Error Analysis, Problem Generation
Jun 2024	M2Lingual ‖ [32]	IF-general	1140	2.48	no	yes	*MT-Bench*				*same as MT-Bench's*
Aug 2024	SysBench [33]	IF-general	500	5	yes	no	yes	no	no	/	Constraint Satisfaction Rate, Instruction Satisfaction Rate, Session Stability Rate
Oct 2024	FB-Bench [34]	IF-general	591	2	yes	yes	no	no	yes	yes	Error Correction, Response Maintenance
Oct 2024	WILT [35]	IF-reasoning	50	< 30	/	/	yes	no	no	/	Accuracy in Deducing the Hidden Functions
Oct 2024	Multi-IF [36]	IF-general	909	3	no	yes	yes	no	no	/	Instruction- and Conversation-Level Accuracy of Instruction Following
Oct 2024	FairMT-Bench [37]	IF-fairness	10k	4	no	yes	no	no	yes	yes	Direct and Implicit Bias
Dec 2024	MMSQL [38]	IF-coding	6493	6	no	yes	yes	no	yes	yes	Exact Matching, Execution Accuracy, Dual Assessment of Question Type Detection, Execution Accuracy, Response Quality Score

Table 1: An Overview of Recent Multi-Turn Instruction Following Bechmark or Datasets and Evaluation Methods.

---

*We use the date of their initial version on arXiv to rank them, allowing the development trajectory and patterns to be revealed.

†Do not include the human evaluation done for agreement check between LLM and human.

‡If LLM-as-a-judge is used, whether there is an agreement check between LLM and human annotators on a subset of the dataset.

§We intentionally include ABC-Eval and BotChat in this table as they represent early explorations in multi-turn interactions. Their presence illustrates the developmental trajectory of the field, although a more detailed discussion can be found in §2.2.

¶15 for easy mode, 30 for hard mode.

‖M2Lingual includes dialogues for 70+ languages. The statistics shown here are restricted to only English multi-turn dialogues.

2 Multi-Turn Interaction Tasks

In this survey, we categorize multi-turn interactions based on tasks rather than capabilities because real-world multi-turn scenarios inherently require integrating multiple capabilities to accomplish a particular objective. Although capabilities such as reasoning, memory, contextual understanding, or adaptability are critical, they rarely function in isolation within conversational contexts. Instead, different capabilities intertwine and complement each other dynamically to achieve specific task goals. By focusing on clearly defined tasks—such as multi-turn instruction-following or conversational engagement—we better reflect the realistic complexities of these interactions, offering readers an intuitive framework for understanding how LLMs practically meet user requirements and perform in real-world multi-turn conversational scenarios. Table 1 provides an overview of recent multi-turn instruction-following benchmarks or datasets and their evaluation methods, offering readers a detailed reference for understanding current trends and practices in multi-turn LLM evaluation.

Before diving into detailed discussions on multi-turn instruction-following and conversational engagement tasks, we briefly clarify our rationale behind this categorization. We distinguish these tasks primarily based on two dimensions: user intention clarity and task complexity. Specifically, multi-turn instruction-following tasks typically involve clear, explicit instructions provided by the user, with well-defined user intentions. Consequently, the LLM's performance in these tasks is evaluated directly by how precisely it adheres to or successfully executes the given instructions. Conversely, multi-turn conversational engagement tasks are often characterized by more open-ended user interactions, in which the user's intention may initially be unclear, partially defined, or dynamically evolving throughout the dialogue. The LLM's role in these interactions generally extends beyond strict compliance with instructions; it assumes roles such as an assistant or consultant—examples include health consultants, teaching assistants, or customer service representatives. Such tasks may involve proactive information-seeking, synthesizing insights across multiple topics, interpreting implied user intentions, and utilizing external knowledge bases or tool calls to curate relevant and contextually appropriate responses throughout multi-turn conversations.

2.1 Instruction Following Tasks

In this section, we specifically focus on multi-turn instruction-following tasks, highlighting their unique characteristics, challenges, and recent benchmark developments. Inspired by task categorization of MT-Bench [21], We group existing multi-turn instruction-following benchmarks into three primary categories: general (mixed) instruction-following, mathematics, and coding. Most benchmarks naturally fall into the general category due to their coverage of diverse task types. Nevertheless, specialized benchmarks targeting math and coding interactions have also emerged prominently, demonstrating unique evaluation dimensions and challenges distinct from general-purpose tasks. Table 1 summarizes these benchmarks, including their evaluation methods, scope, and distinguishing attributes.

2.1.1 Instruction Following Tasks in General

Recent research has increasingly focused on evaluating the multi-turn interaction abilities of large language models (LLMs), aiming to capture the complexities of real-world dialogue that single-turn benchmarks, such as BIG-Bench [39], CSQA [40], MMLU [41], and GSM8K [42], often fail to address.

MT-Bench [21] emerged as one of the first curated benchmarks designed to evaluate multi-turn instruction following capabilities in LLMs. It consists of 80 two-turn dialogues that span eight categories: writing, roleplay, information extraction, reasoning, math, coding, STEM knowledge, and social science. The benchmark evaluates model responses through pairwise comparisons, assessing correctness and helpfulness. A key contribution of MT-Bench (along with AlpacaEval [43]) is its systematic study of **LLM-as-a-judge**, where strong LLMs, such as GPT-4, serve as automated evaluators. The study shows that LLM judges achieve over 80% agreement with human evaluators, making them a scalable alternative to human assessments. It also examines potential biases and limitations of LLM judges and suggests mitigation strategies. Since its introduction, LLM-as-a-judge has become a widely adopted evaluation method, shaping benchmarking practices. However, MT-Bench is limited to two-turn dialogues and a relatively small dataset, underscoring the need for more comprehensive benchmarks.

Building on this foundation, MT-Bench++ [24] extends the original MT-Bench dataset by incorporating additional follow-up turns per dialogue, resulting in eight-turn interactions. These extended interactions, enriched with carefully crafted ellipsis and anaphora, further challenge models' abilities to maintain context and coherence over prolonged exchanges.

Increasing dialogue length alone was insufficient for comprehensive evaluation. MT-Bench-101 [30] introduces a fine-grained taxonomy for multi-turn dialogues. It categorizes interactions into aspects such as perceptivity (context understanding and memory), interactivity (eliciting clarifying questions), and adaptability (reasoning and reflection). With 4,208 turns across 1,388 dialogues and a three-tier ability taxonomy, this benchmark facilitates a detailed assessment of specific interaction skills. Evaluations on 21 prominent LLMs have revealed that even state-of-the-art chat-tuned models exhibit uneven performance across different turns and task types, and that standard alignment techniques (e.g., supervised fine-tuning and RLHF) do not guarantee consistent improvements in multi-turn settings.

MT-Eval [27] takes a different approach by investigating the performance disparities between single-turn and multi-turn interactions. Utilizing 1,170 multi-turn queries derived from human-chat transcripts, MT-Eval categorizes user tasks into four distinct types: follow-up (building upon previous responses), refinement (modifying prior requests), expansion (elaborating on earlier topics), and recollection (retrieving information from earlier turns). The study demonstrates that most LLMs suffer significant performance degradation in multi-turn scenarios, with errors compounding over successive exchanges and the temporal distance from the relevant context further exacerbating the decline This interaction taxonomy has since been adopted by subsequent research, notably M2Lingual [32], which extends multi-turn evaluation into the multilingual domain. M2Lingual applies these interaction categories across 12 diverse languages, revealing concerning cross-lingual brittleness in context retention abilities. The benchmark demonstrates that even advanced models struggle to maintain contextual understanding across language boundaries, with performance deteriorating more severely in non-English interactions, highlighting a critical gap in the multilingual capabilities of current LLMs for sustained dialogues.

M2Lingual is not the only research highlighting such cross-lingual challenges. Multi-IF [36], a benchmark for multi-turn and multilingual instructions, introduces a challenging evaluation set for LLMs involving multi-turn, verifiable writing instructions (e.g., style or format requirements) across 8 languages. The Multi-IF dataset contains 4,501 dialogues created by expanding a single-turn English benchmark (IFEval [12]) into more challenging multi-turn sequences and translating them into 7 additional languages. Concurrently, FairMT-Bench [37] focuses explicitly on fairness in multi-turn dialogues. FairMT-Bench is the first comprehensive benchmark designed to evaluate fairness in open-domain multi-turn dialogues for LLMs, formulating a task taxonomy that targets LLM fairness capabilities across three stages: context understanding, user interaction, and instruction trade-offs—challenges discussed further in Section 4. Based on this taxonomy, the authors constructed fairness datasets, FairMT-1K and FairMT-10K, encompassing two major bias types (stereotype and toxicity) and six bias attributes (age, gender, race, religion, disabled, and appearance), covering nearly all bias categories commonly addressed in fairness evaluation.

While most multi-turn benchmarks focus on context retention and reasoning across turns, FB-Bench [34] introduces a novel dimension by measuring LLMs' responsiveness to human feedback in multi-turn interaction settings. This Chinese-language benchmark evaluates two crucial aspects of feedback handling: error correction (the ability to fix mistakes when prompted) and response maintenance (the ability to maintain correct responses when challenged). FB-Bench spans diverse task categories, including mathematics, reasoning, coding, text extraction, text error correction, text creation, knowledge Q&A, and text translation. Findings from this benchmark reveal that leading LLMs demonstrate comparable capabilities in error correction across tasks, but their performance varies significantly in response maintenance. Moreover, the research indicates that hinting guidance substantially improves response quality, while exposure to misinformation or fabricated credentials often results in misleading outputs.

Li et al. (2025) [44] recently drew attention to the challenge of maintaining consistent responses in multi-turn LLM interactions, introducing a framework that assumes models should remain "firm" rather than "fickle" when faced with challenging follow-up prompts. Their work introduces the Position-Weighted Consistency (PWC) score, a novel metric that considers the temporal dimension of dialogue by assigning greater penalties

to inconsistencies occurring in earlier interaction rounds. This approach reflects the real-world importance of early stability in establishing user trust. Through experiments with leading models across carefully designed challenge scenarios, the authors demonstrated that even high-performing LLMs can be swayed by various follow-up strategies, including emotional appeals and expert authority claims. To enhance response stability, they propose the Confidence-Aware Response Generation (CARG) framework, which integrates model confidence signals into the generation process. Empirical results demonstrate that CARG significantly improves response consistency without compromising accuracy, highlighting the necessity of incorporating turn-based considerations into LLM evaluations.

In addition to task-oriented evaluations, a complementary line of work has explored abstract reasoning benchmarks designed to test LLMs' core reasoning capabilities in multi-turn settings. AQA-Bench [29] evaluates LLMs in interactive environments requiring sequential decision-making, such as binary search and graph traversal (DFS, BFS). It emphasizes memory maintenance, procedural adherence, and planning over multiple turns, with both algorithmic and "embodied" (narrative) settings. WILT (Wason Inductive Logic Test) [35] by Banatt et al. (2024) further assesses inductive reasoning by asking models to iteratively discover hidden rules through evidence gathering and hypothesis testing. These benchmarks are task-agnostic and explicitly designed to avoid memorization, targeting core skills such as logical consistency, strategic exploration, and hypothesis refinement.

Besides these abstract reasoning evaluations, other studies have addressed multi-turn interactions from specialized task perspectives. For instance, WebLINX [28] tackles the problem of conversational web navigation, where an LLM-based agent must follow user instructions via dialogue to accomplish tasks on real websites. Tasks range from booking tickets to finding information, requiring the agent to understand natural language commands and manipulate a web page accordingly (click links, fill forms, etc.) over multiple turns. MULTITURNINSTRUCT [45] proposed a systematic benchmark to probe LLMs' ability to handle sequential, potentially conflicting instructions in a conversation. SysBench [33] is a benchmark specifically designed to evaluate how well LLMs adhere to system-level instructions (the hidden directives guiding an AI assistant) in multi-turn interactions. Emphasizing clear and explicit user intentions, we categorize two recent recommendation system works in this subsection: SAPIENT [46] and ECR [47]. SAPIENT [46] is a multi-turn conversational recommender system that integrates a planning module for strategic dialogue management, while ECR [47] introduces an "empathetic"conversational recommender that enhances traditional recommendation dialogues with emotional awareness.

Together, these works underscore ongoing challenges in developing robust multi-turn instruction-following interactions in LLMs, particularly regarding context retention, dialogue coherence, multilingual interactions, fairness considerations, and responsiveness to user feedback. To effectively address the nuanced demands of specialized domains such as mathematics and coding, recent research has introduced targeted benchmarks and approaches. The following subsections explore these specialized areas, highlighting how they deepen our understanding of multi-turn instruction-following tasks.

2.1.2 Instruction Following Tasks in Math

LLMs have demonstrated impressive performance in solving math problems via single-turn prompts, often generating detailed, step-by-step "chain-of-thought" solutions. For example, providing a few worked examples allows a 540B-parameter model to achieve state-of-the-art accuracy on the GSM8K math word problems [48]. However, complex mathematical tasks frequently require LLMs to engage in multi-turn, instruction-following interactions that involve incremental reasoning, clarification questions, and iterative refinement based on interactive feedback [31, 49]. In these instruction-following math tasks, users iteratively guide LLMs by providing clarifications, corrections, or additional contextual instructions. Such dynamic interactions not only enhance the models' reasoning processes but also facilitate effective use of external computational tools, enabling the models to perform sophisticated tasks like simulating scenarios or executing complex calculations through code [49, 50]. Instruction-following tasks in math also encompass advanced skills such as follow-up questioning, errors diagnosing and correcting, and educational feedback delivering, thus capturing broader capabilities essential for deploying LLMs in diverse educational and problem-solving contexts.

Several approaches have leveraged multi-turn dialogue with LLMs to enhance mathematical reasoning. For example, Wu et al. (2024) [20] introduced MathChat-Agent, a framework where an LLM collaborates with a user-proxy agent (responsible for tool use, such as a Python solver) via iterative conversation. This approach solved competition-level math problems more effectively, improving accuracy by roughly 6% over standard single-turn chain-of-thought prompts. Similarly, Keating (2024) [51] employed a multi-agent strategy in which two GPT-4 agents engage in debate-style interactions to reach a solution. This dual-agent zero-shot method achieved about 62.7% accuracy on the MATH benchmark—surpassing single-agent baselines and illustrating the benefits of peer deliberation in reasoning. Moving further, Xiong et al. (2024) [52] propose a method to train LLMs to better combine tool usage with their own reasoning for complex tasks. It gathers trajectory-level feedback from users over multiple turns, to refine problem-solving strategies. The learning process is modeled as a Markov decision process (MDP) and adapts direct preference learning algorithms to multi-turn interactions that include external messages. In practice, the model is trained on multi-turn chats where a user first asks a question and then gives Python outputs in later turns. The accuracy of the method is evaluated on the GSM8K and MATH test sets.

Benchmarks & Evaluation in Math   Several benchmarks now target multi-turn math dialogues. MathChat-Bench [31] extends GSM8K with four new tasks: follow-up QA, problem generation, error correction, and error analysis. The original problems are modified using GPT-4 to meet specific requirements. For instance, in the follow-up QA task, three rounds of dialogue are created: first, GSM8K test problems with ground-truth answers are presented; then GPT-4o generates two follow-up questions; finally, GPT-4 produces the final answers, which are verified or revised by two other LLMs and human annotators. LLMs are evaluated by both accuracy and scores, assigned by GPT-4, which measure their instruction follow-up abilities. Evaluations of state-of-the-art models on MathChat showed that while they excel at standard one-shot math questions, performance drops sharply on these multi-turn interactions requiring sustained reasoning and dialogue comprehension. To address this gap, the authors also released MathChatSync, a synthetic dialogue dataset for fine-tuning LLMs on conversational math problem solving. Fine-tuning on MathChatSync yielded notable improvements in multi-turn performance.

More broadly, the MINT benchmark [23] evaluates multi-turn tool use and user feedback across domains (including math reasoning). MINT provides an automated framework where the LLM can call a Python interpreter and receive natural language feedback (simulated by GPT-4) in successive turns. Findings from MINT show that multi-turn tool-aided dialogues consistently improve problem-solving success (each tool use or feedback turn yields additional 1–17% accuracy gains). Interestingly, this evaluation also found that some models fine-tuned only on single-turn instructions (via standard supervised tuning or RLHF) underperform in multi-turn settings, suggesting that multi-turn-specific training is needed to excel in interactive math tasks.

Noticeably, several studies have found that LLMs struggle with generalizing to new problems. For example, Liang et al. (2024) [31] show that math-specific LLMs lack adaptive behavior, and their difficulty with generating novel problems highlights their rigidity. Similarly, Macina et al. (2023) [53] report that dialogue tutoring models do not generalize well to unseen math problems. To improve multi-turn math problem solving and generalization, many researchers propose using SFT [31, 53, 23] or RLHF [23, 52]. Although RLHF can affect LLM–tool interactions and leveraging feedbacks [23], studies consistently find that fine-tuning with preference data and instructions boosts downstream performance. Macina et al. (2023) [53] demonstrate that small, fine-tuned models perform significantly better, in terms of correctness and equitable tutoring, than prompting a large model like ChatGPT.

2.1.3   Instruction Following Tasks in Coding

LLMs often struggle to produce correct code and perform self-debugging in a single pass, frequently requiring multi-turn or iterative interactions [54, 55, 56]. Instruction-following tasks in coding contexts commonly involve collaborative, iterative interactions, where users provide detailed instructions that the LLM translates into executable code. Through subsequent turns, the LLM iteratively integrates feedback, refines initial solutions, strategically plans modifications, executes and tests submodules, and performs debugging until satisfying the given specifications. Such iterative instruction-following tasks are essential to realistically evaluate LLM performance and robustness in dynamic coding scenarios.

Benchmarks & Evaluation in Coding   Different frameworks and playgrounds are developed for evaluaing code generation qualities [22, 57]. InterCode [22] introduces a generation pipeline for evaluating LLM coding quality through multi-turn interactions that simulate a real-world coding environment. The InterCode framework requires as input a natural language prompt paired with either an answer or a correct code block. The LLM is evaluated using one of three strategies: "Try-again", where the execution output is fed back as an observation; ReACT, which terminates once the thought chain is complete; or Plan & Solve, which terminates when the plan is fully executed. The experiments involve three programming environments with Spider [58], MBPP [59] and NL2Bash [60] datasets. Zheng et al. (2024) [57] propose a framework for systematically evaluating various prompting techniques for multi-turn code generation by LLMs. Their evaluation is conducted in a zero-shot setting using two competitive coding benchmarks, CodeContests [61] and TACO [62]. PyBench [63] proposes a unified benchmark to evaluate LLM Python coding ability in several categories such as chart analysis, software development, etc. For each task, LLMs are interacting with a code interpreter for a few turns before making a formal response. The LLMs are evaluated on the success rates and average turns to complete a unit test for each task.

Toward effective code generations, CodeGEN [18] and CodeGEN2 [64] are a family of LLMs designed to generate programs from natural language descriptions in multiple turns. The models are evaluated using the Pass Rate metric on their Multi-Turn Programming Benchmark (MTPB), which comprises 115 expert-written problems. Each problem includes multi-step natural language prompts, created by human annotators who decompose the problem into sequential steps. Models are required to generate the complete solution from scratch. Chen et al. (2025) [65] propose to fine-tune existing LLMs with SFT and DPO. They introduce CodeSteer, a framework that guides LLMs through multiple rounds of interaction to generate code. In this system, the primary model (TaskLLM) produces responses, both in natural language and in code, while a supervisory agent (CodeSteerLLM) reviews these outputs using symbolic reasoning and self-answer checking to ensure correctness and provide refined guidance. They are fine-tuned and evaluated on subsets of SymBench, which comprises 37 symbolic tasks with adjustable complexity and includes a synthesized dataset of multi-round guidance/generation trajectories and guidance comparison pairs. OpenCodeInterpreter [66] proposes to fine-tune LLMs with CodeFeedback, a dataset of challenging LeetCode questions, incorporating multi-turn execution feedback (with code interpreters) and dialogues of human (synthesized by GPT-4). CodeAct [67] collect an instruction-tuning dataset, CodeActInstruct, which contains 7,000 multi-turn interactions. PyInstruct [63] is used in PyBench for continuous pretraining and fine-tuning. Zheng et al. (2024) [57] explores a wide range of prompting strategies for effective code generation, focusing on automatic re-prompting over multiple runs.

SQL generation is a subcategory of coding generation that focuses more on data acquisition through large-scale data warehouses. Two benchmarks for SQL generation multi-turn LLMs are identified during the literature search. MMSQL [38] focuses on text-to-SQL generation tasks and introduces a Multi-type and Multi-turn text to-SQL test suite, which is a comprehensive benchmark engineered to evaluate the proficiency of LLMs in handling multi-turn text-to-SQL tasks across diverse question types. MMSQL contains a multi-agent framework anchored by a core Question Detector and Question Decomposer tasked with identifying question types and determining appropriate answering strategies. The framework includes two supportive agents: the Schema Selector, which identifies and provides the essential subset of a database schema, and the SQL Refiner, which is dedicated to refining SQL queries. EHRAgent [26] demonstrates a specific application of SQL generation in healthcare settings, which speeds up the extraction and interaction of clinician information within electronic health record (EHR) systems. EHRAgent [26] translates EHR question-answering into a tool-use planning process, which integrates query-specific medical information and formulates executable code plans through multi-turn dialogues. The model is evaluated based on their ability to reason across multiple tables and generate accurate, actionable insights from complex EHR data.

For education applications, TreeInstruct [68] transforms LLMs into instructor agents that guide users in debugging and writing better code. Acting as a Socratic educator, it plays two roles: the instructor generates tree-based sequential questions, and the verifier identifies tasks to help students understand, assess, and correct their code. To evaluate TreeInstruct, the authors created MULTIDEBUG, a dataset derived from popular LeetCode problems with expert-injected syntactical and conceptual bugs. Evaluation measures include both qualitative factors (relevance, indirectness, logical flow) and quantitative metrics (success rate).

Several compelling research questions arise as common themes in adapting LLMs for effective code generation through multi-turn interactions. For instance, how can an LLM decide whether to employ textual reasoning or programmatic solutions when explicit cues are absent [69, 55]? What iterative interaction protocols enable the model to refine its solution before submitting a final answer [22, 57, 18, 64, 67]? And how should we evaluate LLMs' coding performance across different programming languages and tasks in diverse domains [22, 63]?

2.1.4   Discussions

Based on our analysis and the overview presented in Table 1, several noteworthy trends emerge in the evolution of multi-turn instruction following benchmarks and evaluation methodologies.

Dataset Evolution   We observe a clear trajectory toward larger and more comprehensive datasets, with benchmark sizes expanding from just 80 examples in MT-bench to over 1,000 in newer benchmarks like MT-Bench 101, M2Lingual, and FairMT-Bench. This growth reflects a recognition that robust evaluation requires broader coverage of interaction patterns and abilities. Simultaneously, data curation methodologies have evolved from primarily human-generated content toward automated and LLM-assisted generation processes, addressing scalability challenges while maintaining quality. Despite this expansion, current benchmarks predominantly limit conversations to 10 or fewer turns, leaving the domain of extended multi-turn interactions (dozens or hundreds of turns) largely unexplored.

Evaluation Methodologies   The evaluation landscape has diversified considerably, transitioning from relatively simple rule-based metrics toward more nuanced and fine-grained assessment frameworks. This evolution parallels the increasing sophistication of LLM capabilities and application scenarios. LLM-as-judge approaches have emerged as particularly promising, offering cost-effective evaluation solutions that can scale with the growing complexity of benchmarks. MT-bench [21] pioneered this approach while acknowledging inherent biases, subsequent studies have further elucidated critical issues.

Recent studies highlight several concerning limitations: Preference Leakage, as demonstrated by Li et al. (2025) [70], shows bias toward responses from models sharing architectural or training lineage with the judge model, compromising evaluation fairness. Contextual Sensitivity, revealed by Xu et al. (2025) [71], manifests as performance degradation when evaluating outputs dependent on external context, such as retrieval-augmented generation, where even state-of-the-art judges struggle with consistency. Reference Dependence is another issue, as evaluations often exhibit brittleness when reference solutions are unavailable or when multiple valid approaches exist, a common scenario in open-ended multi-turn interactions.

These challenges underscore the continued importance of human evaluation validation. Human-AI agreement checks are necessary to establish the reliability of automated evaluation methods, yet our analysis reveals that relatively few benchmarks incorporate substantial human agreement verification, representing a critical weakness in current evaluation frameworks, potentially allowing biases and inconsistencies to persist undetected.

2.2   Conversational Engagement

Conversational engagement tasks in multi-turn dialogue have been catalyzed by several pioneering studies that established how to define and measure an LLM's capacity to sustain interactions over multiple turns. One early effort is the ABC-Eval framework by Finch et al. (2023) [19], which introduced a dimensional human evaluation scheme for open-domain chat. ABC-Eval defines 16 fine-grained conversational behavior categories (spanning aspects like factual accuracy, consistency, relevance, and empathy) and uses these as binary turn-level labels to quantify dialogue quality. Although this benchmark relies on labor-intensive human evaluations, its initial findings underscore the inherent challenges in assessing iterative interactions, thereby motivating the development of more scalable and nuanced evaluation frameworks.

Shortly afterward, Duan et al. (2023) [25] introduced the BotChat evaluation paradigm, which reduces reliance on costly human judges by leveraging LLMs for both conversation generation and evaluation. In BotChat, models are prompted to extend real-world dialogue seeds (ChatSEED prompts) into extended multi-turn

conversations, subsequently evaluated by top-tier LLMs (e.g., GPT-4) serving as automated judges. Notably, GPT-4-generated dialogues were found to be nearly indistinguishable from human conversations, successfully fooling discriminator models, whereas other contemporary LLMs exhibited shortcomings in instruction adherence and conciseness, underscoring specific challenges in maintaining human-like coherence across multi-turn dialogues.

Most recently, Sirdeshmukh et al. (2025) [72] introduced MultiChallenge, a benchmark designed to rigorously test conversational persistence and context management in frontier LLMs. MultiChallenge comprises four realistic scenarios—long-term instruction retention, implicit information recall, iterative revision, and consistent responses without sycophancy—requiring simultaneous instruction-following, context tracking, and reasoning.

Building upon these frameworks, recent work explores conversational engagement within specialized real-world contexts, including immersive role-playing, healthcare consultations, educational interactions, and adversarial jailbreak scenarios. Extending evaluations into these practical domains provides deeper insights into several key applications of conversational capabilities essential for modern LLMs.

2.2.1   Conversational Engagements in Roleplay

Role-playing significantly enhances conversational engagement by immersing users in specific scenarios, making interactions feel authentic and contextually relevant. Incorporating explicit roles into dialogue systems encourages users to perceive the interactions as genuine, thereby increasing engagement and satisfaction. Within conversational AI, there is an emerging research domain specifically dedicated to role-playing, which aims to create realistic and persona-consistent interactions through LLMs.

Early persona-grounded dialogue systems aimed at maintaining consistent character personas over multi-turn conversations using architectures like memory networks and transformers. Significant contributions included Li et al. (2016) [73] introducing persona embeddings, Kottur et al. [74] using personalized memory networks, and notably, Zhang et al. (2018) [75]'s PersonaChat dataset and memory-based models that set foundational benchmarks for persona consistency. These initial works primarily trained models from scratch, facing challenges in sustained persona adherence across interactions. The recent survey [76] extensively covers role-playing before and after the advent of LLMs. Within the scope of this survey, we focus specifically on role-playing interactions under multi-turn LLM settings.

In-Context Learning   Early LLMs demonstrated an ability to impersonate user-defined personas through prompting alone. Users could supply descriptions like "You are a wise old wizard..." in a system or context prompt, and models like GPT-3 would attempt to respond "in character." PersonaLLM [77] introduced a benchmark for personalization and highlighted that simply prefixing instructions with high-level persona descriptions yields limited diversity; instead, it proposed simulating nuanced user preferences via prompt-based reward models. It showed that prompting can move beyond trivial traits to tailor outputs to idiosyncratic user needs. Similarly, CharacterChat [78] used role-playing prompts with behavior presets and dynamic memory to maintain a character's persona over long conversations. It constructed an MBTI-based persona bank and prompted ChatGPT to produce dialogues between a "seeker" and a compatible "supporter," injecting preset behavioral tendencies and retrieving context-specific memory each turn to keep interactions coherent.

Beyond persona style, prompting has been used to improve reasoning. Role-play prompting [79] showed that instructing an LLM to "pretend to be" a domain expert can implicitly trigger step-by-step reasoning. In zero-shot settings across 12 tasks, prompting models with a role (e.g. "You are an excellent math teacher...") led to significantly higher accuracy than a vanilla prompt. The method uses a two-stage prompting framework: first having the model generate an "immersive" backstory or persona acknowledgement, then using that along with the query. These studies collectively demonstrate prompting-based role-play as a powerful lever: it can induce consistent persona adherence (persona and behavior presets) and even boost cognitive performance (reasoning via expert roles) without any parameter updates.

Supervised Fine-Tuning   To achieve more robust in-character behavior, researchers introduced instruction tuning and fine-tuning with role-playing data. PIPPA [80] released a partially synthetic corpus of over 1 million

role-play messages, crowdsourced from an online community of role-play enthusiasts. By fine-tuning on PIPPA's diverse persona-conditioned conversations, small LLMs dramatically improved at staying in character, underscoring that sheer volume and diversity of persona-rich dialogues can teach consistent role-playing. UltraChat [81] constructed 1.5M multi-turn dialogues via self-chat with GPT-4, covering broad topics and user types. Fine-tuning LLaMA on this yielded UltraLLaMA, which surpassed previous open models in general conversation quality (including user engagement and coherence). Other data efforts target specific role-play domains: PRODIGy [82] built a dialogue dataset from movie scripts aligned with detailed character profiles (biographies, personality traits). Fine-tuning models on these profile-grounded movie dialogues significantly improved consistency when emulating those characters – e.g. including a character's backstory and speaking style led to higher human preference for in-character responses.

A parallel direction creates models specialized for particular characters or customizable personas. ChatHaruhi [83] focused on anime characters, compiling 54k dialogues for 32 characters by combining original script lines with simulated conversations. The model fine-tuned on this data, augmented with a "memory" of past events for each character, was able to "revive" characters like Haruhi Suzumiya, accurately quoting lore and personality in new interactions.In another example, CharacterGLM [84] built on the Chinese GLM model to allow explicit profile injection for any character: it fine-tuned ChatGLM variants on a corpus of dialogues with richly annotated character profiles (covering identity, style, relationships) and achieved state-of-the-art human-likeness and consistency for customized personas. RoleCraft-GLM [85] further extended this concept by crafting original non-celebrity personas with emotional depth and fine-tuning a model on dialogues involving those characters. This yielded more nuanced emotional consistency, validating that meticulous character development during fine-tuning yields agents that are engaging and lifelike in their persona.

Recent work also explores how to fine-tune effectively for role consistency. Ditto [86] introduced a self-alignment pipeline where the model generates its own role-play dialogues for 4,000 distinct characters and then trains on them. By leveraging the model's internal knowledge to create training data (with feedback to ensure each character stays distinct), Ditto achieved strong persona fidelity across a wide range of roles, outperforming other fine-tuned baselines on a Role-play benchmark. This highlights a trend of using LLMs themselves to amplify training: both UltraChat and Ditto use models to generate synthetic data, but Ditto's focus is specifically on persona diversity and consistency (it treats role-play generation as a reading-comprehension task to avoid style collapse, then fine-tunes on the result). Another notable work, CharacterLLM [87], demonstrated a pipeline where for each target character (especially historical figures), one uses Wiki bios and documents to prompt an LLM to produce dialogues involving that character. Fine-tuning on this synthetic dialogue allowed a single agent to robustly play many famous roles, effectively transferring factual knowledge into conversational skill. In summary, supervised fine-tuning for role-play has evolved from harvesting large-scale data to increasingly clever data generation and specialization techniques.

Personalization and Rapid Adaptation   Even with instruction tuning, a given model has limits in the number of distinct characters or styles it can perfectly emulate. Thus, a key theme is personalization: adapting an LLM to a new persona with minimal data or effort. Recent research has explored parameter-efficient tuning modules that allow rapid persona swapping without retraining the entire model. PersonaPKT [88] proposed representing each persona as a continuous embedding vector that can be learned from a small set of that user's dialogues. By keeping the pre-trained model fixed and only training a tiny persona-specific vector (less than 0.1% of parameters), PersonaPKT efficiently imbues the model with that user's speaking style and preferences. In a related vein, PPlug [89] introduced a plug-and-play user encoder that on-the-fly computes a user embedding from their conversation history . Instead of a static learned vector, it employs a small model to read a user's past messages and output a "personal embedding" summarizing their quirks and facts. This embedding is then prepended to the LLM's input to personalize the response. Such design lets the LLM be dynamically personalized each turn based on context, and experiments showed significant gains in personalization metrics across tasks.

As LLMs are used to power multiple characters simultaneously, it becomes important to switch personas quickly and even maintain many personas at once. Neeko [90] addresses this with a dynamic LoRA (Low-Rank Adapter) framework for multi-character role-play. It pre-trains a separate LoRA module for each character and uses a gating network to activate the appropriate one based on the dialogue context. This incremental

ability means an unlimited number of personas can gradually accrue, each encapsulated in a plug-in adapter. An agent can seamlessly swap roles by toggling adapters, which showed superior consistency when one model needed to portray many characters in a group chat. Another challenge is preserving persona over multiple dialogue rounds. Standard fine-tuning often splits dialogues into independent turns, which can break character memory. MIDI-Tuning [91] proposed to explicitly model the user and system roles with separate adapters and a round-level state. In their framework, an LLM-based agent is trained by alternating between a "user adapter" (processing user utterances) and an "agent adapter" (generating responses), carrying hidden state forward through turns.

Reinforcement Learning   While supervised learning can teach a model a persona, it doesn't explicitly punish lapses. Reinforcement learning (RL) provides a way to directly optimize consistency and other long-horizon behaviors. Shea et al. (2023) [92] demonstrated this by applying offline RL to a dialogue model for persona consistency. They took an existing high-quality chatbot and defined a reward that penalizes persona breaks (e.g. contradicting provided profile or previous statements) and rewards in-character responses. Rather than interact with humans live, they performed RL on a static dataset of conversations – adjusting the model's policy to maximize the persona consistency reward while leveraging off-policy data. The result was a chatbot that, in human evaluation, more reliably adhered to its given persona description and avoided contradictions compared to its purely supervised counterpart. This work bridged the gap between static fine-tuning and RLHF: it shows one can inexpensively refine a model to be more in-character by offline RL on existing dialogues, getting some benefits of RL (direct control of behavior) without an expensive online loop.

Another aspect of multi-turn role-play that benefits from RL-like thinking is maintaining long-term coherence. COMEDY [93] approached this via a compressive memory mechanism that can be seen as the model "reinforcing" important memory content over a conversation. Instead of a traditional retrieval pipeline, COMEDY has the LLM periodically summarize and compress the dialogue history (including user persona hints and past events) into a concise memo, which is fed back into itself for future responses. This one-model architecture learns through supervised fine-tuning to generate useful summaries (e.g. remembering the user's preferences or the agent's own backstory) and to consult them when answering. While not an RL in algorithm, COMEDY's design implicitly optimizes a long-term reward: the compressed memory serves to avoid contradictions and boring repetition, much like an RL agent maximizing a reward for user engagement would learn to recall relevant facts.

Benchmarks & Evaluation   As role-playing agents become more advanced, evaluating their effectiveness requires moving beyond standard metrics like BLEU or response fluency. Instead, the field has introduced a suite of specialized benchmarks to assess whether an LLM can faithfully embody a persona, maintain consistency over time, interact appropriately in social settings, and remain aligned with ethical norms.

One line of work focuses on general personalization. The LaMP benchmark [94] evaluates whether LLMs can adapt to user-specific profiles across a range of tasks—such as rewriting text in a personalized tone or classifying based on individual preferences. While not limited to dialogue, LaMP highlights the broader need for systems that understand and leverage identity cues, and confirms that retrieval-based methods are especially effective for on-the-fly personalization.

A more targeted category of benchmarks examines character-specific fidelity. CharacterEval [95] is a Chinese benchmark featuring 77 characters from novels, each with multi-turn dialogues and detailed profiles. It defines 13 metrics across four dimensions – including character consistency, behavior realism, and conversational quality – and provides human and model-based evaluation for each. CharacterEval revealed that even GPT-4, when role-playing in Chinese, could be outperformed by fine-tuned local models on consistency, indicating that specialized training made a measurable difference. On the English side, RoleEval [96] poses factual and commonsense questions about 300 well-known characters, assessing whether models accurately retain and apply character-specific knowledge. Results show that global models like GPT-4 excel with internationally known figures, while locally fine-tuned models do better with culturally specific roles. Meanwhile, TimeChara [97] explores a new dimension—temporal consistency—by checking if a model role-playing a character at a given point in a story inadvertently leaks future events. Even the strongest models often violate timeline

boundaries, suggesting the need for narrative-aware mechanisms to constrain temporal knowledge. SimulBench [98] is a benchmark assessing LLM performance in interactive simulation scenarios – imaginative, role-playing and tool-use tasks that unfold over multiple turns. The benchmark includes tasks like acting as a Linux terminal, playing text-based games, and complex, long-horizon simulations requiring dynamic interaction with a user.

Moving from factual to psychological evaluation, InCharacter [99] assesses if a role-playing agent truly internalizes a character's personality. It uses an interview-style personality test: the agent (in role) is asked a battery of questions akin to a psychological survey, and its answers are compared to the expected personality profile of the character. Complementing this, RoleInteract [100] evaluates social interaction skills at two levels: one-on-one conversation quality (empathy, politeness, etc.) and group dynamics (how well an agent plays its role in a multi-agent conversation). RoleInteract includes 500 characters and diverse scenarios (e.g., an office meeting with several personas). One finding was that some agents that excel in bilateral chat struggled in group settings – sometimes a normally consistent character would conform or get sidetracked when other AI characters were present, indicating influence and social pressure effects.

A different angle on evaluation is to test how well models understand a character from source material. Yuan et al. (2024) [101] argue that a truly aligned role-play model should be able to read a narrative and produce a coherent character profile. They created the CROSS dataset of expert-written character profiles (covering attributes, relationships, events, and personality) for characters in novels. Models are evaluated on how well they can generate similar profiles after "reading" the novel (or being given chapters as input). In addition to intrinsic metrics (overlap with the expert profile), they evaluate extrinsically via a motivation recognition task: given a scenario from the story, does the model's profile help it answer why the character acted a certain way. Such evaluation drives home that role-play is not only about output style, but also about the model's internal model of the character.

Overall, the rapid advancements in LLM-based role-playing research demonstrate a clear evolution: moving from basic persona adherence to sophisticated, adaptable, multi-agent interactions, supported by advanced evaluation and continuous alignment efforts. This evolution highlights ongoing efforts toward creating engaging, realistic conversational agents capable of sustained, immersive role-play interactions. Such role-playing also plays a crucial role throughout all conversational engagement cases discussed in the remaining subsections.

2.2.2   Conversational Engagements in Healthcare

Healthcare is one of the key domains where multi-turn conversational large language models (LLMs) demonstrate significant potential. A defining characteristic of these models in the medical domain is their ability to emulate a doctor's role by engaging in task-oriented, context-aware dialogues with patients. Unlike single-turn medical knowledge question-answering systems, such as BenTsao [102] (formerly named HuaTuo), ChatMed, ShenNong-TCM, MING [103], and DoctorGLM [104], responding to isolated queries with full information at a time [105], multi-turn healthcare LLMs operate in a setting where patient information is often incomplete or ambiguous at the outset.

Ideally, multi-turn healthcare conversational LLMs should be capable of proactively generating a sequence of questions to refine their understanding of the patient's condition through iterative inquiry. This concept has been referred to by different names in various studies, including chain of questions (CoQ) [106], proactive questioning [106], symptom inquiry [107, 108], proactivity [109], and information seeking [110]. Despite the differences in terminology, they all describe the model's ability to **dynamically gather relevant details over multiple exchanges, ultimately leading to a more accurate diagnosis.** Additionally, multi-turn healthcare LLMs need to retain dialogue history, so that enabling seamless continuity in conversations, and integrate medical knowledge to ensure responses are accurate, reliable, and contextually appropriate.

While prior surveys [15, 16] provide a broad overview of medical dialogue systems, including both pre-LLM and LLM-based approaches, our focus is specifically on multi-turn LLMs and related evaluation framework. Traditional medical dialogue systems typically rely on rule-based logic or task-specific neural networks, which can be rigid and require extensive manual engineering. We explore the specific tasks these models address, the architectural advancements in their development, and the characteristics of the datasets used for pre-training and fine-tuning. In addition, we discuss the evaluation frameworks used to assess these models, highlighting

key metrics, established benchmarks, the growing focus on their information-seeking capabilities, and the challenges faced in online conversational healthcare systems.

Conversational Medical LLMs Development   The development of multi-turn healthcare LLM methods has seen significant progress. The following collected LLMs demonstrate the efficacy of customized training methodologies, curated data sets, and evaluation metrics, in improving healthcare domain performance, showing the potential of LLMs in advancing medical and psychological applications.

A collection of studies have aimed to make medical LLM communication more interactive and accessible. For example, Clinical Camel [111] is an open-source medical language model that introduces dialogue-based knowledge encoding (DBKE) to transform dense medical texts into conversational formats. This methodology enhances the model's ability to engage in multi-turn dialogues, aligning its responses to a conversational format. Fine-tuned from LLaMA-2 using QLoRA, Clinical Camel outperforms other LLMs, including GPT 3.5, GPT 4, and Med-PaLM2, across benchmarks such as USMLE Sample Exam [112], PubMedQA [113], and MedQA [114]. T-Agent [115] enhances medical dialogue generation by incorporating a term-aware approach by integrating a term extraction tool and a term prediction model within a two-stage training framework. This design aims to enhance the model's performance in dialogue term status extraction and generation tasks. Their experimental result demonstrates the improvements of T-Agent in the ROUGE and term extraction F1 scores. In contrast, APP (Ask Patients with Patience) [116] serves as a reasoning and guidance layer atop an LLM to support diagnostic decision-making during online medical consultations. Rather than altering dialogue format, APP emphasizes grounded reasoning, entropy minimization, and patient-centered communication without requiring further fine-tuning.

A majority of studies have focused on building medical LLMs that generate accurate and reliable responses in dialogue settings through SFT on healthcare-specific data. DISC-MedLLM [109] is fine-tuned on DISC-Med-SFT, a 400K-sample Chinese medical instruction dataset covering single-turn Q&A, multi-turn consultations, and multiple-choice Q&A, which is evaluated based on rule-based accuracy for single-turn and GPT-4 scoring for multi-turn conversations on Proactivity, Accuracy, Helpfulness, and Linguistic Quality. BianQue [106] integrates multi-turn doctor-patient Q&A datasets to enable a Chain of Questioning (CoQ) approach, emulating real consultations where doctors iteratively inquire to fully understand a patient's condition. Fine-tuned on ChatGLM-6B using the 2.4M-sample BianQueCorpus, BianQue improves CoQ by balancing questions and suggestions, addressing earlier models' limitations in interactive questioning. BiMediX [117], a bilingual medical mixture-of-experts LLM, enables seamless interaction in both English and Arabic. Fine-tuned using QLoRA, BiMediX outperformed existing models like Med42 and Meditron in English-based medical evaluations and significantly surpassed the generic bilingual LLM Jais-30B in Arabic medical and bilingual assessments. CPsyCounX [118] and PsycoLLM [119] are two healthcare conversational LLMs focusing on psychological counseling, where CPsyCounX is fine-tuned over InternLM2-7B-Chat with CPsyCounD, which contains 3,134 high-quality multi-turn consultation dialogues, and PsycoLLM is fine-tuned with psychological single-turn Q&A, multiturn dialogues, and knowledge-based Q&A on based on Qwen1.514B-Chat. Besides, the SMILE [120] method (Single-turn to Multi-turn Inclusive Language Expansion) uses ChatGPT to rewrite single-turn counseling QA exchanges into multi-turn dialogues. Starting from a mental health QA dataset (PsyQA [121]), SMILE injects diverse dialogue topics and prompts ChatGPT to produce realistic counselor–client conversations, then filters out any outputs that don't meet format or turn-count requirements. This process yielded SMILECHAT, a corpus of 55k multi-turn counseling dialogues in Chinese

Beyond solely relying on supervised fine-tuning (SFT), some studies have adopted a comprehensive pipeline encompassing pre-training, SFT, and reinforcement learning from human feedback (RLHF). Zhongjing [122], the first Chinese medical LLaMA-based large language model, exemplifies this approach through continuous pre-training, targeted SFT, and RLHF optimization. Particularly emphasizing multi-turn interactions, the authors constructed CMtMedQA, a specialized Chinese medical dialogue dataset comprising approximately 7,000 QA pairs derived from authentic doctor-patient exchanges across 14 clinical departments and spanning over 10 medical scenarios, including disease diagnosis, medication guidance, health consultation, and general medical knowledge inquiries. To address the inconsistency and brevity prevalent in real-world doctor responses, a self-instruct methodology [123] was utilized, standardizing replies into a uniform, professional, and empathetic communication style. Furthermore, the external medical knowledge graph CMeKG [124] was integrated

17

to verify and enhance the medical accuracy and safety of the dialogue interactions. HuaTuoGPT [107] is a medical consultation LLM, fine-tuned using both data distilled from ChatGPT and real-world data from medical professionals to combine fluent and informative responses with authentic diagnostic capabilities. The authors employ Reinforcement Learning from AI Feedback to align the model's outputs with the strengths of both data sources. HuatuoGPT is evaluated using rule-based NLP metrics for similarity check with reference answers, LLM-based assessment for language quality, symptom inquiry, treatment effectiveness, and patient helpfulness, and human evaluation by medical experts for diagnosis accuracy, treatment recommendation accuracy, and prescription knowledge, which demonstrates superior performance compared to baseline models under different criteria. The same research team later developed HuaTuoGPT II [125], which adopts a one-stage domain adaptation protocol that unifies heterogeneous data from traditional pre-training and supervised stages into a simple instruction-output pair format, facilitating efficient knowledge injection. The model achieved competitive performance with GPT-4 across multiple benchmarks, notably excelling in Chinese medical evaluations and the latest pharmacist license examinations.

Several papers also apply Direct Preference Optimization (DPO) within the full pipeline to generate responses that better align with human preferences. Aquila-Med [126] is a medical LLM that undergoes continued pre-training, SFT, and reinforcement learning with DPO to enhance its medical consultation capabilities. 12,727 DPO preference pairs are utilized for fine-tuning, which demonstrates significant improvements in handling single-turn and multi-turn medical consultations, indicating enhanced fluency, relevance, completeness, and proficiency. Similarly, Qilin-Med [127] utilizes a multi-stage training approach that combines Domain-specific Continued Pre-training (DCPT), SFT, and DPO. This method leverages the ChiMed dataset [128], encompassing question answering, plain texts, knowledge graphs, and dialogues, resulting in an accuracy improvement with the baseline Baichuan-7B model on the CMExam [129] test set during the SFT phase. In the DPO phase, Qilin-Med achieved scores of 16.66 in BLEU-1 and 27.44 in ROUGE-1 on the Huatuo-26M test set, indicating further enhancements over the SFT phase.

In addition to the mainstream approaches, Google's Articulate Medical Intelligence Explorer (AMIE) [130] represents a comprehensive advancement in conversational medical AI for training and evaluation. AMIE utilizes a chain-of-reasoning strategy within a simulated environment, incorporating self-play and automated feedback mechanisms to enhance its diagnostic dialogue capabilities across diverse medical conditions. Its training data includes both human-curated and AI-generated datasets, encompassing multiple-choice medical questions, long-form medical reasoning queries, clinical note summaries, and simulated dialogues based on various medical conditions. Evaluated through a randomized, double-blind crossover study using Objective Structured Clinical Examination scenarios, AMIE demonstrated superior diagnostic accuracy compared to primary care physicians and received higher ratings from both specialist physicians and patient actors on multiple assessment criteria.

Multi-turn Medical LLM Evaluation framework   Beyond the development of large language models (LLMs), we also found a growing body of literature focused on designing evaluation frameworks specifically for medical LLM systems, as traditional LLM evaluation methods often fall short in addressing the complexity and safety requirements of clinical applications.

Recent studies have proposed evaluation frameworks for medical LLMs that emphasize interactive quality, clinical reasoning, and human-centered communication. MedGPTEval [131] assesses LLMs like ChatGPT and Dr.PJ using 27 multi-turn dialogue cases and 7 case reports, measuring the accuracy, empathy, and clinical logic of the LLMs' responses. Similarly, Liao et al. (2023) [132] proposes an automated evaluation framework, emphasizing assessing LLM abilities, such as recognizing knowledge limitations, gathering relevant information, and improving diagnostic accuracy. The researchers reformulated medical multiple-choice questions from the USMLE into consultation tasks, creating a specialized benchmark for assessment. Additionally, they verify that fine-tuning with a consultation-specific dataset reduced hallucinations and improved benchmark performance.

Besides, a growing number of studies emphasize simulation-based interactive evaluation to approximate real-world clinical consultation. For instance, Liao et al. (2024) [108] introduced the Automated Interactive Evaluation (AIE) framework featuring the State-Aware Patient Simulator (SAPS), which incorporates a state tracker, memory bank, and response generator to support dynamic, multi-turn evaluation. Similarly, MMD-Eval

(Multi-turn Medical Dialogue Evaluation) [133] offers a locally deployable, task-oriented dialogue simulator, providing more resource-efficient and consistent evaluations than LLM-based scoring. MediQ [110] introduces an interactive benchmark for medical evaluation by simulating clinical interactions between a patient system and an adaptive expert system, emphasizing the assessment of information seeking ability. It employs abstention strategies to better estimate confidence and determine when to seek additional information. The benchmark converts existing datasets like MedQA and Craft-MD into interactive formats, simulating clinical interactions for evaluation purposes.

Complementing these efforts, MedFuzz [134] focuses on evaluating the adversarial robustness of healthcare LLMs by introducing ambiguous or unexpected inputs that could lead to clinical misconceptions and harmful decisions. It questions the assumption that high benchmark scores reflect real-world reliability by introducing complexities like ambiguous patient traits and biased data into medical QA benchmarks. Results show that GPT-3.5, GPT-4, and Med-PaLM 2 perform worse on the "MedFuzzed" benchmark, showing their vulnerability to clinical biases, demographic stereotypes, and incomplete data interpretation.

Recently, OpenAI introduced *HealthBench* [135], an open-source benchmark constructed from real-world healthcare conversations designed explicitly for evaluating LLM performance. HealthBench includes 5,000 multi-turn conversations averaging 2.6 turns per dialogue (ranging from 1 to 19 turns), capturing diverse and realistic patient-clinician interactions. It emphasizes clinical accuracy, appropriate communication depth, handling of uncertainty, and context awareness. Unique to HealthBench is its extensive use of physician-developed rubrics, consisting of over 48,000 distinct evaluation criteria, making it a robust tool for assessing nuanced aspects of clinical dialogue. Moreover, HealthBench introduces specialized subsets such as *HealthBench Consensus*, validated across multiple experts, and *HealthBench Hard*, comprising notably challenging interactions that test the limits of current LLM capabilities. This systematic, clinician-validated approach significantly enriches the landscape of multi-turn medical LLM evaluation by explicitly aligning automated assessment closely with clinical judgment.

In the development of healthcare conversational LLMs, a widely adopted strategy is to construct multi-turn medical dialogue datasets and fine-tune models using SFT, sometimes followed by RL-based methods. Interestingly, while some studies in general multi-turn dialogue suggest that SFT+RLHF may not yield significant improvements, this pipeline has proven surprisingly effective in the healthcare domain. This may be attributed to the relatively high quality and domain specificity of medical data, which contrasts with the noisier, less structured data used in broader multi-turn applications. Evaluation metrics in this area include traditional rule-based NLP measures (e.g., BLEU and ROUGE), diagnostic accuracy, and both AI-based and human assessments using more nuanced, subjective, domain-knowledge judgment criteria. However, very few papers conduct the human and AI agreement check [130, 108, 125], even if a considerable of literature use both of them. Notably, information-seeking and clinical reasoning are essential capabilities in this domain and are increasingly prioritized in newer benchmarks, which now place more emphasis on clinical relevance than conventional NLP metrics. Please refer to the appendix for details on the papers discussed.

2.2.3 Conversational Engagements in Education

Multi-turn conversational systems are increasingly central in education, enabling dynamic back-and-forth interactions that mimic human tutoring. Recent work can be grouped into three key areas: Intelligent Tutoring Systems, which use dialogic exchanges to teach or guide students; Automated Grading & Feedback, where AI provides iterative evaluation and comments on student work; and Scenario Simulation, which involves simulating students or classrooms with AI agents. Below, we survey advances in each category, highlighting how multi-turn conversation enhances educational effectiveness.

2.2.3.1 Intelligent Tutoring System

Socratic and Strategy-Guided Tutoring Early LLM-based tutors often fell into a simple question-answer pattern, providing direct answers and explanations that made students passive. To address this, researchers have developed Socratic and guided approaches that engage learners in multi-turn dialogue. SocraticLM [136] is a notable example, proposing a "thought-provoking" teaching paradigm in which the tutor asks open-ended

questions and prompts the student's reasoning instead of giving away solutions. SocraticLM was trained on SocraTeach, a new dataset of 35k multi-turn dialogues where a simulated teacher guides students with diverse cognitive states through math problems. Similarly, Kargupta et al. (2024) [68] tackle the "answer-too-direct" issue in the coding domain with TreeInstruct, an LLM-based tutor that plans a hierarchy of questions to help students debug code. TreeInstruct models the student's knowledge state and asks targeted questions for each error, effectively guiding learners to independently correct mistakes. It achieved state-of-the-art results on code debugging benchmarks and demonstrated in a user study that students could fix bugs with minimal direct hints.

Beyond specific algorithms, others have looked at controlling LLM tutors via high-level pedagogical strategies. For example, StratL [137] introduces a pedagogical steering framework: instead of letting an LLM tutor freely generate replies, StratL optimizes prompts to make the tutor follow a predefined teaching plan represented as a graph. In a case study on Productive Failure (an educational strategy where students first struggle with problems before instruction), StratL successfully steered an LLM tutor to withhold answers and encourage productive trial-and-error. Notably, in a field experiment with 17 high school students, the LLM tutor guided by StratL adhered to the desired strategy and helped students discover solutions on their own. These efforts collectively show a shift from straightforward Q&A to multi-turn, strategy-aware dialogue, making LLM tutors behave more like human teachers who ask, hint, and adapt rather than just tell.

Adaptive Tutoring Traditional intelligent tutoring systems often struggled with a "one size fits all" design that failed to accommodate individual learners. In contrast, many new approaches use conversational turns to tailor teaching to student needs. PACE (PersonAlized Conversational tutoring agEnt) [138], introduced by Liu et al. (2025) under the motto "One Size Doesn't Fit All," explicitly models student learning styles and personas to personalize the dialogue. PACE's multi-turn tutor adapts its explanations and questions according to the Felder–Silverman learning style model – for instance, giving concrete examples for sensory learners versus abstract prompts for intuitive learners. It also incorporates the Socratic method (asking guiding questions instead of lecturing) to stimulate critical thinking. To enable this, a new dataset of personalized tutoring dialogues was created, simulating students with diverse backgrounds and personalities and pairing them with an LLM tutor.

Moreover, LLM tutors can track a student's step-by-step reasoning in multi-turn exchanges and provide targeted help at the right moment. This ability to model the student within the conversation marks a key advantage of multi-turn tutoring. An illustrative real-world deployment is JeepyTA [139], a GPT-based virtual teaching assistant used in an online course forum. JeepyTA monitors student questions on course material and responds in seconds with context-specific help, adapting its style to the informal, conversational tone of forum discussions. It distinguishes logistical queries from conceptual ones and provides hints or explanations accordingly. By continuously engaging with students' follow-up questions, such an always-on conversational TA exemplifies how personalization and instant adaptivity can scale via LLMs.

Evaluating LLM Tutoring in Mathematics Recent research further explores ways to deepen assessments of LLM tutoring capabilities, particularly regarding their subject expertise and pedagogical effectiveness in mathematics. An influential early example is the work of Macina et al. (2023) [53], who introduced *MathDial* to systematically evaluate LLMs' abilities as tutors, emphasizing faithful and equitable teaching. Their approach involved collecting high-quality, teacher–student dialogues through human–LLM interactions, with InstructGPT simulating student behaviors—including common misconceptions—and expert annotators assuming the teacher role. Diverse student responses were elicited using temperature sampling, prompting the LLM to generate realistic errors. Human teachers then employed scaffolding strategies to guide the simulated students toward solutions. Dialogue quality was ensured through rigorous human annotation. Evaluation metrics included the simulated student's success rate and the telling@k score, indicating how frequently teachers prematurely revealed answers. Models fine-tuned on the MathDial dataset demonstrated improved student outcomes by emphasizing hints and prompting over direct answers. Similarly, Ding et al. (2024) [140] introduced *SocraticLLM*, a knowledge-enhanced model emulating Socratic questioning to foster critical thinking and self-discovery. To support this method, they developed the publicly available *SocraticMATH* dataset, containing structured Socratic dialogues that cover 513 primary-school math topics.

Another benchmark, *MathTutorBench* by Macina et al. (2025) [141], specifically targets one-on-one math tutor–student dialogues, assessing LLM tutors across multiple dimensions, including Math Expertise (accuracy in problem-solving), Student Understanding (ability to diagnose and correct misconceptions), and Teacher Response Quality (effectiveness in providing hints and Socratic guidance).

Most recently, retrieval-augmented generation (RAG) techniques have further advanced the field. For instance, Feng et al. (2024) [142] developed *CourseAssist*, a system that grounds tutor-generated responses explicitly in course syllabi and lecture notes, ensuring alignment with instructor expectations. Levonian et al. (2025) [143] explored alignment methods to enhance generative tutors' outputs, making responses safer, pedagogically appropriate, and closely relevant to student queries, significantly improving multi-turn algebra tutoring dialogues. Expanding upon this theme, Scarlatos et al. (2025) [144] argue for the necessity of adaptive AI tutors that subtly adjust responses across dialogues, strategically guiding students toward correct understanding without explicit intervention. They operationalized this idea by generating multiple potential tutor replies at each conversational turn and then assessing them with a student model—which predicts students' subsequent correctness—and a pedagogical evaluator. By ranking tutor responses based on these dual criteria, they successfully fine-tuned a new tutor model through direct preference optimization, thereby rewarding interactions that consistently promoted student learning.

Bridging Research and Real Classrooms The emerging trend in conversational tutors involves integrating research insights directly into classroom practice, facilitated by the increasing availability of large language models tailored specifically for education. Major industry players, including Google with LearnLM [145] and Anthropic with Claude for Education [146], have actively embraced this approach, developing specialized educational versions of their flagship AI models. These initiatives reflect a convergence between academic research and industry applications, emphasizing personalized, adaptive tutoring methods grounded in evidence-based dialogue strategies and instructional principles.

2.2.3.2 Automated Feedback & Grading Support

Automated Feedback Support Large language models are increasingly leveraged to generate automated feedback and grading, aiming to support instructors with large-scale assessments [147]. Similarly, in open-ended writing tasks, LLMs can produce fluent and plausible comments that appear insightful, yet often include content not grounded in the student's work [148]. This lack of faithfulness (e.g. fabricated critiques or irrelevant suggestions) is a critical limitation, as unfaithful feedback can mislead or confuse learners. Improving the accuracy and alignment of feedback with students' actual mistakes has therefore become a central research focus.

To address these issues, recent work has introduced strategies to make LLM feedback more interactive, adaptive, and pedagogically grounded. One approach is to incorporate a verification step before feedback delivery: Daheim et al. (2024) have an LLM "verifier" analyze the student's reasoning step-by-step to pinpoint errors, which then guides a tutor model to give targeted hints [149]. This stepwise verification significantly reduces hallucinated advice and yields feedback more precisely tailored to the student's misunderstanding. Another line of research optimizes feedback through learning from human pedagogical preferences. Scarlatos et al. (2024) [150] propose a rubric-based evaluation of feedback quality (checking for correctness, encouragement, misconception-addressing, etc.), use GPT-4 to label LLM outputs along these criteria, and then apply reinforcement learning to tune the model . The result is an LLM that produces feedback with measurably higher factual correctness and better alignment to effective tutoring practices. There are even efforts to "close the loop" by having the LLM anticipate student revisions: Nair et al. (2024) [151] introduce a system where the model generates feedback, simulates how a student would revise their essay in response, and iteratively refines its feedback to maximize the improvement between drafts . This led to greater actual improvements in student writing compared to static feedback, and the optimized feedback exhibited enhanced pedagogical qualities.

Studies are also examining how well LLM feedback aligns with real classroom needs and how students/instructors perceive it. Initial deployments of LLM-based feedback systems in university courses show mixed but encouraging results. In a graduate-level computer science class, an automated feedback tool generated

paragraph-level comments on project reports, which students found generally helpful for improving their work [152]. The course instructor noted, however, that the AI's comments did not always match the assignment's pedagogical objectives or emphasis [153]. In fact, the instructor preferred to use the LLM's output as a draft – a starting point that the instructor would edit – rather than sending it raw to students. This highlights an important limitation: current LLMs may need human oversight to ensure feedback is consistent with the teacher's goals and curriculum. On the other hand, domain-focused applications of LLM feedback have shown clear benefits when carefully integrated. Riazi and Rooshenas developed an LLM-driven tutor for a databases course that could analyze a student's entity-relationship diagram and generate detailed, context-specific critiques and follow-up questions [154]. Likewise, in medical education, LLMs have been used to generate explanatory feedback for answers to medical board-style multiple-choice questions. Experts evaluating such feedback found it relevant and useful – not a replacement for human feedback, but a valuable supplement to the usual numeric scores students receive [152].

Efforts like Lohr et al. (2025) [155]'s, which prompt LLMs to produce specific types of feedback from established educational taxonomies (e.g. an error-identification vs. a hint vs. an elaboration), further illustrate the push toward more controlled and purposefully designed feedback messages. In addition to feedback itself, researchers are beginning to evaluate LLMs on related teaching skills such as asking good questions – for instance, the Dr.Academy benchmark assesses whether LLMs can generate high-quality, higher-order questions in line with Bloom's taxonomy, finding that models like GPT-4 already show strong capability in formulating deep conceptual questions. Together, these advances show a clear developmental trajectory: from basic feedback generation that often wandered off-target, to increasingly faithful, adaptive, and pedagogically-aware feedback loops facilitated by LLMs' multi-turn interaction capacity.

Automated Grading Support   In parallel with feedback generation, researchers have started leveraging LLMs to grade student work and provide evaluative judgments (scores, ratings, or rubric-based assessments). Automatic grading by AI is not entirely new, but LLMs offer a unified, flexible approach that can handle open-ended responses more like a human grader. Recent studies suggest that, for certain types of assignments, LLM graders can approach human-level performance in both consistency and accuracy. Capdehourat et al. (2025) [156] explored LLMs for scoring short free-response questions in Spanish, a scenario involving language complexity beyond the typical English-centric datasets. They found that state-of-the-art models (including GPT-4 and advanced open-source LLMs) could predict expert graders' scores with over 95% accuracy in a three-tier grading scale, and even 98% accuracy on simpler right/wrong judgments.

Another study compared ChatGPT directly against university instructors for grading full exam papers in higher education. Out of 463 Master's-level exam responses graded, about 70% of the AI's assigned scores fell within a 10% margin of the human-given score, and 31% were within a 5% margin [157]. Teachers involved in the experiment expressed surprise at how closely ChatGPT's evaluations matched their own in these exams. However, important discrepancies were observed. The AI grader tended to be more conservative, avoiding very high or very low scores on individual questions.

In the domain of essay scoring, which demands understanding of content, organization, style, and often providing feedback, LLMs still struggle to meet human-level nuance. Kostić et al. (2024) [158] conducted a case study using GPT-4 to evaluate German-language student essays at a business school . The LLM could generate a score and some comments, but it often failed to apply the rubric criteria consistently and lacked the depth of feedback that human lecturers provided. Complex aspects of writing quality (critical analysis, creativity, etc.) proved difficult for the model to judge correctly, highlighting a gap between what current LLMs can do and the "nuanced requirements" of real essay evaluation.

2.2.3.3   Scenario Simulation

Recent advances have explored scenario simulation as a means to leverage LLMs for enacting multi-turn educational interactions between virtual teachers and students. Early work focused on using LLM-simulated student profiles to evaluate learning materials. For example, Generative Students introduced a prompt-based architecture (grounded in the Knowledge-Learning-Instruction framework) to instantiate diverse student profiles

defined by mastered vs. confused knowledge components [159]. Each simulated student (powered by GPT-4) answered multiple-choice questions, producing responses that aligned with its knowledge profile. Notably, these generative students exhibited answer patterns highly correlated with real student performance, correctly flagging many of the same difficult questions. This result suggested that realistic virtual learners can serve as proxies for human students in content evaluation, allowing instructors to identify problematic questions before deployment.

Subsequent research broadened the fidelity of simulated students by incorporating individual differences in ability and personality. Liu et al. (2024) [160] developed a personality-aware simulation framework that enriches student profiles with both cognitive level (e.g. language proficiency) and noncognitive traits (e.g. conscientiousness). In a language tutoring scenario, an LLM could then produce diverse student utterances consistent with a given persona, which in turn successfully triggered the tutor's adaptive scaffolding strategies. Building on this idea, Jin et al. (2024) [161] introduced TeachTune, a system enabling teachers to test their pedagogical conversational agents (PCAs) against diverse simulated students. Teachers specify a student's presumed prior knowledge and motivation, and an LLM-driven student agent engages in a multi-turn chat with the PCA. This automated student–teacher dialogue reveals how well the PCA adapts its explanations and feedback to different learner needs, going beyond single-turn Q&A tests. The TeachTune pipeline ensured that each simulated student's behavior remained faithful to its profile, with measured deviations under 5–10%. These efforts highlight that richly modeled virtual students can support teacher agents and tutoring systems by surfacing potential shortcomings in adaptive instruction in a cost-effective manner.

Researchers have also scaled scenario simulation to multi-party settings. Zhang et al. (2024) [162] proposed SimClass, a framework in which multiple LLM-based agents assume typical classroom roles. A novel class-level control mechanism orchestrates the agents' turn-taking and topic flow to emulate a live classroom lesson. In user trials with real students, SimClass was able to simulate dynamic classroom interactions featuring both teacher–student exchanges and student–student discussions. The emergent group behavior was strikingly human-like – the student agents would ask and answer each other's questions and collaboratively debate topics, creating an enlivened atmosphere. These collective simulations improved the human participant's learning experience by maintaining engagement and peer-like dialogue support. The success of SimClass demonstrates that LLMs can collectively model complex social dynamics of a classroom, opening the door to virtual class rehearsals and large-scale peer interaction scenarios.

Later on, scenario simulation has been extended to address specialized pedagogical needs for both learners and instructors. To better support students with low academic performance or poor metacognitive skills, Li et al. (2025) [163] devised a pipeline for generating struggling student agents and evaluating their realism. This approach automatically creates a spectrum of student profiles with varying learning deficiencies and filters them through a two-round LLM evaluation (validated by human experts) to ensure the simulated learning struggles are authentic. By assembling a set of high-fidelity "at-risk" student agents, educators and intelligent tutoring systems can safely experiment with interventions to foster self-regulation and reflective thinking, without ethical concerns of testing on real students. On the instructor side, Hu et al. (2025) [164] explored using LLM-based teaching simulation to improve lesson planning. In their method, an LLM is prompted to play out a full classroom lesson based on a teacher's draft plan – simulating the teacher's instruction and the students' reactions – and then to generate a reflective critique of that session. The insights from this simulated teacher–student interaction are used to iteratively refine the lesson plan. Additionally, Wang et al. (2024) [165] proposed Book2Dial, a framework to automatically generate synthetic teacher-student dialogues from textbook content to address data scarcity in developing educational chatbots. Three dialogue-generation approaches—Multi-turn QG-QA, Dialogue Inpainting, and LLM-based Role-Playing—are introduced and evaluated using automated metrics (e.g., coherence, answerability, factual consistency) and human judgment (specificity). The results demonstrate that LLM-based Role-Playing performs best, highlighting a cost-effective method for chatbot training in educational domains.

In summary, multi-turn conversational simulations serve as a valuable sandbox for educational innovation. By leveraging LLMs to generate realistic student and teacher behaviors, researchers and practitioners can prototype and test interventions rapidly, ethically, and inexpensively. These simulations complement live studies: they can uncover issues and inform design decisions before real students are involved, and suggest which approaches

merit real-world trials. From generative students for item analysis to full classrooms for teacher training, the common thread is that rich, multi-turn interactions are the fabric of these simulations.

2.2.4   Conversational Engagements in Jailbreak

In multi-turn settings, LLMs face not only heightened requirements for consistency but also an increased risk of malicious exploitation. Although LLMs excel at various tasks, their vulnerabilities can lead to harmful outputs—such as generating dangerous instructions—that highlight significant limitations compared to human judgment. The phenomenon of multi-turn jailbreaking, where adversaries bypass guardrails over a series of exchanges, has thus emerged as a critical area of concern.

Most prior research has focused on single-turn jailbreaks, in which adversaries use a single prompt—often few-shot—to trigger harmful content. For example, optimization-based methods in [166, 167] utilize gradient-derived gibberish suffixes to elicit such outputs, but these techniques depend on internal token probability knowledge and are not applicable to closed-source models. To explore alternative strategies, subsequent studies [168, 169] have investigated unconventional communication patterns, such as role-playing scenarios, and developed multi-turn conversational methods that leverage the entire dialogue history. These approaches demonstrate that multi-turn jailbreaking, where even seemingly innocuous prompts contribute to later interventions, presents a far more complex challenge.

Crescendo and ActorAttack   Multi-turn jailbreak can employ more diverse strategies. "Crescendo" [170, 171] is one of the strategies, which implicitly instruct victim LLMs to provide harmful information. The intuition behind this strategy is that LLMs agreeing to a small, initial request increases the likelihood of complying with subsequent, larger demands. By asking innocuous but implicitly suggestive questions, the token distribution of LLMs shifts towards the direction where tokens containing more harmful information are more likely to be generated. The multi-turn interactions as a whole jailbreaks LLMs instead of one specific questions plays the role. Besides, the implicitness of questions and the order to present the questions to LLMs are important to the success of jailbreak. Following the crescendo idea, Ren et al. (2024) [171] generalizes the crescendo method used in [170], where in [170] fixed and human-crafted seed instances are needed to generate attacks, making it challenging to generate diverse attacks. In [171], the novelty focuses on the process of self-discovering diverse attack clues inside the model's prior knowledge via network structures and the classification of the clues, by constructing the two-layer relation tree according to Latour's actor-network theory.

Decomposition   Another important strategy for multi-turn jailbreaks is to decompose the one harmful prompt into several pieces where each contains less malicious contents [172, 173, 174]. Therefore, language models can incrementally generate harmful content through multi-turn dialogue. By decomposing the original malicious query into several less harmful sub-questions can evade the guardrail of LLMs and induce harmful responses. Due to the in-context learning capabilities, harmful knowledge can be gathered together in the final turn. Using such decomposition, alignment is quite successful for each turn in a multi-turn dialogue, except the final turn when all responses are gathered and combined. However, the cumulative harmful content across the dialogue results in an overall alignment failure. Following the idea, Wang et al. (2025) [175] trains a red-team jailbreak agent through the interactions with target LLMs to generate decomposed yet coherent jailbreak prompts. [176] proposes a scenario-based jailbreak method to disguise the malicious intent from guardrails of LLMs, assuming LLMs can only detect direct harmful intent but would be misled if the attacker creates a scenario claiming that others are planning harmful actions and positioning the attacker as the protector. The attack is decomposed into multi-turns, where the attacker firstly describes others' harmful intent and seeks prevention, secondly asks about possible evidence items, finally requests an example harmful plan for comparison.

Datasets for Multi-Turn Jailbreaks   AdvBench [166] and HarmBench [177] are two important benchmarks for jailbreaking LLMs. AdvBench contains a set of 500 harmful behaviors formulated as instructions, as well as a collection of 500 strings that reflect harmful or toxic behavior. HarmBench contains 510 unique harmful behaviors, split into 400 textual behaviors and 110 multimodal behaviors. Both datasets are not intentionally

curated for multi-turn jailbreaks, but offer representative jailbreak tasks. Based on AdvBench and HarmBench, Russinovich et al. (2024) [170] manually crafted crescendo multi-turn prompts, and Ren et al. (2024) [171] scales the "attack chain" generation process by employing LLMs via self-talk and six clues from Latour's actor-network theory. Ren et al. (2024) [171] then compiles the data into SafeMTData. Gibbs et al. (2024) [172] employs word substitution cipher approach [178] to process data from HarmBench with Mixtral-8x7b to isolate the impact of the multi-turn prompting structure, and [173] uses decomposed AdvBench for jailbreaking. In constrast, [174] uses HarmfulQ dataset [179], which comprises 200 explicit harmful questions in English, to process for multi-turn jailbreak. In [175], AdvBench is used for training the jailbreak model and JBB [180] is used for evaluation. In [176], the authors leverage the Beavertails dataset [181], which contains malicious queries across 14 categories designed to test a model's refusal capability, and employ sentence transformers [182] to generate multi-turn dialogues and harmful actions for GPT-4o.

Defenses against Multi-Turn Jailbreaks While researchers have developed numerous multi-turn methodologies to jailbreak LLMs, effective defense strategies specifically designed against such multi-turn attacks remain scarce. Conventional defense methods (e.g., perplexity filters, input/output filters, rephrase/retokenize, rand-drop) demonstrate limited efficacy against these complex, extended conversational interactions. To address this challenge, Yu et al. (2024) [183] shows that system prompts and Chain-of-Thought (CoT) [48] can partially counteract attacks by refusing to answer harmful queries—albeit with lower helpfulness scores—while Gibbs et al. (2024) [172] finds that NeMoGuardrails [184] can be overzealous even with benign prompts, and Liu et al. (2024) [174] notes that stronger defenses often compromise usability. To summarize, robust guardrails against multi-turn jailbreaks remain scarce, and even the limited defenses available often compromise usability by hindering benign interactions. This stark trade-off underscores the urgent need for adaptive, context-aware strategies that can maintain both security and user-friendly performance.

3 Improvements

Recent advances in enhancing multi-turn interactions with Large Language Models (LLMs) have pursued diverse approaches that address the unique challenges of extended conversations. As illustrated in our taxonomy (Fig. 1), current improvement methods can be categorized into three main branches: (1) Model-Centric Approaches, which focus directly on adapting and refining LLMs to better handle sequential dialogue dynamics through strategies such as in-context learning, supervised fine-tuning, reinforcement learning, and novel architectures (§3.1); (2) External Integration Approaches, which augment LLM capabilities by incorporating external resources, including memory structures, retrieval mechanisms, and knowledge graphs, to mitigate context limitations and ensure factual consistency (§3.2); and (3) Agent-Based Approaches, which represent a paradigm shift toward treating LLMs as proactive, iterative agents that interact either individually or collaboratively to manage complexity and enhance reasoning over sustained interactions (§3.3). Collectively, these methods represent an evolving toolkit that significantly expands the potential for sophisticated, context-aware, and reliable multi-turn interactions with LLMs.
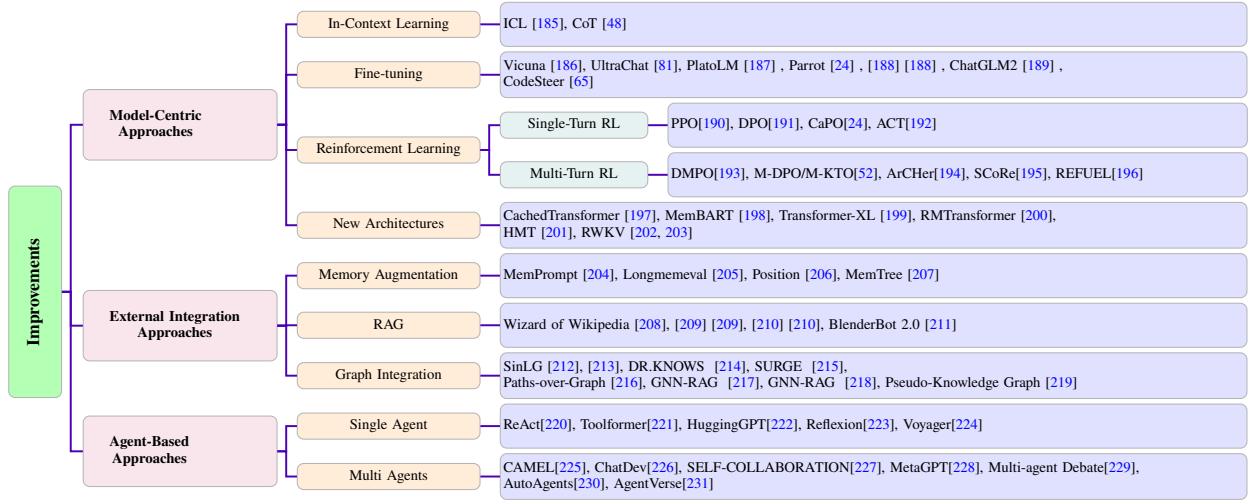

Figure 1: Taxonomy of Improvements Methodologies in Multi-turn LLM Interactions.

3.1 Model-Centric Approaches

This section surveys key model-centric strategies aimed at improving the performance of LLM on multi-turn interaction tasks. As dialogue tasks shift from static, single-turn queries to dynamic, multi-turn exchanges, traditional modeling techniques often fall short. We examine four major approaches that address these challenges from different angles: (1) In-Context Learning, which explores prompt-based adaptation using multi-turn exemplars (§3.1.1); (2) Supervised Fine-Tuning, which focuses on data curation and training strategies for maintaining coherence and context over multiple rounds (§3.1.2); (3) Reinforcement Learning, which aligns multi-turn behaviors with human preferences through trajectory-level optimization (§3.1.3); and (4) New Architectures, which reimagine the Transformer design to better support long-range memory and dialog flow (§3.1.4). Together, these techniques represent an evolving toolkit for building LLMs that can reason, remember, and respond effectively in sustained interactions.

3.1.1 In-Context Learning

In-context learning (ICL) [185] in multi-turn settings yields nuanced outcomes according to recent empirical studies. While providing exemplars usually aids single-turn tasks, naive in-context prompting with interactive multi-turn examples can sometimes hurt performance in sequential dialog scenarios [29]. For instance, on

the AQA-Bench sequential reasoning benchmark, certain models (e.g. LLaMA2 and DeepSeek-LLM in a coin-guessing game environment) actually saw their accuracy drop when given a single in-context demonstration, recovering only as more examples were provided. This surprising inversion of the typical few-shot benefit is attributed to the multi-round nature of the examples, which may cause overfitting to specific interaction trajectories. In other words, multi-turn example formats behave differently from standard one-turn Q&A prompts. Consistent with this, dedicated dialogue benchmarks like MT-Bench [21] and its fine-grained successor MT-Bench-101 [30] highlight that multi-turn interactions demand evaluating facets such as context retention, proactivity, and adaptability that one-shot QA tests often overlook.

Across diverse domains, a variety of prompting techniques have been explored to improve multi-turn LLM interactions, with mixed empirical results. In code generation, interactive multi-turn prompting has proven beneficial: the InterCode [22] benchmark showed that strategies like step-by-step reasoning (e.g. Chain of Thought [48], ReAct) or iterative plan-and-refine prompting yield concrete performance gains on coding tasks by leveraging compiler feedback over turns. In the clinical domain, role-playing a domain expert during training enabled notable gains – Clinical Camel [111], which uses dialogue-based knowledge encoding to infuse medical texts into a Q&A format, surpasses GPT-3.5 on multiple medical benchmarks in few-shot evaluations. By contrast, persona-driven prompts do not universally boost objective accuracy: a systematic study found that adding a persona (e.g. instructing "You are a helpful assistant. . . ") in system prompts generally failed to improve factual question-answering performance [232]. Nevertheless, persona-centric role-play can enhance conversational quality and engagement in the right context. For example, CharacterChat [78] employs MBTI-aligned personas with preset behaviors and dynamic memory; its role-play prompts facilitated effective personalized support dialogues, demonstrating remarkable efficacy in providing tailored social support. Finally, explicitly steering the dialogue structure has shown promise in certain multi-turn tasks. The StratL [137] approach, which optimizes prompts to guide an LLM tutor along a predefined teaching plan, successfully induced a more pedagogically effective multi-round tutoring strategy (Productive Failure) in practice.

These findings collectively suggest that the impact of ICL and prompt design on multi-turn performance is highly context-dependent – improvements emerge when the prompting strategy is well-aligned with the task's interactive dynamics, whereas misaligned or naive prompts may even degrade performance in complex multi-turn environments.

3.1.2   Supervised Fine-Tuning

Classic supervised fine-tuning (SFT) methods like InstructGPT [190] and FLAN [233], along with parameter-efficient techniques such as AdapterDrop [234] and LoRA [235], have driven major progress in single-turn LLM tasks. For a survey on SFT, see [236]. However, to enhance LLMs' multi-turn interaction ability, SFT must be modified; as Wang et al. (2023) [23] point out, treating each round independently can lead to context forgetting and incoherent responses can lead to context forgetting and incoherent responses in multi-turn dialogues. This limitation has motivated the development of tailored strategies that both leverage realistic multi-turn data and adjust training methods to fully exploit such data.

Moreover, SFT remains one of the most commonly used techniques to boost LLMs' multi-turn performance. Several studies, as discussed in §2, suggest that incorporating domain-specific multi-turn interaction datasets into SFT can further improve LLM performance. For example, research in multi-turn instruction following (including math reasoning and multi-lingual context) [30, 32, 31], role-play [81, 82] and clinical dialogues [111, 107, 109, 122, 132, 127, 126, 117, 165, 53, 144] demonstrates promising gains. These findings underscore the need for further research to improve fine-tuning efficacy specifically in multi-turn settings.

Realistic Multi-Turn Dialogue Data Curation   Recent works have addressed multi-turn challenges by generating datasets that capture natural dialogue flows. For example, Vicuna [186] fine-tunes on user-shared ChatGPT conversations to retain genuine multi-turn interactions, though scaling such data remains costly. Other methods, such as those in UltraChat [81], use self-chat to create extensive multi-turn data; however, these auto-generated dialogues can be overly scripted and lack diversity. To mitigate these issues, approaches like PlatoLM [187] incorporate sophisticated user simulators (e.g., the "Socratic" agent) to produce more dynamic and realistic multi-round dialogues, thereby enhancing topic shifts and follow-up naturalness. In

addition, Parrot [24] generates multi-turn data by first mimicking human asking styles for query generation, then constructing negative responses that simulate context neglect or misunderstanding, and finally combining this with context-aware preference signals for fine-tuning.

Optimized SFT Approaches   In parallel with dataset curation, new fine-tuning strategies have been proposed to fully leverage multi-turn data, including aspects of modified loss functions, longer context length, and more efficient training. For example, Vicuna [186] employs a modified loss function and extended context lengths, supported by gradient checkpointing [237] and FlashAttention [238], to handle long dialogue histories efficiently. Similarly, ChatGLM2 [189] extends context length with comparable techniques while also reducing GPU memory costs through multi-query attention [239] and causal masking strategies. More generally, Teng et al. (2024) [188] propose a fine-tuning method that optimizes cross-entropy loss combined with KL divergence across all dialogue turns. Instead of focusing solely on the final turn, this approach leverages the entire conversation, resulting in more coherent outputs and efficient training. Additionally, along with CodeSteer, Chen et al. (2025) [65] address the issue of multi-round gradient cancellation, where gradients from early rounds may cancel out those from later, more informative rounds, by doubling the weights of the final two rounds during fine-tuning. This strategy ensures that the most influential guidance steps drive the model update, improving its ability to select the optimal initial step.

In summary, these advances illustrate two complementary avenues for enhancing multi-turn interactions in LLMs with SFT: generating realistic dialogue data and optimizing fine-tuning strategies. Together, these approaches promote better context retention and coherent multi-turn responses, marking a significant step toward more robust conversational models.

3.1.3   Reinforcement Learning

Reinforcement learning (RL) has become central to aligning LLMs with human preferences, improving their safety, helpfulness, and coherence. While initial RL successes focused on optimizing single-turn interactions using human or AI feedback, real-world conversations often span multiple turns, introducing complexities beyond single-turn methods. Addressing these challenges has motivated recent developments in specialized multi-turn RL algorithms designed explicitly for extended conversational interactions.

3.1.3.1   Single-Turn RL

Reinforcement learning from human feedback (RLHF) aligns LLMs with human preferences by using reward models trained on human-generated rankings, significantly improving response quality and safety. For instance, a 1.3B-parameter InstructGPT fine-tuned via RLHF outperformed a 175B-parameter GPT-3 in human evaluations, exhibiting greater truthfulness and reduced toxicity [190]. Reinforcement learning from AI feedback (RLAIF) extends RLHF by employing AI-generated feedback signals instead of human labels. Anthropic's Constitutional AI [240] notably trains models using AI-authored principles and self-critique, successfully reducing harmful behaviors without direct human supervision. Both RLHF and RLAIF have proven effective for aligning single-turn LLM responses with human values.

Early RLHF successes largely utilized Proximal Policy Optimization (PPO), an on-policy method offering stable and sample-efficient fine-tuning of large models. For example, OpenAI's InstructGPT [190] employed PPO to optimize human-preference-based reward models, significantly enhancing helpfulness and reducing harmful outputs with minimal performance loss on traditional NLP benchmarks.

While PPO-based RLHF is effective, it can be complex due to reward modeling and tuning. Direct Preference Optimization (DPO) [191] simplifies this by converting RLHF into a supervised classification task, analytically deriving optimal policies from preference models. Context-aware Preference Optimization (CaPO, [24]) extends DPO by integrating context-aware preferences, promoting correct context utilization. Additionally, Action-Based Contrastive Self-Training (ACT) [192] applies DPO quasi-online, enhancing LLMs' clarifying behaviors in ambiguous contexts while preserving DPO's simplicity.

Many previous studies have investigated RLHF methods for enhancing multi-turn interactions in LLMs [24, 30, 23, 28, 46, 20, 22, 65, 109, 122, 127, 125, 150]. By horizontally comparing these works, however, a nuanced and somewhat contradictory picture emerges regarding RLHF's effectiveness. On one hand, specialized domains such as medical QA [122, 127] and coding tasks [22, 65] demonstrate substantial performance gains from carefully implemented RLHF approaches like DPO and interactive feedback loops. Similarly, the Parrot framework [24] underscores how explicitly training on negative examples can guide models towards better contextual alignment and fewer repetitive errors.

On the other hand, broader analyses from MINT [23] and MT-Bench 101 [30] reveal that generalized RLHF methods frequently yield limited or even negative impacts in multi-turn scenarios. This discrepancy highlights the inherent limitations of RLHF methods in terms of generalizability, emphasizing the necessity for tailored RL strategies explicitly designed for multi-turn interactions.

3.1.3.2   Multi-Turn RL

Recent research has therefore extended preference optimization and reinforcement learning techniques to optimize entire conversation trajectories rather than single responses. Below, we review key developments organized by technique.

Multi-Turn DPO and Variants   A straightforward way to align multi-turn behavior is to generalize the single-turn preference optimization objective across an entire dialogue trajectory. Direct Multi-Turn Preference Optimization (DMPO) [193] is one such approach that adapts DPO to multi-turn agent tasks. A technical obstacle in multi-turn DPO is that the partition function (normalization factor) no longer cancels out between preferred and dispreferred trajectory pairs, complicating the loss computation. DMPO addresses this by re-formulating the optimization: it replaces the per-response probability ratio with a state-action occupancy measure and normalizes for sequence length differences. The result is a novel DMPO loss that comes with theoretical justification for multi-turn settings.

Related work by Xiong et al., (2024) [52], which we covered in §2.1.2, introduced an iterative multi-turn preference learning framework with two implementations: Multi-turn DPO (M-DPO) and Multi-turn KTO (M-KTO). (Kahneman-Tversky Optimization (KTO) is a recently proposed RLHF approach inspired by prospect theory, designed to better align model outputs with human preferences by explicitly modeling decision-making biases [241].) Their focus was on mathematical reasoning agents that utilize tools (code interpreters) across multiple turns. Because single-turn preference methods did not fully capture multi-step reasoning quality, they extended DPO and the prospect-theoretic KTO loss to handle entire solution trajectories. In this framework, a reward model (augmented with tool feedback) judges the quality of a full reasoning trace, and the model is tuned to prefer better traces. Both M-DPO and M-KTO yielded substantial performance gains on math problem benchmarks: for instance, a 7B model's accuracy on GSM8K math questions jumped from 77.5% to 83.9% after multi-turn preference optimization.

Hierarchical RL and Credit Assignment   Multi-turn interactions intensify the credit assignment problem—attributing outcomes to specific decisions made across several dialogue turns. Hierarchical reinforcement learning addresses this challenge by structuring decision-making across multiple abstraction levels. ArCHer [194] employs a two-level architecture: a high-level module managing dialogue-turn granularity and a low-level module generating tokens within each turn. Its high-level off-policy learner computes Q-values based on accumulated dialogue rewards, guiding low-level updates via actor-critic methods. ArCHer achieves substantial sample efficiency, outperforming non-hierarchical baselines by approximately 100× in terms of training samples.

DeepMind's SCoRe [195], an extension of ArCHer, explicitly targets self-correction by training models on synthetic dialogues to recognize and rectify errors across two-turn sequences. SCoRe significantly advances benchmarks for self-correction in math and coding tasks. Overall, hierarchical frameworks like ArCHer and SCoRe effectively address multi-turn credit assignment through structured training and specialized rewards.

Off-Policy Value Optimization   Another promising direction frames multi-turn LLM alignment as an off-policy value-learning problem. REFUEL [196] addresses covariate shift issues in multi-turn RLHF by iteratively training a Q-value function on accumulated trajectories, directly regressing future cumulative rewards. Unlike standard on-policy methods, REFUEL continuously leverages self-generated dialogues, ensuring accurate value estimation for inference-time states. Empirically, REFUEL outperformed state-of-the-art methods like DPO on long dialogue benchmarks; notably, an 8B-parameter REFUEL model surpassed a 70B-parameter model fine-tuned via single-turn methods, highlighting the value of explicit long-term reward modeling for improved conversational coherence.

Benchmarks & Evaluation   Effectively assessing multi-turn RL algorithms requires specialized benchmarks that capture long-term interaction and credit assignment. LMRL-Gym [242] introduces a suite of interactive language tasks—ranging from open-ended dialogues to text-based games—to measure critical skills such as intentionality, information-gathering, and strategic planning over extended interactions. By providing both offline and online RL evaluation frameworks, LMRL-Gym enables systematic benchmarking of multi-turn RL improvements. Similarly, SWEET-RL [243] introduces ColBench, a set of collaborative human-AI tasks emphasizing multi-turn dialogue and reasoning, where traditional RL struggles due to delayed, sparse feedback. SWEET-RL addresses this by training a turn-wise advantage critic using additional training-time signals, significantly improving performance on collaborative tasks such as coding and UI design conversations. Together, LMRL-Gym and SWEET-RL highlight the importance of tailored benchmarks and evaluation methods to reliably advance LLMs' multi-turn conversational abilities.

3.1.4   New Architectures

Some researchers have questioned whether inherent limitations of the Transformer architecture itself might be responsible for observed performance degradation in complex, multi-turn scenarios [244]. Motivated by this concern, in addition to advances in contextual learning, supervised fine-tuning, and reinforcement learning, recent efforts have explored optimizing LLM architectures to specifically enhance performance in multi-turn interactions.

Cached Transformers [197] introduce a novel model architecture that extends the traditional Transformer by incorporating a Gated Recurrent Cache (GRC) into its self-attention mechanism. This differentiable memory cache compresses historical token representations into fixed-length vectors that are continuously updated through gating, enabling the model to attend efficiently to both past and current tokens. By effectively capturing long-range dependencies without significant computational overhead, Cached Transformers improve performance on various language tasks. In multi-turn conversations, this mechanism can help LLMs maintain coherent and contextually rich dialogue histories by providing a persistent, efficient memory of earlier interactions, thereby enhancing the model's ability to reference and build upon past conversational turns.

Beyond caching, researchers are exploring stateful transformer designs that maintain an internal dialogue state. For example, Wu et al. (2023) [198] introduced a memory-augmented transformer (MemBART) that carries a "memory state" alongside the normal model hidden state, updated at each turn. MemBART employs a dual attention stream to separately handle memory reading and writing, along with a residual gated update mechanism that determines how much past information to retain versus update at each timestep. This design allows the model to efficiently store and retrieve important historical context without needing excessively large input windows, thereby enhancing its ability to maintain coherent multi-turn conversations with lower computational overhead and improved latency.

Recent advances in long-context language modeling have also leveraged recurrence to extend Transformers' effective context. Transformer-XL [199] reuses hidden states from previous segments with a novel relative positional encoding to mitigate context fragmentation. The Recurrent Memory Transformer [200] augments this idea by inserting dedicated memory tokens into the sequence, which are recurrently updated to store global information. He et al. (2024) [201] presents a transformer framework that mimics the brain's memory hierarchy by segmenting information into sensory, short-term, and long-term layers, thereby facilitating the processing of lengthy contexts with lower computational overhead. Similarly, RWKV [202] reformulates attention in an RNN-like manner, achieving linear complexity while retaining Transformer parallelism. Enhancing

RWKV-based models [203], recent work introduces adaptive gating and position-aware convolutional shifts to dynamically regulate inter-token information flow. All four approaches share the common goal of overcoming fixed-length limitations by propagating information across segments, thereby capturing long-term dependencies more efficiently while balancing training parallelism with inference efficiency.

Integrating memory and recurrence mechanisms allows transformers to capture long-term dependencies across segments, improving multi-turn dialogue coherence. These advances enhance model performance and efficiency, making conversational AI more robust and scalable.

3.2   External Integration Approaches

Beyond model-centric approaches, another prominent strategy involves external integration methods, where LLMs are augmented with additional resources to enhance their performance in multi-turn interactions. These approaches incorporate external tools such as memory augmentation, retrieval-augmented generation (RAG), and knowledge graphs to facilitate external information retrieval, verification, and reasoning. By leveraging these external integrations, LLMs can mitigate compounding errors and misinformation propagation commonly encountered in extended interactions, thereby significantly improving their reliability, accuracy, and consistency in multi-turn settings.

3.2.1   Memory-Augmented Methods

Memory-augmented methods address the challenge of maintaining context over extended conversations by equipping LLMs with mechanisms to store and recall past interactions. These techniques help models correct misinterpretations, reduce repeated errors, and adapt to evolving dialogue, ultimately fostering more coherent multi-turn conversations.

MemPrompt [204] demonstrates an early external memory approach where the system records pairs of misunderstood inputs and corresponding user corrections in a dynamic memory bank. When a similar query arises later, the stored corrective feedback is retrieved and appended to the prompt, guiding the model toward a more accurate interpretation. In a related effort, Wu et al. (2024) [205] proposes a unified framework that decomposes memory design into indexing, retrieval, and reading stages. Their work introduces detailed optimizations—including session decomposition, fact-augmented key expansion, and time-aware query expansion—that significantly enhance memory recall and downstream question-answering accuracy, even for models designed with extended contexts.

Taking inspiration from human cognition, Pink et al. (2025) [206] argues for the integration of episodic memory into LLMs. Their framework emphasizes long-term storage, explicit reasoning, and instance-specific detail capture, outlining four research directions: discretizing continuous interactions into episodes, retrieving relevant past experiences, consolidating episodic traces into generalized knowledge, and establishing benchmarks for evaluation. Meanwhile, hierarchical memory structures offer another avenue for improvement. Rezazadeh et al. (2024) [207] introduces MemTree, a dynamic tree-based system that aggregates dialogue content into hierarchical nodes to enable efficient retrieval and improved long-term reasoning.

Together, these works illustrate that integrating external, episodic, and hierarchical memory mechanisms can substantially enhance the consistency and contextual understanding of LLMs in multi-turn conversations.

3.2.2   Retrieval Augmented Generated

Retrieval-augmented generation (RAG) is an advanced framework introduced by Lewis et al. (2020) [245] that enhances the capabilities of NLP models by integrating external knowledge sources into the generation process. The RAG integration allows LLMs to access up-to-date and domain-specific information, improving the accuracy and relevance of their responses. It also helps mitigate AI hallucinations by grounding outputs in reliable data.

Studies have applied retrieval-based architectures to multi-turn dialogue systems, enabling the generation of more informative and factual responses by conditioning on both the user's input and relevant external documents. For instance, Wizard of Wikipedia [208] incorporated retrieval into each dialogue turn, resulting in significantly

higher factual accuracy and user engagement compared to non-retrieval baselines. Similarly, Komeili et al. (2021) [209] explicitly generates search queries from the dialogue context to pull in up-to-date knowledge (e.g. via internet search) and then conditions the response on the retrieved results. Other variants, like [210], retrieve from a fixed knowledge base or enterprise documents, using dense vector search to find passages related to the user's query, where the generator model then conditions on both the dialogue context and the fetched evidence to produce a response. BlenderBot 2.0 [211] extended the RAG idea by integrating both internet search and a long-term memory component, allowing the model to sustain coherent conversations across multiple turns and sessions while retrieving past facts when needed. This allows the system to handle context dependencies that go beyond the immediate dialogue window. Overall, RAG offers a principled mechanism for overcoming the context length and memory limitations of LLMs, making it a valuable technique for improving dialogue coherence, answer accuracy, and user trust in multi-turn interactions.

To assess the effectiveness of RAG systems in multi-turn conversational settings, benchmarks such as MTRAG [246], CORAL [247], and RAD-Bench [248] have been developed. MTRAG comprises 110 conversations averaging 7.7 turns each, totaling 842 tasks across four domains and incorporates diverse question types (factoid, comparison, explanation, etc.), varying answerability (answerable, partially answerable, unanswerable), and multi-turn dynamics (follow-up and clarification questions), providing a comprehensive evaluation framework for multi-turn RAG systems. The benchmark emphasizes active retrieval, where relevant documents are dynamically fetched based on user inquiries throughout the conversation, simulating a more realistic conversational experience. Similarly, CORAL offers a large-scale benchmark designed to assess RAG systems in realistic multi-turn conversational settings, supporting tasks such as passage retrieval, response generation, and citation labeling. RAD-Bench provides a frame to assess multi-turn LLMs' capabilities in augmented generation with retrieved context in multi-turn scenarios with both Retrieval Synthesis and Retrieval Reasoning mechanisms. The evaluation framework contains 89 multi-turn question samples sampled across six practical scenarios inspired by human-LLM multi-turn dialogue interactions requiring retrieved context to complete tasks.

3.2.3   Knowledge Graph Integration

Graph neural networks (GNN), especially when integrated with knowledge graphs (KG), have emerged as effective approaches for enhancing the multi-turn reasoning and interaction capabilities of LLMs. They notably improve tasks such as tracking entities and resolving coreferences, managing dialogue context structures, and enabling more robust reasoning over structured knowledge.

Many research in this field focuses on deriving graph-structured embeddings (typically using GNN), which are then integrated during the continuous pre-training or fine-tuning stages of LLMs [212, 213, 214]. Some of them utilize existing KGs for obtaining commonsense or domain knowledge that is lacking between conversations. Wang et al. (2024) [212] address the rational response candidate selection problem, where commonsense—often necessary—is explicitly omitted during human-LLM conversational interactions. The system builds KGs based on external commonsense sources. Both the GNN and LLM are jointly fine-tuned using a response selection loss, which measures how effectively the model ranks the correct response among several candidates. Jain et al. (2024) [213] construct a dynamic graph representation of both the ongoing conversation and Wikidata KG (with text, tables, and infoboxes) [249]. The method jointly trains GNN and LLM with additional memory module, helping LLM to maintain context across multiple dialogue turns and enhances its reasoning capabilities for QA answering from heterogeneous sources. In the healthcare domain, Gao et al. (2025) [214] integrates LLMs with Unified Medical Language System–based KGs [250] using a graph isomorphism network to rank and identify knowledge pathways relevant to the clinical contexts of patients, thus enhancing diagnostic processes.

Being able to refer entities and coreferences from past dialogues and conversations can help LLMs with reasoning and instruction following abilities. Consequently, several research initiatives are now focusing on constructing or modifying KGs to integrate historical dialogue data. This integration helps to create richer, context-aware models that not only understand isolated queries but also leverage prior conversational context for improved performance. The works of Wang et al. (2024)[212] and Jain et al. (2024)[213], as discussed above, append existing KGs with entities and relationships from past conversations. SURGE [215] builds KGs that specifically encode relevant knowledge for ongoing conversation, with triplets consisting of entities and their relations as items. The GNN is incorporated to perform multi-hop reasoning and connect disparate

pieces of information. Tan et al. (2025) [216] propose to prune existing KGs to remove irrelevant information, incorporating improved graph structures with additional prompting and LLMs to find candidate paths in multi-hop and multi-entity QAs.

Besides directly intergating graphs into training or finetuning LLMs, several research focus on Graph RAG for enhancing reasoning and flexibility in dynamic conversations. Typically, Graph RAG extracts entities and their relationships from documents or dialogues to build a KG. Then, it traverses the knowledge graph to retrieve subgraphs. The retrieved graph information is then integrated into the LLM's prompt, which allows LLM to generate a response that is richer, more coherent, and better grounded in factual knowledge [219, 217, 218].

3.3 Agent-Based Approaches

  An emerging paradigm in enhancing multi-turn interactions is using Large Language Models (LLMs) as agents, a framework commonly termed as LLM-based agents. In contrast to traditional static use of language models (where an LLM passively responds to inputs), an LLM-based agent proactively engages in iterative loops of reasoning, planning, and interacting with external resources or environments to accomplish complex goals over multiple conversational turns.

This subsection reviews recent works on LLM-based agents, grouped into two broad categories: (1) single-agent systems (one LLM agent interacting iteratively with an environment or tools), and (2) multi-agent systems (multiple LLM agents collaboratively interacting or engaging in structured multi-turn dialogues).

3.3.1 Single Agent Approaches

Single-agent approaches use one LLM as a sole agent that iteratively interacts with external tools, environments, or its own internal reasoning trace to improve multi-turn performance. These methods enhance the agent's ability to answer complex queries, perform decision-making, or solve long-horizon tasks by breaking problems into iterative steps. Key themes include interleaving reasoning with actions, using tools or external APIs, and self-refinement via feedback.

One influential example is the ReAct by Yao et al. (2023) [251], which integrates explicit reasoning steps with action executions, enabling the agent to interact dynamically with knowledge bases or external APIs. In this paradigm, the agent proactively decides when to retrieve external information, significantly reducing hallucination issues in open-domain question answering and interactive decision-making tasks. Empirical evaluations showed notable performance gains on benchmarks such as HotpotQA [252], FEVER [253], ALFWorld [254], and WebShop [255] compared to static prompting or reinforcement learning baselines.

Extending this concept, the Toolformer framework [221] trains an LLM to autonomously determine the necessity and timing of external tool usage, such as calculators or web search APIs, by inserting specialized API-call tokens within its generations. By learning to integrate tool use in a self-supervised manner, Toolformer achieves substantial accuracy improvements on zero-shot arithmetic, knowledge retrieval, and translation tasks. This method demonstrates how explicit training for proactive tool invocation can significantly enhance multi-turn problem-solving without increasing model size. Similarly, HuggingGPT [222] employs an LLM as a centralized orchestrator that autonomously decomposes complex user requests into manageable sub-tasks delegated to specialized external models. Although HuggingGPT primarily emphasizes multi-modal and multi-model coordination, its relevance here lies in showcasing an LLM's capability for sophisticated planning and iterative decision-making across multi-step interactions, reinforcing the power of agent-based decomposition strategies.

The Reflexion framework introduced by Shinn et al. (2023) [223] further pushes the concept of single-agent improvement by enabling self-reflective feedback loops. Instead of relying on traditional gradient-based learning methods, Reflexion employs verbal reinforcement learning, allowing the LLM to record textual reflections of its previous mistakes into episodic memory. Subsequent interactions utilize these reflections for iterative self-improvement, significantly boosting the agent's performance on coding and sequential decision-making benchmarks, surpassing even highly advanced models like GPT-4 in single-pass accuracy on code-generation tasks such as HumanEval. Extending this iterative self-improvement paradigm, the Voyager agent by Wang et al. (2023) [224] exemplifies lifelong learning capabilities within an open-ended virtual environment (Minecraft). Voyager autonomously engages in continuous loops of planning, code generation,

environment-based execution, and observation, incrementally refining its skill repository through persistent, multi-turn interactions. This approach demonstrates impressive exploration efficiency, rapid skill generalization, and improved cumulative task success compared to prior single-agent baselines, highlighting the effectiveness of iterative, experience-driven knowledge acquisition in complex, open-ended environments.

For comprehensive agent evaluation, Liu et al. (2023) [256] introduced AgentBench, a rigorous benchmark designed to assess LLMs on agentic tasks—contexts requiring LLMs to make decisions and execute actions within interactive environments to accomplish specific goals. AgentBench features eight diverse simulated environments, spanning embodied navigation, interactive games, tool utilization, reasoning puzzles, and web interaction, enabling systematic evaluation of LLMs' reasoning, planning, and decision-making capabilities across extended interaction sequences. The authors conducted an extensive evaluation of 27 LLMs, including both open-source and proprietary API-based models, revealing substantial performance differentials. Their findings demonstrate that leading proprietary models such as GPT-4 and Claude exhibit remarkable proficiency in functioning as coherent agents in complex, long-horizon tasks, often successfully completing scenarios that confound other models. Nevertheless, even these advanced systems demonstrate considerable limitations, with a significant performance gap persisting between them and the most capable open-source alternatives.

Collectively, these single-agent works underscore a paradigm shift toward proactive, iterative interaction of LLMs with external tools, environments, and internal memory states, enabling significantly enhanced reasoning, decision-making, and self-improvement capabilities across diverse multi-turn interaction settings.

3.3.2   Multi-Agents Approaches

Multi-agent approaches involve multiple LLM-based agents collaboratively interacting to jointly solve complex problems. Drawing inspiration from human teamwork and structured debate, these methods leverage collective intelligence to surpass single-agent capabilities. Recent works can be grouped into three main categories: (1) role-based collaborative agents, (2) debate-based approaches, and (3) dynamic agent composition.

Role-based Collaborative Agents   Role-based frameworks assign distinct roles to agents, guiding structured multi-turn dialogues. CAMEL [225] by Li et al. (2023) uses inception prompting to enable autonomous role-playing between agents (e.g., user and assistant), successfully generating conversational data and revealing emergent cognitive behaviors. Extending structured cooperation, ChatDev [226] by Qian et al. (2024) organizes multiple specialized agents into virtual software-development teams (architect, coder, tester) that sequentially interact, significantly improving software quality over single-agent baselines such as GPT-4. Similarly, Dong et al. (2024) [227] proposed structured collaboration among analyst, coder, and tester agents, greatly enhancing code-generation accuracy. Further emphasizing clear workflows, MetaGPT [228] encodes human-like Standard Operating Procedures into prompts, ensuring systematic verification and substantially reducing cascading errors, thus outperforming simpler multi-agent setups. Similarly, the Mixture-of-Search-Agents (MoSA) approach [257] leverages multiple LLMs that propose and refine solutions in tandem. Each agent can independently suggest a next step or critique another agent's partial solution, and through this multi-turn collaboration the group avoids single-model blind spots. MoSA demonstrated higher accuracy on challenging math sets (e.g. a MATH benchmark subset) than any single model working alone.

Chen et al. (2024) [258] propose BUTTON ("Bottom-Up then Top-Down"), a systematic method to train LLMs for executing multi-step tool use or function calls in conversational contexts. In the bottom-up phase, BUTTON generates simple atomic instruction–function pairs derived from real-world scenarios. The top-down phase creates a simulated environment involving interactions among user agents, assistant agents, and tool agents, emulating realistic multi-turn dialogues. Using this approach, the authors develop BUTTONInstruct, a dataset comprising 8,000 multi-turn dialogues with compositional function-calling tasks. Models fine-tuned on BUTTONInstruct exhibit significant improvements in planning and correctly executing complex sequences of API calls.

Debate-based Approaches   Debate-based methods enhance reasoning accuracy by structuring iterative critiques among agents. Du et al. (2023) [229] showed multi-agent debate significantly improves factual

correctness and logical reasoning, outperforming single-agent solutions on mathematical and strategic tasks through structured back-and-forth critique. Complementarily, Generative Agents by [259] explore multi-turn social simulations, demonstrating emergent realistic behaviors (planning, social interactions) that arise naturally through iterative dialogue, highlighting the broader implications of structured deliberation in multi-agent interactions.

Dynamic Agent Composition   Dynamic agent-composition methods create flexible agent teams tailored to specific tasks. AutoAgents [230] automatically generates specialized agents, accompanied by an observer agent that monitors and adjusts the interaction dynamically, surpassing fixed-role approaches on heterogeneous tasks. Similarly, AgentVerse [231] provides a versatile platform for agents to dynamically join or leave teams, enhancing adaptability and performance, while simultaneously producing beneficial emergent behaviors like negotiation and consensus formation.

Overall, these multi-agent frameworks collectively demonstrate how structured roles, deliberative debates, and dynamic compositions significantly enhance multi-turn interactions, paving the way toward robust, adaptive, and human-aligned AI collaboration.

Despite their promising results, multi-agent LLM systems still exhibit notable challenges and limitations. Recent critiques commonly highlight recurring issues such as role misassignments [260, 261], where agents frequently misunderstand or deviate from their assigned responsibilities, causing redundancy, confusion, or omission of crucial tasks. Inefficient communication overhead and compounded errors [260, 261, 262] are also prevalent, stemming from extensive interactions required for coordination among agents, which not only increases computational costs but also propagates and amplifies mistakes made by individual agents. Additionally, inadequate verification mechanisms [260, 261] lead to unreliable or inconsistent outcomes, as agents often fail to effectively validate intermediate or final results, causing errors to remain undetected or improperly resolved. Emergent risks such as miscoordination, conflicts, and unintended collusion among autonomous agents [262] further complicate multi-agent interactions, highlighting deeper issues related to the unpredictability and complexity inherent in systems of multiple interacting entities. Collectively, these insights underline the necessity for future research to prioritize robust coordination protocols, scalable memory-sharing mechanisms, adaptive verification strategies, and comprehensive evaluation frameworks to effectively harness the full potential of multi-agent collaborative intelligence, enhancing multi-turn interactions and paving the way toward robust, adaptive, and human-aligned AI collaboration.

4 Open Challenges

Despite remarkable advancements in large language models that have silently solved many previously formidable AI obstacles—such as understanding physical rules of the real world, generating human-like text, extending context memory length, and demonstrating creativity—significant challenges persist specifically in multi-turn interactions that limit their robustness, reliability, and alignment with user expectations. While earlier sections of this survey have reviewed common multi-turn tasks and discussed state-of-the-art methods aimed at improving performance, it is crucial to recognize that existing approaches fall short of addressing all complexities comprehensively. As illustrated in Figure 2, we systematically categorize these open challenges into six major areas: Context Understanding, Complex Reasoning, Adaptation & Learning, Evaluations, and Ethical & Safety Issues, each with their associated sub-challenges. By highlighting these critical limitations and under-explored areas, we aim to guide future research efforts and encourage the development of LLM multi-turn systems that can maintain coherence, context-awareness, adaptability, and ethical soundness over prolonged interactions.

Figure 2: Illustration of open challenges in multi-turn LLM interactions, categorized into six major areas: Context Understanding, Complex Reasoning, Adaptation & Learning, Evaluations, Ethical & Safety Issues, and associated sub-challenges.

4.1   Context Understanding & Management

4.1.1   Context Retention & Coherence

LLMs struggle to maintain long-term context, leading to incoherence or contradictions in extended dialogues. As conversations grow, models often forget or confuse earlier details, causing lapses in consistency. Recent evaluations show that increasing the distance between a query and its relevant prior context degrades model performance [27]. Even advanced chat-oriented models exhibit only modest gains in multi-turn coherence despite larger context windows and alignment tuning [30]. Multi-turn challenge benchmarks reveal that instruction retention and self-coherence remain difficult for current models – they frequently fail to remember instructions or maintain a consistent narrative over several turns [72]. This indicates that preserving conversational state across turns is an open problem.

4.1.2   Anaphora & Ellipsis Resolution

Multi-turn dialogue requires resolving pronouns, omissions, and references to earlier utterances. Today's LLMs can easily misinterpret sentences like "That one looks good" or "I did it," especially in complex dialogues with many entities. Models often falter at linking such utterances to the correct antecedents in prior context. For example, keeping track of characters or user preferences mentioned only implicitly can lead to confusion in later turns. Recent dialogue evaluations emphasize inference memory of user-provided information as a key gap: models frequently forget or mix up attributes (e.g. a user's name or previously stated facts) when referred to later [72]. Robust anaphora and ellipsis resolution, akin to co-reference resolution in dialogue, remains an open challenge for maintaining coherence across turns.

4.1.3   Ambiguity Recognition & Clarification

When user inputs are ambiguous or underspecified, aligned LLMs tend to either over-hedge (give vague answers) or implicitly guess the user's intent, rather than ask clarifying questions. This is a fundamental limitation in context understanding. Chen et al. (2024) [192] report that current conversational agents often do not adequately disambiguate – if faced with an unclear request, they guess or provide a generic response instead of seeking clarification. For instance, an instruction like "Tell me about that report" might prompt an arbitrary guess about which report, rather than a question to identify the reference. The lack of proactive clarification leads to misunderstandings that compound over a dialogue. Developing LLMs that can recognize ambiguity and ask targeted follow-up questions (as humans do) is an open research direction.

4.2   Complex Reasoning Across Turns

4.2.1   Error Propagation & Compounding

Errors or misunderstandings made by the model (or user) in one turn can be carried into subsequent turns, often amplifying into larger reasoning failures. If an LLM answers a question incorrectly or the user introduces a false premise, the misinformation may persist through the dialogue. Studies have found that most LLMs perform worse in multi-turn settings than in single-turn, partly due to sensitivity to dialogue history – once a mistake enters the context, the model is likely to build on it in later reasoning [27]. This compounding error effect has been noted in sequential decision tasks as well, where minor mistakes accumulate over interaction trajectories. In other words, the model lacks a robust mechanism to correct or "forget" errors mid-dialogue. This challenge is pronounced when the model misinterprets a user question early on or hallucinates a fact – without intervention, subsequent turns will continue down the wrong path. Designing models that can detect and self-correct such errors (or accept user corrections) remains an open problem.

4.2.2   Topic Switching & Discontinuous Reasoning

Real conversations often shift topics or return to earlier subjects after a detour. Such discontinuities pose difficulties for LLMs, which may either wrongly carry over context from the previous topic or fail to recall the relevant earlier context when a topic is resumed. For example, after discussing topic A, then briefly B, a question about A again might confuse the model or lead to a non-sequitur response. Current models have

trouble with this kind of context management – they do not truly pause and resume threads of conversation like humans. Instead, models either treat each turn in isolation (losing the prior thread) or conflate unrelated context. Effective attention re-allocation is required for handling topic switches, yet remains unsolved. Recent benchmarks like MT-Bench [21] and MT-Bench-101 [30] contain category shifts to test this, and results indicate that even top-tier models often break coherence when the conversation flow is non-linear. Better mechanisms for tracking multiple conversation topics and context segmentation are needed.

4.2.3   Proactive Information Seeking

In complex interactive settings such as medical diagnosis, troubleshooting, or tutoring, an ideal conversational agent should ask follow-up questions and guide the dialogue to gather missing information. Today's general-purpose LLMs, however, are largely reactive — they answer questions when asked but rarely volunteer clarifying or exploratory queries unless explicitly prompted. This limits their effectiveness in diagnostic and problem-solving dialogues, where the burden of inquiry should not fall solely on the user. One challenge is that LLMs have been primarily trained on single-turn Q&A or straightforward instruction-following, so they lack exposure to multi-turn dialogue policies (the strategy of when to ask, when to inform) [192]. Data for such scenarios is scarce, and models tend to default to giving an immediate answer even if the query is underspecified or potentially incomplete. For instance, in a medical context, a user might say "I feel sick," and a good agent would ask about specific symptoms; many LLMs instead try to provide a diagnosis or generic advice without sufficient clarification. Some recent work attempts to teach LLMs to clarify and confirm (e.g. asking follow-ups in ambiguous text-to-SQL tasks [192], but robust proactive dialogue behavior in arbitrary settings (medical, legal, technical diagnostics) is still an open challenge.

4.2.4   Multilingual & Code-Switching Scenarios

Multi-turn interactions that involve multiple languages (or mixing languages) introduce additional complexity. An LLM may handle a single-turn query in a non-English language reasonably well, yet maintaining context across languages in a conversation is far more difficult. Code-switching – where users alternate between languages in successive turns or even within a turn – often confuses models, leading to incorrect translations or lost context. For example, a user might ask a question in English, then follow up in Spanish; current models might not consistently link the follow-up to the previous question or could respond in the wrong language. Moreover, evaluation of multi-turn dialogue abilities has been heavily skewed toward English, and many models show performance drop-offs in less dominant languages. Cultural and linguistic knowledge may not transfer seamlessly across turns. There are also safety implications: mixing languages can bypass certain content filters or exploit weaknesses in a model's multilingual alignment. Yoo et al. (2024) [263] found that code-switching in prompts can elicit undesirable behaviors from LLMs that would not surface with English-only inputs, revealing gaps in multilingual understanding and alignment. This underscores that handling multi-turn dialogues in a culturally and linguistically robust way is still an open challenge.

4.3   Adaptation & Learning

4.3.1   Dynamic Preference & Objective Adaptation

Unlike human assistants, LLMs have no true long-term learning or personalization within a conversation – they rely only on the provided context window. Adapting to a user's preferences, tone, or goals over the course of an interaction is thus difficult. Current chat models will follow explicit style instructions (e.g. "please reply formally") but they often miss subtler cues or changes in user preferences over time. For example, if a user seems confused, an ideal assistant might proactively simplify its language in subsequent turns; today's LLM might continue at the same level unless directly told to adjust. Personalization is largely constrained to what the user explicitly repeats every turn. Developing dynamic adaptation is challenging because it requires the model to infer and remember user preferences or context that persist across turns (or even across sessions) without explicit reminders. Some initial research has explored allowing models to update a pseudo-persona or profile as they converse, but ensuring this happens reliably and safely is unresolved. The key difficulty is that the model's parameters are fixed during inference – any adaptation must come from processing the conversation context itself. This leaves LLMs prone to either forgetting user preferences or overshooting (applying a style

inappropriately broadly). Effective techniques for on-the-fly adaptation (short of fine-tuning) are still in early stages.

4.3.2   Knowledge Adaptation

Human conversation partners learn and adapt during dialogue – for instance, absorbing new facts the user provides or adjusting to corrections. Current LLMs, in contrast, have a fixed knowledge base at deployment and lack the ability to truly update their beliefs or knowledge states during a conversation. If a user informs the chatbot of a new piece of information in turn 5, the model might use it temporarily in subsequent responses (since it resides in the dialogue context), but this information isn't incorporated into any lasting memory. Once the session resets, the "learning" is lost. Enabling continual learning or persistent memory in interactive settings is an important open challenge [264]. Approaches like retrieval-augmented generation partially address this by allowing the model to fetch relevant facts from an external database, but they still rely on pre-established knowledge sources. True on-the-fly learning would mean the model can update its internal representations based on user-provided data or feedback, without retraining from scratch. This is difficult due to the risk of catastrophic forgetting (updating a neural model on new data can degrade prior knowledge) and the danger of model misuse (users could insert false or malicious information). Some recent studies have explored using external memory modules or dynamic context augmentation as a surrogate for learning, effectively writing important new facts to a scratchpad that persists through the dialogue. Nonetheless, achieving a good balance between plasticity (learning new information) and stability (not overriding established correct knowledge or safety protocols) in real-time is an unsolved problem.

4.3.3   Robustness to Misinformation & Adversarial Inputs

In multi-turn settings, users may intentionally or unintentionally introduce misleading information, adversarial prompts, or attempts to derail the model (so-called jailbreaks). A major challenge is making LLMs resilient to these tactics. Over multiple turns, a malicious user can gradually manipulate the context or employ social-engineering style approaches to trick the model into breaking rules. Recent studies demonstrate that even top-tier aligned models (including GPT-4) can be gradually coerced into unsafe or policy-violating outputs through clever multi-turn strategies [265]. Attackers obscure harmful intents across several turns and lead the model astray with fabricated context, succeeding where a single-turn attack would fail. In other cases, a user might feed the model subtly incorrect facts each turn; the model typically lacks the fact-checking capability to catch these and will propagate the falsehoods. There is also the issue of prompt leakage – models revealing hidden system instructions or private content when pressed across turns (as discussed in the context of privacy below). All these adversarial scenarios highlight that sustained interactions open new vectors for exploitation. Current LLMs do not have robust defenses against multi-turn manipulation, beyond static safety training which can be sidestepped when the attack is staged gradually. Building conversational agents that can detect inconsistent or malicious user input and respond safely (or refuse) without derailing the interaction remains an open research problem.

4.4   Evaluations

4.4.1   Scalable Data Curation

Achieving scalable data collection for multi-turn interactions remains a significant challenge at the intersection of data engineering and model training. The community continues to explore a range of promising approaches, including leveraging domain-specific real dialogues, establishing collaborations to access new data sources, and developing innovative pipelines for synthetic data generation. Despite these efforts, scalability in curating conversational datasets is still an open and pressing issue.

As we discussed in §2.1.1 and §2.2.2, as dataset sizes continue to grow, ensuring scalable, high-quality conversational data has proven crucial yet remains challenging. Among the domains we've examined, healthcare uniquely benefits from access to extensive real-world multi-turn, patient-doctor conversational datasets. These datasets, derived directly from authentic patient-doctor interactions, inherently provide rich conversational diversity and realism, making them highly suitable for continuous pre-training, SFT, and RLHF. However, replicating such high-quality data collection in other domains is challenging. To meet the rising

demand for specialized conversational data across various fields, increased collaboration is required among researchers, governmental bodies, and institutional organizations to develop scalable systems for data collection. Unlike healthcare, most other domains lack established protocols for recording and anonymizing real-world conversations, representing a significant obstacle in dataset development.

Addressing these limitations, recent research has explored strategies for enriching and expanding existing datasets. For example, single-turn or limited-turn interactions can be converted into rich multi-turn dialogues either manually or via automated frameworks. Notable advancements include specialized methods and frameworks [120, 266, 165], as well as synthetic generation and rewriting techniques leveraging LLMs [122, 81, 32, 267]. However, these methods still fall short of ideal performance, and multiple studies have documented the drawbacks of synthetic data [268, 269, 270]. Even the most sophisticated scalable data approaches suffer from inherent limitations: they often lack the unexpected nuances and diversity of real-world conversations, risk compounding errors (model-generated dialogues might amplify factual inaccuracies or unnatural patterns if not monitored), introduce subtle inconsistencies, fail to capture genuine user spontaneity, or overrely on generic conversational patterns.

Going forward, research must focus on not just increasing the volume of multi-turn data, but doing so in a way that maintains or improves quality. This includes developing better automated dialogue generation that mimics human conversational nuances, creating robust filtering and refinement processes to catch errors in synthetic data, and establishing shared repositories of multi-turn dialogues that cover a broad range of domains and interaction styles. Only through such comprehensive efforts can we hope to meet the data demands of ever more sophisticated multi-turn LLM systems.

4.4.2   Metric Designing

Beyond datasets, the metrics and criteria for judging multi-turn interactions remain open for improvement. Traditional metrics (e.g. single-turn accuracy or BLEU for responses) fail to capture the nuanced qualities of a dialogue. There is a need for more fine-grained metrics that evaluate specific aspects of multi-turn performance, as well as holistic metrics for entire conversations:

Capability Evaluation at A Fine-Grained Level   Multi-turn benchmarks like MT-Bench-101 emphasize evaluating discrete abilities (e.g. logical consistency, factual recall, politeness) at the turn level [30]. Fine-grained scoring can reveal which skills degrade over a conversation (for example, a model might maintain grammar and politeness but lose factual accuracy after many turns). However, designing such granular rubrics is challenging – it requires identifying and weighting many sub-skills, and often necessitates expert annotation or LLM-as-judge for each aspect. The field is exploring rubrics and checklists for each turn or response (as in MultiChallenge's instance-level rubrics [72], but no standardized set of fine-grained metrics has been universally adopted yet.

Evaluation on Long-Term Effectiveness   Evaluating the outcome or global quality of a multi-turn exchange is still an open problem. Ideally, metrics should reflect whether the conversation as a whole was successful (e.g. the user's goal achieved, or the model remained helpful and coherent throughout). This is hard to reduce to a single number. Some works look at conversational return-on-investment, measuring if additional turns are actually helping or just causing more confusion [27]. Others consider memory retention tests across turns or consistency checks at different conversation lengths. Still, we lack metrics for "did the model sustain high performance across 10+ turns?" or "did the conversation eventually converge to a correct/ useful outcome?". Developing evaluation measures for long-range, interactive effectiveness (perhaps analogous to task success rates in dialogue systems) is an ongoing research area.

Cultural & Sociolinguistic Diversity   Most evaluation setups to date focus on a narrow band of interaction styles (often Western, English-centric dialogues). This limits our ability to assess model performance for diverse users. A challenge is to create evaluation metrics and scenarios that account for cultural differences in conversation, multilingual nuances (as noted earlier), and dialect or style variations. For example, a model

might perform well on a formal Q&A dialogue, but fare poorly on a casual, code-mixed conversation or when the user employs idiomatic expressions from a particular culture. Incorporating such diversity into benchmarks is crucial for a fair assessment. FairMT-Bench is a step in this direction, examining bias and fairness across demographic attributes in multi-turn settings [37]. The open problem is defining metrics that can quantify model behavior across a wide sociolinguistic spectrum – possibly by measuring bias, respectfulness, or user satisfaction for different user profiles. Ensuring our evaluation metrics are culturally inclusive and robust to different communication styles remains an important challenge.

4.4.3 LLM-based VS. Human Judging

The evaluation of multi-turn LLM dialogues has increasingly turned to LLM-as-a-judge frameworks, where a strong model (e.g. GPT-4) assesses response quality in place of human annotators. Recent benchmarks like MT-Bench and platforms such as Chatbot Arena employ LLM judges to rank chatbot responses, achieving evaluation at scale with high consistency and low cost [21]. Indeed, studies have found that GPT-4 based evaluators can align with human preferences roughly 80% of the time, approaching inter-annotator agreement levels [21]. However, replacing human judgment introduces notable biases and limitations: an AI evaluator may display self-enhancement bias (preferring answers it or a similar model generated) and verbosity bias (rewarding unnecessarily long responses), among other systematic errors [21]. Automated judges can also struggle with factual correctness and nuanced context - for instance, they might overlook subtle flaws or value judgments that a human reviewer would catch [271]. To mitigate these issues, researchers have proposed hybrid approaches that combine LLM-based scoring with human oversight. Examples include prompting AI judges with explicit rubrics or checklists to guide their evaluations, and incorporating periodic human-in-the-loop audits of model ratings. This way, the field hopes to leverage the scalability of LLM evaluators while maintaining reliability and fairness close to human standards.

4.5 Ethical & Safety

4.5.1 Bias Amplification

Multi-turn dialogues can inadvertently magnify a model's biases with each turn. If a biased assumption slips into an early response, subsequent interactions might build on it, reinforcing stereotypes or unfair perceptions. Furthermore, the interactive nature of dialogue can lead the model to adapt to the user's biases, resulting in amplified problematic content. Recent work on fairness in conversational AI underscores this risk. Fan et al. (2024) [37] find that LLMs are more prone to generating biased or discriminatory responses in multi-turn settings, showing greater bias accumulation than in single-turn prompts. For example, a slight gender bias in a first answer can become more extreme if the conversation continues in that vein. FairMT-Bench specifically evaluates such scenarios and reveals significant variability in how different models handle them – many state-of-the-art models show degraded fairness when faced with back-and-forth discussions involving sensitive attributes [37]. This challenge calls for improved bias mitigation techniques that operate across the dialogue timeline. It may not be enough to detoxify single responses; models might need mechanisms to self-monitor and adjust if they find their outputs drifting toward bias during a conversation. Additionally, data augmentation with diverse user profiles and enforcing consistency with ethical guidelines over multiple turns are potential avenues to address this. Bias in multi-turn AI interactions is an ethical concern with direct user impact, and it remains open how best to evaluate and reduce it without stifling free-flowing conversation.

4.5.2 Privacy Leakage

Prolonged conversations increase the risk of LLMs revealing sensitive information, either about the user or memorized from training data. One concern is that as a dialogue progresses, a user might share personal details that the model could later inadvertently expose or use inappropriately. Another well-documented issue is model memorization: large models sometimes recall specific training examples (such as private facts or copyrighted text) and may output them given the right prompts. Nasr et al. (2023) [272] demonstrated that LLMs can memorize chunks of their training data (like personal phone numbers, addresses, or secrets) and that adversaries can craft sequences of prompts to extract this hidden information. Multi-turn interactions give an attacker more opportunities to perform such extraction gradually, by establishing context and probing iteratively.

Even if a single-turn query doesn't trigger a leak, a series of cleverly steered queries might corner the model into revealing something it "knows". Another aspect of privacy is system prompt or context leakage – in a chat setting, the model may have system or developer instructions and conversation history in its hidden context. Through multi-turn manipulation, users have managed to get models to divulge these hidden instructions or previous messages (a notable example was prompt leaks from ChatGPT revealing its formatting rules). These vulnerabilities are exacerbated over long interactions as more sensitive info accumulates in the context. Mitigating privacy leakage requires techniques like 1) reducing memorization during training (so the model generalizes rather than stores exact data points), 2) real-time filters to catch when a response may contain sensitive content, and 3) perhaps limiting the duration or content of contexts the model can access if they contain private data. As of now, completely preventing an LLM from regurgitating memorized secrets is an open problem – efforts are ongoing to quantify and bound such leakage [272] and to develop safer training regimes. In summary, multi-turn use of LLMs raises serious privacy considerations, and researchers must ensure that extended dialogues do not become a loophole for extracting confidential information or personal data.

4.5.3   Hyper-Realism and User Perception

An emerging concern with highly advanced dialogue models is fidelity – the interactions can become so human-like that users may struggle to remember (or realize) that they are conversing with an AI. On one hand, human-like fluency and emotional resonance are goals of natural language interfaces; on the other hand, when an AI system crosses a threshold of hyperrealism, it can lead to deception (even if unintentional) and overtrust [273]. Studies have documented users developing strong personal bonds or even a therapeutic alliance with chatbots, with measured alliance levels in some cases approaching those of human therapists [274]. Such emotional engagement can provide comfort or companionship, but it also heightens the danger of overtrust and dependence, potentially displacing human relationships or giving AI undue influence over vulnerable individuals [275, 276]. In response, regulators have emphasized transparency; for example, the proposed EU AI Act mandates that AI systems interacting with people must clearly disclose their artificial identity to prevent deception [276]. Mitigation strategies in design likewise include persistent identity reminders and constrained, less too-human stylistic choices to maintain healthy user boundaries and reduce the illusion of personhood [273].

5 Conclusion

This survey has provided a comprehensive overview of the rapidly evolving landscape of multi-turn interactions with Large Language Models, making several key contributions to the field. First and foremost, we have introduced a novel task-oriented taxonomy for analyzing multi-turn LLM interactions, departing from the capability-oriented approaches prevalent in existing literature. While previous surveys have focused on isolated capabilities (such as reasoning, memory, or contextual understanding), our framework recognizes that real-world multi-turn applications require the complex collaboration of multiple capabilities working in concert. By categorizing interactions according to tasks—instruction following and conversational engagement—rather than isolated abilities, we offer a more practical taxonomy that better reflects how LLMs are deployed and evaluated in authentic contexts. This task-oriented perspective has enabled us to provide substantial analysis of critical real-world application domains, including role-playing scenarios, healthcare consultations, educational scenario, and LLM jail-breaking where the interplay of multiple capabilities determines performance.

Our analysis demonstrates that multi-turn interactions represent a fundamental paradigm shift in how we utilize and evaluate LLMs. Unlike single-turn interactions, which have dominated early LLM benchmarks, multi-turn settings more closely mirror real-world applications—from sustained dialogues to complex iterative problem-solving. These contexts demand not only factual knowledge but also context retention, coherent reasoning across turns, adaptive behavior, and robust handling of ambiguous or changing user intentions.

The improvement methodologies we've extensively reviewed span model-centric approaches (in-context learning, fine-tuning, reinforcement learning, and architectural innovations), external integration strategies (memory augmentation, retrieval mechanisms, and knowledge graphs), and agent-based frameworks (both single-agent and multi-agent systems). By providing this detailed analysis of improvement techniques, we address a significant gap in existing surveys, which have often overlooked the nuanced methodological landscape of multi-turn enhancements.

Despite remarkable progress, significant challenges remain. Context understanding issues persist, with even state-of-the-art models struggling to maintain coherence across extended conversations. Complex reasoning across turns frequently suffers from error propagation and topic-switching difficulties. Adaptation capabilities remain limited, especially regarding dynamic preference learning and knowledge updates during interaction. Our detailed organization of these open challenges provides a roadmap for future research that was previously lacking in the literature.

In conclusion, multi-turn interaction capabilities represent both a frontier challenge and a transformative opportunity for LLM research. By more closely approximating the dynamic, context-dependent nature of human communication, advances in this domain will not only enhance technical performance but also significantly improve the practical utility and trustworthiness of these systems across diverse real-world applications. Our task-oriented taxonomy, extensive examination of real-world domains, detailed analysis of improvement methodologies, and in-depth organization of open challenges collectively provide a foundation for future work in this critical area.

References

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

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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

[6] Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194*, 2024.

[7] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024.

[8] Elliot L Epstein, Kaisheng Yao, Jing Li, Xinyi Bai, and Hamid Palangi. Mmmt-if: A challenging multimodal multi-turn instruction following benchmark. *arXiv preprint arXiv:2409.18216*, 2024.

[9] Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*, 2022.

[10] Junhao Cheng, Baiqiao Yin, Kaixin Cai, Minbin Huang, Hanhui Li, Yuxin He, Xi Lu, Yue Li, Yifei Li, Yuhao Cheng, et al. Theatergen: Character management with llm for consistent multi-turn image generation. *arXiv preprint arXiv:2404.18919*, 2024.

[11] Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. *arXiv preprint arXiv:2502.10810*, 2025.

[12] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. *arXiv preprint arXiv:2311.07911*, 2023.

[13] Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating Large Language Models at Evaluating Instruction Following. *arXiv preprint arXiv:2310.07641*, 2023.

[14] Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*, 2024.

[15] Mina Valizadeh and Natalie Parde. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[16] Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. Medical Dialogue System: A Survey of Categories, Methods, Evaluation and Challenges. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2840–2861, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[17] Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. A Survey on Multi-Turn Interaction Capabilities of Large Language Models. *arXiv preprint arXiv:2501.09959*, 2025.

[18] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis. *ICLR*, 2023.

[19] Sarah E. Finch, James D. Finch, and Jinho D. Choi. Don't forget your ABC's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15044–15071, Toronto, Canada, July 2023. Association for Computational Linguistics.

[20] Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. MathChat: Converse to Tackle Challenging Math Problems with LLM Agents. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[21] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[22] John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. InterCode: Standardizing and Benchmarking Interactive Coding with Execution Feedback. *Advances in Neural Information Processing Systems*, 36:23826–23854, 2023.

[23] Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback. *arXiv preprint arXiv:2309.10691*, 2023.

[24] Yuchong Sun, Che Liu, Kun Zhou, Jinwen Huang, Ruihua Song, Wayne Xin Zhao, Fuzheng Zhang, Di Zhang, and Kun Gai. Parrot: Enhancing Multi-Turn Instruction Following for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9729–9750, 2024.

[25] Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. BotChat: Evaluating LLMs' Capabilities of Having Multi-Turn Dialogues. *arXiv preprint arXiv:2310.13650*, 2023.

[26] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C. Ho, Carl Yang, and May Dongmei Wang. EHRAgent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22315–22339, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[27] Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. MT-Eval: A Multi-Turn Capabilities Evaluation Benchmark for Large Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[28] Xing Han Lù, Zdeněk Kasner, and Siva Reddy. WebLINX: Real-World Website Navigation with Multi-Turn Dialogue, September 2024. arXiv:2402.05930 [cs].

[29] Siwei Yang, Bingchen Zhao, and Cihang Xie. AQA-Bench: An Interactive Benchmark for Evaluating LLMs' Sequential Reasoning Ability. *arXiv preprint arXiv:2402.09404*, 2024.

[30] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. MT-Bench-101: A Fine-Grained Benchmark for Evaluating Large Language Models in Multi-Turn Dialogues. *arXiv preprint arXiv:2402.14762*, 2024.

[31] Zhenwen Liang, Dian Yu, Wenhao Yu, Wenlin Yao, Zhihan Zhang, Xiangliang Zhang, and Dong Yu. MathChat: Benchmarking Mathematical Reasoning and Instruction Following in Multi-Turn Interactions. *arXiv preprint arXiv:2405.19444*, 2024.

[32] Rishabh Maheshwary, Vikas Yadav, Hoang Nguyen, Khyati Mahajan, and Sathwik Tejaswi Madhusudhan. M2Lingual: Enhancing Multilingual, Multi-Turn Instruction Alignment in Large Language Models. *arXiv preprint arXiv:2406.16783*, 2024.

[33] Yanzhao Qin, Tao Zhang, Yanjun Shen, Wenjing Luo, Haoze Sun, Yan Zhang, Yujing Qiao, Weipeng Chen, Zenan Zhou, Wentao Zhang, et al. Sysbench: Can large language models follow system messages? *arXiv preprint arXiv:2408.10943*, 2024.

[34] Youquan Li, Miao Zheng, Fan Yang, Guosheng Dong, Bin Cui, Weipeng Chen, Zenan Zhou, and Wentao Zhang. FB-Bench: A Fine-Grained Multi-Task Benchmark for Evaluating LLMs' Responsiveness to Human Feedback. *arXiv preprint arXiv:2410.09412*, 2024.

[35] Eryk Banatt, Jonathan Cheng, Skanda Vaidyanath, and Tiffany Hwu. WILT: A multi-turn, memorization-robust inductive logic benchmark for LLMs. *arXiv preprint arXiv:2410.10998*, 2024.

[36] Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-IF: Benchmarking LLMs on Multi-Turn and Multilingual Instructions Following. *arXiv preprint arXiv:2410.15553*, 2024.

[37] Zhiting Fan, Ruizhe Chen, Tianxiang Hu, and Zuozhu Liu. FairMT-Bench: Benchmarking Fairness for Multi-turn Dialogue in Conversational LLMs. *arXiv preprint arXiv:2410.19317*, 2024.

[38] Ziming Guo, Chao Ma, Yinggang Sun, Tiancheng Zhao, Guangyao Wang, and Hai Huang. Evaluating and Enhancing LLMs for Multi-turn Text-to-SQL with Multiple Question Types. *arXiv preprint arXiv:2412.17867*, 2024.

[39] Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. BigBench: Towards an Industry Standard Benchmark for Big Data Analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208, 2013.

[40] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[41] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[42] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

[43] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

[44] Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. Firm or fickle? evaluating large language models consistency in sequential interactions. *arXiv preprint arXiv:2503.22353*, 2025.

[45] Chi Han. Can language models follow multiple turns of entangled instructions? *arXiv preprint arXiv:2503.13222*, 2025.

[46] Hanwen Du, Bo Peng, and Xia Ning. SAPIENT: Mastering Multi-turn Conversational Recommendation with Strategic Planning and Monte Carlo Tree Search, October 2024. arXiv:2410.09580 [cs].

[47] Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. Towards empathetic conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 84–93, 2024.

[48] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[49] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

[50] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

[51] Vivian Keating. Zero-shot mathematical problem solving with large language models via multi-agent conversation programming. In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI*, 2024.

[52] Wei Xiong, Chengshuai Shi, Jiaming Shen, Aviv Rosenberg, Zhen Qin, Daniele Calandriello, Misha Khalman, Rishabh Joshi, Bilal Piot, Mohammad Saleh, et al. Building Math Agents with Multi-Turn Iterative Preference Learning. *URL https://arxiv. org/abs/2409.02392*, 2024.

[53] Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore, December 2023. Association for Computational Linguistics.

[54] Li Zhong, Zilong Wang, and Jingbo Shang. Debug like a human: A large language model debugger via verifying runtime execution step by step. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 851–870, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[55] Yongchao Chen, Harsh Jhamtani, Srinagesh Sharma, Chuchu Fan, and Chi Wang. Steering large language models between code execution and textual reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[56] Yuling Shi, Songsong Wang, Chengcheng Wan, and Xiaodong Gu. From code to correctness: Closing the last mile of code generation with hierarchical debugging, 2024.

[57] Kunhao Zheng, Juliette Decugis, Jonas Gehring, Taco Cohen, Benjamin Negrevergne, and Gabriel Synnaeve. What makes large language models reason in (multi-turn) code generation? *arXiv preprint arXiv:2410.08105*, 2024.

[58] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[59] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021.

[60] Xi Victoria Lin, Chenglong Wang, Luke Zettlemoyer, and Michael D. Ernst. NL2Bash: A corpus and semantic parser for natural language interface to the linux operating system. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[61] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, December 2022.

[62] Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. Taco: Topics in algorithmic code generation dataset, 2023.

[63] Yaolun Zhang, Yinxu Pan, Yudong Wang, and Jie Cai. Pybench: Evaluating llm agent on various real-world coding tasks, 2024.

[64] Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. CodeGen2: Lessons for Training LLMs on Programming and Natural Languages. *ICLR*, 2023.

[65] Yongchao Chen, Yilun Hao, Yueying Liu, Yang Zhang, and Chuchu Fan. CodeSteer: Symbolic-Augmented Language Models via Code/Text Guidance. *arXiv preprint arXiv:2502.04350*, 2025.

[66] Tianyu Zheng, Ge Zhang, Tianhao Shen, Xueling Liu, Bill Yuchen Lin, Jie Fu, Wenhu Chen, and Xiang Yue. Opencodeinterpreter: Integrating code generation with execution and refinement, 2025.

[67] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better llm agents, 2024.

[68] Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani Tur, and Jiawei Han. Instruct, not assist: LLM-based multi-turn planning and hierarchical questioning for socratic code debugging. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9475–9495, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[69] Yongchao Chen, Harsh Jhamtani, Srinagesh Sharma, Chuchu Fan, and Chi Wang. Steering Large Language Models between Code Execution and Textual Reasoning. *arXiv preprint arXiv:2410.03524*, 2024.

[70] Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*, 2025.

[71] Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. Does context matter? contextual-judgebench for evaluating llm-based judges in contextual settings. *arXiv preprint arXiv:2503.15620*, 2025.

[72] Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs. *arXiv preprint arXiv:2501.17399*, 2025.

[73] Jiwei Li et al. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.

[74] Satwik Kottur et al. Exploring personalized neural conversational models. *arXiv preprint arXiv:1706.07503*, 2017.

[75] Saizheng Zhang et al. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, 2018.

[76] Nuo Chen, Yan Wang, Yang Deng, and Jia Li. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*, 2024.

[77] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*, 2023.

[78] Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*, 2023.

[79] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[80] Zihan Tu and et al. Pippa: A large-scale dataset for persona-driven dialogues. *arXiv preprint arXiv:2308.05884*, 2023.

[81] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore, December 2023. Association for Computational Linguistics.

[82] Daniela Occhipinti, Serra Sinem Tekiroğlu, and Marco Guerini. PRODIGy: a PROfile-based DIalogue generation dataset. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3500–3514, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[83] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023.

[84] Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*, 2023.

[85] Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. RoleCraft-GLM: Advancing Personalized Role-Playing in Large Language Models, April 2024. arXiv:2401.09432 [cs].

[86] Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[87] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore, December 2023. Association for Computational Linguistics.

[88] Xu Han, Bin Guo, Yoon Jung, Benjamin Yao, Yu Zhang, Xiaohu Liu, and Chenlei Guo. Personapkt: Building personalized dialogue agents via parameter-efficient knowledge transfer. *arXiv preprint arXiv:2306.08126*, 2023.

[89] Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. Llms+ persona-plug= personalized llms. *arXiv preprint arXiv:2409.11901*, 2024.

[90] Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. *arXiv preprint arXiv:2402.13717*, 2024.

[91] Jian Wang, Chak Tou Leong, Jiashuo Wang, Dongding Lin, Wenjie Li, and Xiaoyong Wei. Instruct once, chat consistently in multiple rounds: An efficient tuning framework for dialogue. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3993–4010, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[92] Ryan Shea and Zhou Yu. Building persona consistent dialogue agents with offline reinforcement learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1795, Singapore, December 2023. Association for Computational Linguistics.

[93] Nuo Chen, Hongguang Li, Juhua Huang, Baoyuan Wang, and Jia Li. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. *arXiv preprint arXiv:2402.11975*, 2024.

[94] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[95] Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[96] Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*, 2023.

[97] Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. *arXiv preprint arXiv:2405.18027*, 2024.

[98] Qi Jia, Xiang Yue, Tianyu Zheng, Jie Huang, and Bill Yuchen Lin. SimulBench: Evaluating Language Models with Creative Simulation Tasks, September 2024. arXiv:2409.07641 [cs].

[99] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv preprint arXiv:2310.17976*, 2023.

[100] Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. Roleinteract: Evaluating the social interaction of role-playing agents. *arXiv e-prints*, pages arXiv–2403, 2024.

[101] Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. Evaluating character understanding of large language models via character profiling from fictional works. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[102] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*, 2023.

[103] Yusheng Liao, Shuyang Jiang, Yu Wang, and Yanfeng Wang. Ming-moe: Enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts. *arXiv preprint arXiv:2404.09027*, 2024.

[104] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*, 2023.

[105] Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12):5456, 2021.

[106] Yirong Chen, Zhenyu Wang, Xiaofen Xing, huimin zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. BianQue: Balancing the Questioning and Suggestion Ability of Health LLMs with Multi-turn Health Conversations Polished by ChatGPT, December 2023.

[107] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. HuatuoGPT, towards taming language model to be a doctor. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore, December 2023. Association for Computational Linguistics.

[108] Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. Automatic Interactive Evaluation for Large Language Models with State Aware Patient Simulator, July 2024.

[109] Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. DISC-MedLLM: Bridging General Large Language Models and Real-World Medical Consultation, August 2023.

[110] Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning, November 2024.

[111] Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical Camel: An Open Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. *arXiv preprint arXiv:2305.12031*, 2023.

[112] NBME FSMB. United states medical licensing examination sample test questions. https://www.usmle.org/exam-resources/step-1-materials/step-1-sample-test-questions, 2023. Accessed: 2025-03-31.

[113] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[114] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

[115] Zefa Hu, Haozhi Zhao, Yuanyuan Zhao, Shuang Xu, and Bo Xu. T-Agent: A Term-Aware Agent for Medical Dialogue Generation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, June 2024.

[116] Jiayuan Zhu and Junde Wu. Ask Patients with Patience: Enabling LLMs for Human-Centric Medical Dialogue with Grounded Reasoning, February 2025.

[117] Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. BiMediX: Bilingual Medical Mixture of Experts LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16984–17002, 2024.

[118] Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling, June 2024.

[119] Jinpeng Hu, Tengteng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. PsycoLLM: Enhancing LLM for Psychological Understanding and Evaluation, December 2024.

[120] Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[121] Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online, August 2021. Association for Computational Linguistics.

[122] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-world Multi-turn Dialogue, December 2023.

[123] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[124] Ao Dema, Yang Yunfei, Sui Zhifang, et al. Preliminary study on the construction of chinese medical knowledge graph [j]. *Journal of Chinese Information Processing*, 33(10):1–9, 2019.

[125] Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. HuatuoGPT-II, One-stage Training for Medical Adaption of LLMs, September 2024.

[126] Lulu Zhao, Weihao Zeng, Xiaofeng Shi, Hua Zhou, Donglin Hao, and Yonghua Lin. Aqulia-Med LLM: Pioneering Full-Process Open-Source Medical Language Models, June 2024.

[127] Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, Fenglin Liu, Meng Cao, Ziming Wang, Xuxin Cheng, Zhu Lei, and Zhenhua Guo. Qilin-Med: Multi-stage Knowledge Injection Advanced Medical Large Language Model, April 2024.

[128] Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song. Chimed: A chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 250–260, 2019.

[129] Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452, 2023.

[130] Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards Conversational Diagnostic AI, January 2024.

[131] Jie Xu, Lu Lu, Sen Yang, Bilin Liang, Xinwei Peng, Jiali Pang, Jinru Ding, Xiaoming Shi, Lingrui Yang, Huan Song, Kang Li, Xin Sun, and Shaoting Zhang. MedGPTEval: A Dataset and Benchmark to Evaluate Responses of Large Language Models in Medicine, May 2023.

[132] Yusheng Liao, Yutong Meng, Hongcheng Liu, Yanfeng Wang, and Yu Wang. An Automatic Evaluation Framework for Multi-turn Medical Consultations Capabilities of Large Language Models, September 2023.

[133] Ruoyu Liu, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. Interactive Evaluation for Medical LLMs via Task-oriented Dialogue System. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4871–4896, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[134] Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. MedFuzz: Exploring the Robustness of Large Language Models in Medical Question Answering, September 2024.

[135] OpenAI. Healthbench: A benchmark for evaluating large language models in healthcare, 2025.

[136] Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. Socraticlm: exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 37:85693–85721, 2024.

[137] Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure, 2024.

[138] Ben Liu, Jihan Zhang, Fangquan Lin, Xu Jia, and Min Peng. One size doesn't fit all: A personalized conversational tutoring agent for mathematics instruction. *arXiv e-prints*, pages arXiv–2502, 2025.

[139] Xiner Liu, Maciej Pankiewicz, Tanvi Gupta, Zhongtian Huang, and Ryan S Baker. A step towards adaptive online learning: Exploring the role of gpt as virtual teaching assistants in online education. *Manuscript under review*, 2024.

[140] Yuyang Ding, Hanglei Hu, Jie Zhou, Qin Chen, Bo Jiang, and Liang He. Boosting large language models with socratic method for conversational mathematics teaching. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3730–3735, 2024.

[141] Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors. *arXiv preprint arXiv:2502.18940*, 2025.

[142] Ty Feng, Sa Liu, and Dipak Ghosal. Courseassist: Pedagogically appropriate ai tutor for computer science education. In *Proceedings of the 2024 on ACM Virtual Global Computing Education Conference V. 2*, pages 310–311, 2024.

[143] Zachary Levonian, Owen Henkel, Chenglu Li, Millie-Ellen Postle, et al. Designing safe and relevant generative chats for math learning in intelligent tutoring systems. *Journal of Educational Data Mining*, 17(1), 2025.

[144] Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. Training llm-based tutors to improve student learning outcomes in dialogues, 2025.

[145] Kyle Wiggers. Learnlm is Google's new family of AI models for education. https://techcrunch.com/2024/05/14/learnlm-is-googles-new-family-of-ai-models-for-education/, 2024.

[146] Anthropic. Introducing Claude for Education. https://www.anthropic.com/news/introducing-claude-for-education, 2025.

[147] Priscylla Silva and Evandro Costa. Assessing large language models for automated feedback generation in learning programming problem solving. *arXiv preprint arXiv:2503.14630*, 2025.

[148] Qinjin Jia, Jialin Cui, Ruijie Xi, Chengyuan Liu, Parvez Rashid, Ruochi Li, and Edward Gehringer. On assessing the faithfulness of llm-generated feedback on student assignments. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 491–499, 2024.

[149] Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. Stepwise verification and remediation of student reasoning errors with large language model tutors. *arXiv preprint arXiv:2407.09136*, 2024.

[150] Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. Improving the validity of automatically generated feedback via reinforcement learning. In *International Conference on Artificial Intelligence in Education*, pages 280–294. Springer, 2024.

[151] Inderjeet Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. Closing the loop: Learning to generate writing feedback via language model simulated student revisions. *arXiv preprint arXiv:2410.08058*, 2024.

[152] Mihaela Tomova, Iván Roselló Atanet, Victoria Sehy, Miriam Sieg, Maren März, and Patrick Mäder. Leveraging large language models to construct feedback from medical multiple-choice questions. *Scientific Reports*, 14(1):27910, 2024.

[153] Qinjin Jia, Jialin Cui, Haoze Du, Parvez Rashid, Ruijie Xi, Ruochi Li, and Edward Gehringer. Llm-generated feedback in real classes and beyond: Perspectives from students and instructors. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 862–867, 2024.

[154] Sara Riazi and Pedram Rooshenas. Llm-driven feedback for enhancing conceptual design learning in database systems courses. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pages 1001–1007, 2025.

[155] Dominic Lohr, Hieke Keuning, and Natalie Kiesler. You're (not) my type-can llms generate feedback of specific types for introductory programming tasks? *Journal of Computer Assisted Learning*, 41(1):e13107, 2025.

[156] Germán Capdehourat, Isabel Amigo, Brian Lorenzo, and Joaquín Trigo. On the effectiveness of llms for automatic grading of open-ended questions in spanish. *arXiv preprint arXiv:2503.18072*, 2025.

[157] Jonas Flodén. Grading exams using large language models: A comparison between human and ai grading of exams in higher education using chatgpt. *British educational research journal*, 51(1):201–224, 2025.

[158] Milan Kostic, Hans Friedrich Witschel, Knut Hinkelmann, and Maja Spahic-Bogdanovic. Llms in automated essay evaluation: A case study. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 143–147, 2024.

[159] Xinyi Lu and Xu Wang. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 16–27, 2024.

[160] Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F Chen. Personality-aware student simulation for conversational intelligent tutoring systems. *arXiv preprint arXiv:2404.06762*, 2024.

[161] Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. *arXiv preprint arXiv:2410.04078*, 2024.

[162] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, et al. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*, 2024.

[163] Haoxuan Li, Jifan Yu, Xin Cong, Yang Dang, Yisi Zhan, Huiqin Liu, and Zhiyuan Liu. Exploring llm-based student simulation for metacognitive cultivation. *arXiv preprint arXiv:2502.11678*, 2025.

[164] Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. Exploring the potential of llm to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10(1):7, 2025.

[165] Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. Book2dial: Generating teacher-student interactions from textbooks for cost-effective development of educational chatbots. *arXiv preprint arXiv:2403.03307*, 2024.

[166] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*, 2023.

[167] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. *arXiv preprint arXiv:2310.04451*, 2023.

[168] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned?, 2024.

[169] Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts, 2024.

[170] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, Now Write an Article About That: The Crescendo Multi-Turn LLM Jailbreak Attack, 2024.

[171] Qibing Ren, Hao Li, Dongrui Liu, Zhanxu Xie, Xiaoya Lu, Yu Qiao, Lei Sha, Junchi Yan, Lizhuang Ma, and Jing Shao. Derail Yourself: Multi-turn LLM Jailbreak Attack through Self-discovered Clues, 2024.

[172] Tom Gibbs, Ethan Kosak-Hine, George Ingebretsen, Jason Zhang, Julius Broomfield, Sara Pieri, Reihaneh Iranmanesh, Reihaneh Rabbany, and Kellin Pelrine. Emerging Vulnerabilities in Frontier Models: Multi-Turn Jailbreak Attacks, 2024.

[173] Zhenhong Zhou, Jiuyang Xiang, Haopeng Chen, Quan Liu, Zherui Li, and Sen Su. Speak out of turn: Safety vulnerability of large language models in multi-turn dialogue, 2024.

[174] Xiao Liu, Liangzhi Li, Tong Xiang, Fuying Ye, Lu Wei, Wangyue Li, and Noa Garcia. Imposter.ai: Adversarial attacks with hidden intentions towards aligned large language models, 2024.

[175] Fengxiang Wang, Ranjie Duan, Peng Xiao, Xiaojun Jia, Shiji Zhao, Cheng Wei, YueFeng Chen, Chongwen Wang, Jialing Tao, Hang Su, Jun Zhu, and Hui Xue. Mrj-agent: An effective jailbreak agent for multi-round dialogue, 2025.

[176] Yifan Jiang, Kriti Aggarwal, Tanmay Laud, Kashif Munir, Jay Pujara, and Subhabrata Mukherjee. Red queen: Safeguarding large language models against concealed multi-turn jailbreaking, 2024.

[177] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024.

[178] Divij Handa, Zehua Zhang, Amir Saeidi, and Chitta Baral. When "competency" in reasoning opens the door to vulnerability: Jailbreaking llms via novel complex ciphers, 2024.

[179] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning, 2023.

[180] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. Jailbreakbench: An open robustness benchmark for jailbreaking large language models, 2024.

[181] Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Chi Zhang, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset, 2023.

[182] Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models, 2021.

[183] Erxin Yu, Jing Li, Ming Liao, Siqi Wang, Zuchen Gao, Fei Mi, and Lanqing Hong. Cosafe: Evaluating large language model safety in multi-turn dialogue coreference. *arXiv preprint arXiv:2406.17626*, 2024.

[184] Traian Rebedea, Razvan Dinu, Makesh Sreedhar, Christopher Parisien, and Jonathan Cohen. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails. *arXiv preprint arXiv:2310.10501*, 2023.

[185] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[186] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 3 2023.

[187] Chuyi Kong, Yaxin Fan, Xiang Wan, Feng Jiang, and Benyou Wang. PlatoLM: Teaching LLMs in Multi-Round Dialogue via a User Simulator. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7841–7863, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[188] Zeyu Teng, Yong Song, Xiaozhou Ye, and Ye Ouyang. Fine-tuning llms for multi-turn dialogues: Optimizing cross-entropy loss with kl divergence for all rounds of responses. In *Proceedings of the 2024 16th International Conference on Machine Learning and Computing*, ICMLC '24, page 128–133, New York, NY, USA, 2024. Association for Computing Machinery.

[189] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[190] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[191] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[192] Maximillian Chen, Ruoxi Sun, Sercan Ö Arık, and Tomas Pfister. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. *arXiv preprint arXiv:2406.00222*, 2024.

[193] Wentao Shi, Mengqi Yuan, Junkang Wu, Qifan Wang, and Fuli Feng. Direct multi-turn preference optimization for language agents. *arXiv preprint arXiv:2406.14868*, 2024.

[194] Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*, 2024.

[195] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

[196] Zhaolin Gao, Wenhao Zhan, Jonathan D. Chang, Gokul Swamy, Kianté Brantley, Jason D. Lee, and Wen Sun. Regressing the relative future: Efficient policy optimization for multi-turn rlhf, 2024.

[197] Zhaoyang Zhang, Wenqi Shao, Yixiao Ge, Xiaogang Wang, Jinwei Gu, and Ping Luo. Cached transformers: Improving transformers with differentiable memory cachde. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16935–16943, 2024.

[198] Qingyang Wu and Zhou Yu. Stateful memory-augmented transformers for efficient dialogue modeling, 2023.

[199] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[200] Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091, 2022.

[201] Zifan He, Yingqi Cao, Zongyue Qin, Neha Prakriya, Yizhou Sun, and Jason Cong. Hmt: Hierarchical memory transformer for efficient long context language processing. *arXiv preprint arXiv:2405.06067*, 2024.

[202] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

[203] Xinghan Pan. Enhancing rwkv-based language models for long-sequence text generation, 2025.

[204] Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. Memory-assisted prompt editing to improve gpt-3 after deployment. *arXiv preprint arXiv:2201.06009*, 2022.

[205] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

[206] Mathis Pink, Qinyuan Wu, Vy Ai Vo, Javier Turek, Jianing Mu, Alexander Huth, and Mariya Toneva. Position: Episodic memory is the missing piece for long-term llm agents, 2025.

[207] Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for llms. *arXiv preprint arXiv:2410.14052*, 2024.

[208] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.

[209] Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*, 2021.

[210] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

[211] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation, 2021.

[212] Yuandong Wang, Xuhui Ren, Tong Chen, Yuxiao Dong, Nguyen Quoc Viet Hung, and Jie Tang. Multi-turn response selection with commonsense-enhanced language models, 2024.

[213] Parag Jain and Mirella Lapata. Integrating large language models with graph-based reasoning for conversational question answering, 2024.

[214] Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670, February 2025.

[215] Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. Knowledge graph-augmented language models for knowledge-grounded dialogue generation, 2023.

[216] Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, and Wenjie Zhang. Paths-over-graph: Knowledge graph empowered large language model reasoning, 2025.

[217] Costas Mavromatis and George Karypis. Gnn-rag: Graph neural retrieval for large language model reasoning, 2024.

[218] Yuxin Yang, Haoyang Wu, Tao Wang, Jia Yang, Hao Ma, and Guojie Luo. Pseudo-knowledge graph: Meta-path guided retrieval and in-graph text for rag-equipped llm, 2025.

[219] Xinyu Wang, Yanzheng Xiang, Lin Gui, and Yulan He. Garlic: Llm-guided dynamic progress control with hierarchical weighted graph for long document qa, 2024.

[220] Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study. *JMIR AI*, 4:e58670, 2025.

[221] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

[222] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

[223] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[224] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[225] Guohao Li, Hasan Abed Al Kader Hammoud Li, Haotian Li, Chaowei Wu, Chen Zhu, Shiyu Liu, Khalid Almubarak, Zhenhailong Zhang, Zhao Ding, Zhilin Qian, et al. Camel: Communicative agents for "mind" exploration of large scale language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[226] Kun Qian, Yining Qian, Weize Xie, Guangyao Chen, Zhou Yu, Xifeng Wang, Qingxiu Xu, Ji Wang, and Chunhua Chen. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[227] Yanjun Dong, Jiale Jiang, Jie Yang, Qingkai Zeng, Tao Xie, and Huanjin Wu. Self-collaboration code generation via chatgpt. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 2024.

[228] Sirui Hong, Xiawu Wu, Shenlei Wang, Zejun Zhang, Guocheng Chen, Yaodong Wang, and Stan Z. Lu. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

[229] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

[230] Guangyao Chen, Kun Qian, Chunhua Chen, Chaoya Zhang, Ji Wang, Qingxiu Xu, Xifeng Wang, Fan Wu, and Zhou Yu. Autoagents: A framework for automatic agent generation. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[231] Weize Chen, Ziyi Li, Xuechen Liu, Yusen Zhang, Zhen Wang, Yadong Zhao, Xipeng Qiu, and Xuanjing Huang. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. *arXiv preprint arXiv:2308.10848*, 2023.

[232] Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15126–15154, 2024.

[233] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[234] Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. AdapterDrop: On the efficiency of adapters in transformers. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[235] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[236] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[237] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost, 2016. *arXiv preprint arXiv:1604.06174*, 2016.

[238] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[239] Noam Shazeer. Fast transformer decoding: One write-head is all you need, 2019. *URL https://arxiv.org/abs*, 1911.

[240] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[241] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[242] Marwa Abdulhai, Isadora White, Charlie Snell, Charles Sun, Joey Hong, Yuexiang Zhai, Kelvin Xu, and Sergey Levine. Lmrl gym: Benchmarks for multi-turn reinforcement learning with language models. *arXiv preprint arXiv:2311.18232*, 2023.

[243] Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*, 2025.

[244] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332, 2023.

[245] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[246] Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems. *arXiv preprint arXiv:2501.03468*, 2025.

[247] Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen, and Zhicheng Dou. CORAL: Benchmarking Multi-turn Conversational Retrieval-Augmentation Generation, October 2024.

[248] Tzu-Lin Kuo, Feng-Ting Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-Shan Shiu. RAD-Bench: Evaluating Large Language Models Capabilities in Retrieval Augmented Dialogues, February 2025.

[249] A. Waagmeester, G. Stupp, S. Burgstaller-Muehlbacher, B.M. Good, M. Griffith, O.L. Griffith, K. Hanspers, H. Hermjakob, T.S. Hudson, K. Hybiske, S.M. Keating, M. Manske, M. Mayers, D. Mietchen, E. Mitraka, A.R. Pico, T. Putman, A. Riutta, N. Queralt-Rosinach, L.M. Schriml, T. Shafee, D. Slenter, R. Stephan, K. Thornton, G. Tsueng, R. Tu, S. Ul-Hasan, E. Willighagen, C. Wu, and A.I. Su. Wikidata as a knowledge graph for the life sciences. *eLife*, 9:e52614, mar 2020.

[250] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004.

[251] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[252] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[253] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[254] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

[255] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

[256] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as Agents, October 2023. arXiv:2308.03688 [cs].

[257] Sen Yang, Yafu Li, Wai Lam, and Yu Cheng. Multi-llm collaborative search for complex problem solving. *arXiv preprint arXiv:2502.18873*, 2025.

[258] Mingyang Chen, Haoze Sun, Tianpeng Li, Fan Yang, Hao Liang, Keer Lu, Bin Cui, Wentao Zhang, Zenan Zhou, and Weipeng Chen. Facilitating multi-turn function calling for llms via compositional instruction tuning. *arXiv preprint arXiv:2410.12952*, 2024.

[259] Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

[260] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024.

[261] Mert Cemri, Melissa Z Pan, Shuyi Yang, Lakshya A Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, et al. Why do multi-agent llm systems fail? *arXiv preprint arXiv:2503.13657*, 2025.

[262] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, et al. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*, 2025.

[263] Haneul Yoo, Yongjin Yang, and Hwaran Lee. Code-switching red-teaming: Llm evaluation for safety and multilingual understanding. *arXiv preprint arXiv:2406.15481*, 2024.

[264] Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024.

[265] Andy Zhou. Siege: Autonomous multi-turn jailbreaking of large language models with tree search. *arXiv preprint arXiv:2503.10619*, 2025.

[266] Baokui Li, Sen Zhang, Wangshu Zhang, Yicheng Chen, Changlin Yang, Sen Hu, Teng Xu, Siye Liu, and Jiwei Li. S2m: Converting single-turn to multi-turn datasets for conversational question answering. In *ECAI 2023*, pages 1365–1372. IOS Press, 2023.

[267] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

[268] Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. How bad is training on synthetic data? a statistical analysis of language model collapse. *arXiv preprint arXiv:2404.05090*, 2024.

[269] Jie Chen, Yupeng Zhang, Bingning Wang, Wayne Xin Zhao, Ji-Rong Wen, and Weipeng Chen. Unveiling the flaws: exploring imperfections in synthetic data and mitigation strategies for large language models. *arXiv preprint arXiv:2406.12397*, 2024.

[270] Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I Abdin. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*, 2024.

[271] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[272] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

[273] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. All too human? mapping and mitigating the risk from anthropomorphic ai. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 13–26, 2024.

[274] Clare Beatty, Tanya Malik, Saha Meheli, and Chaitali Sinha. Evaluating the therapeutic alliance with a free-text cbt conversational agent (wysa): a mixed-methods study. *Frontiers in Digital Health*, 4:847991, 2022.

[275] Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranuta-porn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*, 2025.

[276] Claire Boine. Emotional attachment to ai companions and european law. 2023.