Synthesis and Evaluation of Long-term History-aware Medical Dialogue

Hebin Hu*
South-Central Minzu University
Wuhan, China
hebin9410@gmail.com

Renke Dai*
South-Central Minzu University
Wuhan, China
2024110293@mail.scuec.edu.cn

Ah-Hwee Tan†
Singapore Management University
Singapore, Singapore
ahtan@smu.edu.sg

Yilin Kang†
South-Central Minzu University
Wuhan, China
ylkang@mail.scuec.edu.cn

ABSTRACT

An effective healthcare agent must be able to recall and reason over a patient's longitudinal medical history. However, the absence of datasets with realistic long-term dialogue timelines limits systematic evaluation. Real clinical text is constrained by privacy and ethics, while existing benchmarks focus on isolated interactions, failing to capture cross-session reasoning. We introduce a framework for synthesizing high-quality, long-term medical dialogues with LLMs. Our approach entails a knowledge-guided decomposition into three stages: constructing synthetic patient profiles with diverse disease and complication trajectories, generating multi-turn dialogues per encounter, and integrating them into a coherent longitudinal history dataset, MediLongChat. We establish three benchmark tasks—In-dialogue Reasoning, Cross-dialogue Reasoning, and Synthesis Reasoning—to evaluate the memory capabilities of healthcare agents. To assess data quality, we introduce a multi-dimensional evaluation framework combining vector-based metrics with LLM-as-a-judge assessments. Specifically, we define automatic measures—Faithfulness, Coherence, and Diversity—together with two LLM-based evaluations: Correctness and Realism. Benchmark experiments show that even state-of-the-art LLMs struggle with MediLongChat. These findings highlight the benchmark's applicability and underscore the need for tailored methods to advance healthcare agents.

KEYWORDS

Healthcare agent, Synthetic Dataset, LLM, Medical Dialogue Dataset

ACM Reference Format:
Hebin Hu, Renke Dai, Ah-Hwee Tan, and Yilin Kang. 2026. Synthesis and Evaluation of Long-term History-aware Medical Dialogue. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/EFXQ8322

*The first two authors contributed equally to this work.
†Corresponding author

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/EFXQ8322

1 INTRODUCTION

As large language models (LLMs) have offered significant assistance in diverse areas in the medical domain [21, 25, 26], a central challenge lies in the development of healthcare agents that can engage in long-term, coherent conversations. A core requirement for such agents is the ability to interpret and utilize longitudinal patient history—not just the current utterance, but months or years of interactions concerning prior symptoms, diagnostics, and treatments. We refer to these settings as history-aware longitudinal clinical dialogues. For instance, as shown in Fig. 1, consider a patient who now reports persistent headaches and blurred vision. A history-agnostic agent might default to migraine. In contrast, a history-aware agent that recalls a prior breast-cancer diagnosis and incomplete follow-up imaging will surface altogether different diagnoses and safety actions. This example illustrates a significant point: Synthesize Longitudinal Reasoning-drawing clinically valid inferences from events scattered along a longitudinal trajectory—is fundamental to building safe and reliable healthcare agents.

Despite this importance, existing public benchmarks rarely stress longitudinality. Popular dialogue corpora tend to comprise independent conversations lacking a consistent patient narrative, while medical QA benchmarks emphasize static knowledge rather than dynamic, context-dependent reasoning [7, 16]. Moreover, assembling real longitudinal dialogue data is ethically and operationally daunting: clinical text is privacy-sensitive; de-identification is costly and imperfect; and even approved corpora frequently cover narrow settings and come with stringent governance. As a result, research on long-horizon clinical dialogue remains hampered by data scarcity.

Although the synthetic approach alleviates data scarcity and compliance barriers [4, 8], its quality is constrained by three common bottlenecks:

- Generation Quality and Consistency. The inherent hallucination problem [9, 15] of LLMs is particularly detrimental in the medical field. This issue is compounded by their tendency to produce contradictory details when generating long dialogues [12, 22]. Furthermore, the Mixture of Experts (MoE) architecture used in many LLMs may exhibit instability when generating long texts, leading to inconsistencies in style or knowledge depth.
- Context Window Limitations. A fundamental barrier to generating long, continuous dialogues is the finite context window

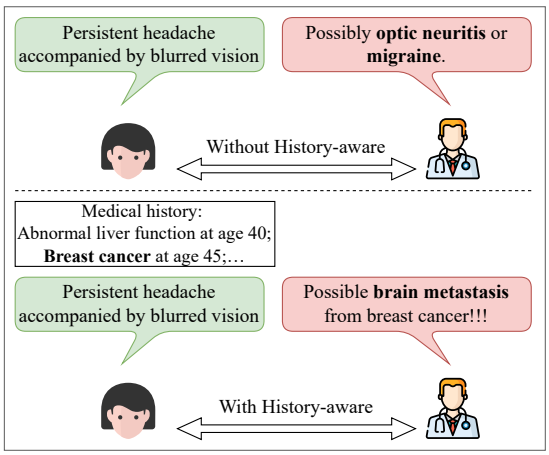

Figure 1: The importance of history-aware capabilities for healthcare agents. A history-agnostic model defaults to common causes for "headache + blurred vision," whereas a history-aware model recalls prior breast cancer, yielding different diagnoses and safety actions.

of LLMs. While nowadays models support longer contexts, utilizing their full capacity is often prohibitively expensive and still falls short of capturing a complete patient history.This exposes a deeper, unsolved problem: how to architect generation processes—from single, guided passes to complex pipelines—to guarantee narrative and logical coherence over the long term.

- Lack of Evaluation Standards. There is a lack of standardized methods for evaluating the quality of synthetic data itself. Current research often relies on limited automatic metrics or small-scale human evaluations, lacking a systematic and scalable framework to comprehensively measure a dataset's medical accuracy, long-term logical consistency, and effectiveness for evaluating agent capabilities.

To address the aforementioned challenges, this paper proposes a systematic pipeline for generating history-aware longitudinal clinical dialogue datasets and a multi-dimensional framework for their evaluation. The core of our method lies in task decomposition guided by structured knowledge. Specifically, we first construct metadata about disease cases and their complications. Based on this, we generate synthetic patient profiles with complete, diverse, and chronologically ordered medical events. Our task decomposition approach breaks down the complex process of generating a patient's lifelong medical history into manageable steps. By progressively creating multi-turn dialogues for each clinical visit, we build a long-term, coherent, and history-aware medical conversation dataset MediLongChat.

To systematically evaluate the memory and reasoning abilities of healthcare assistants, we introduce a benchmark built upon our dataset, featuring three dedicated tasks: In-dialogue Reasoning, Cross-dialogue Reasoning, and Synthesis Reasoning. These tasks respectively assess an agent's ability to recall information from a single encounter, link events across multiple dialogues, and synthesize the complete history for clinical inference. More importantly, we have designed a comprehensive framework that combines automatic metrics with an LLM-as-a-Judge approach to evaluate the quality of the generated dataset by measuring its Faithfulness, Coherence, Correctness, Diversity, and Realism. Our code and dataset are publicly available at: https://github.com/HebinHu/MediLongChat.

The main contributions of this paper are summarized as follows:

- We propose a novel framework for synthesizing long-form, history-aware medical dialogues with explicit longitudinal dependencies, based on knowledge-guided task decomposition, directly addressing challenges of LLM hallucination and inconsistency in long-content generation.
- We propose a comprehensive evaluation framework that establishes a new standard for assessing the quality of synthetic medical dialogue data.
- We construct a new benchmark dataset, MediLongChat, specifically designed to evaluate the longitudinal memory and reasoning capabilities of healthcare agents in multi-session dialogues.

2 RELATED WORK

2.1 Synthetic Medical Datasets

In the medical field, growing concerns over privacy, ethics, and data scarcity have led researchers to increasingly favor synthetic or semi-synthetic data for model training and evaluation. NoteChat [23] generates doctor-patient dialogues from clinical notes using a multi-agent framework, incorporating medical logic control to minimize invalid outputs. SynDial [2] leverages publicly available MTS-Dialogue and MIMIC datasets to generate dialogues via zero-shot prompting, integrating a feedback loop during generation to enhance dialogue quality. This method demonstrates superior performance in extractive and factual consistency compared to simple prompting approaches. SynSUM [18] bridges structured variables and clinical text for information extraction and causal studies, employing a Bayesian network to first generate tabular variables, which are then used to prompt an LLM to produce corresponding clinical text. Holysz et al. propose a multi-stage generation framework that first generates patient profiles and case backgrounds before producing dialogues, striving to improve the diversity and medical plausibility of synthetic data.

Although the aforementioned methods can generate high-quality single-session dialogues or dialogue-note paired samples, they still exhibit typical limitations in cross-turn and cross-dialogue reasoning. Most existing datasets focus on single consultation dialogues or intra-session dialogue-note alignment, with very few encompassing longitudinal records or dialogue histories of the same patient across multiple sessions. This presents a significant gap in evaluating whether healthcare agents possess long-term memory or the ability to comprehend longitudinal patient history.

2.2 Dataset Evaluation

Evaluating the quality of dialogue datasets and models is a critical and long-standing challenge in the field of natural language processing. Prior work can be broadly categorized into three main approaches: human-centric evaluation, traditional lexical and vector-based methods, and more recent LLM based evaluation.

Human Evaluation remains the gold standard for assessing dialogue quality [3]. For single-turn dialogues, metrics like mean opinion score or pairwise comparisons are widely used [3, 10]. However, evaluating multi-turn conversations presents unique challenges. Annotators must consider the entire dialogue history to assess long-term coherence, consistency, and the model's ability to maintain a persona or follow a complex narrative arc. While highly reliable, human evaluation is expensive, time-consuming, and difficult to scale, especially for large datasets with thousands or millions of dialogues. Moreover, achieving high inter-annotator agreement can be challenging due to the subjective nature of dialogue quality.

Traditional Lexical and Vector-Based Metrics were introduced to address the scalability issues of human evaluation. Early approaches focused on lexical overlap. Metrics such as BLEU [17], ROUGE [11], and METEOR [1] measure the n-gram similarity between a model's response and a set of human-written reference responses. While effective for tasks with a limited range of correct answers, these metrics are less suitable for multi-session dialogue, where there can be many valid and novel responses that do not match the reference text. To capture semantic similarity, later methods employed word embeddings and other vector-based techniques. Metrics like embedding average or greedy matching [20] calculate the cosine similarity between the embeddings of the generated response and the reference response. These metrics offer an improvement over lexical overlap by accounting for synonyms and semantically similar words. However, they still struggle to evaluate nuanced aspects of dialogue quality, such as factual consistency, long-term coherence across multiple turns, and the overall conversational flow of a lengthy dialogue.

Recently, the remarkable capabilities of LLMs have led to their use as automated evaluators, often referred to as "LLM-as-a-judge" [27]. The LLM is given a dialogue and a set of instructions, and it returns a score, a ranking, or a detailed critique. This method offers several advantages: it is significantly faster and cheaper than human evaluation, and it can be designed to assess more complex attributes like nuance, factual accuracy, and conversational flow. However, LLM-based evaluation is not without its limitations. The judgments can be susceptible to biases, such as position bias [27] or verbosity bias. While LLM-as-a-judge has shown high correlation with human judgments in many cases, it is still a developing field, and the reliability and robustness of these methods for complex, multi-turn conversations remain an active area of research.

3 MEDILONGCHAT SYNTHESIS PIPELINE

Our goal is to construct a high-quality, long-sequence medical dialogue dataset together with a benchmark that evaluates long-range memory and reasoning in healthcare agents. As illustrated in Fig. 2, the framework comprises three stages: (1)Knowledge-Guided Generation of Patient Medical Records; (2)Multi-turn Dialogue Generation Based on Task Decomposition; (3)Benchmark Generation.

3.1 Stage 1: Knowledge-Guided Generation of Patient Medical Records

A medical record contains complete information about a specific patient. Each medical record consists of four components: personal information, lifestyle habits, past medical history, and additional information. Medical records represent fictitious yet realistic lifelong patient profiles, thus avoiding privacy concerns. Notably, medical records serve as intermediate data during the generation process and are not included explicitly in the final dataset. Also, they remain hidden from the LLMs during evaluation benchmarks. High-quality datasets start from high-quality priors. To mitigate medically incorrect hallucinations, we generate complete fictional patient records under explicit knowledge guidance in three steps, ensuring both medical plausibility and narrative diversity.

3.1.1 Patient Persona Construction. For each fictional patient, we create detailed patient basic information, including demographics (e.g., age, sex, occupation), lifestyle habits (dietary preference, exercise frequency, smoking/alcohol history), and additional information, such as family history. We operationalize personas through prompts that explicitly state these attributes—for instance, "a sedentary young software engineer" versus "a retired teacher with balanced nutrition"—so that subsequent disease trajectories remain credible under varying contexts.

3.1.2 Disease-and-Complication Metadata Curation and Review. We compile metadata linking common diseases with their typical complications and temporal patterns. Because models may misestimate timing or likelihood, we incorporate a targeted human-in-the-loop review to verify: (i) evidence-based disease–complication associations; (ii) clinically plausible temporal ordering; and (iii) reasonable spacing between events across conditions. This step substantially strengthens the reliability of the underlying medical knowledge.

3.1.3 Generation of Sequential Medical Records. We fuse audited disease cases with patient personas to generate a coherent, clinically grounded timeline that enumerates all key medical events and their order. The timeline defines when patient–agent conversations should occur and encodes the longitudinal links among events across the patient's life course. This event timeline serves as the narrative backbone for subsequent dialogue generation.

3.2 Stage 2: Multi-turn Dialogue Generation Based on Task Decomposition

3.2.1 Definition and Scope of Chat History. Chat History constitutes the core component of our dataset, encapsulating all medical consultations related to diseases experienced by a patient throughout their lifetime. It consists of multiple dialogues, with each dialogue representing a consultation regarding a particular disease, simulating realistic patient-physician interactions. Each dialogue captures details such as date and location, greetings, symptom inquiry, routine examinations, and treatment discussions of the consultation. Typically, each dialogue contains around 50 conversational exchanges between the doctor and the patient, with an average length of approximately 3000 tokens. A complete patient chat history consists of 15-20 dialogues, implying the total token count of all consultations about 50K tokens.

Directly asking an LLM to produce a long-term, coherent history from the full record risks context loss, information mixing, and severe hallucinations. We therefore decompose "generate the patient's entire dialogue history" into simpler sub-tasks: generate

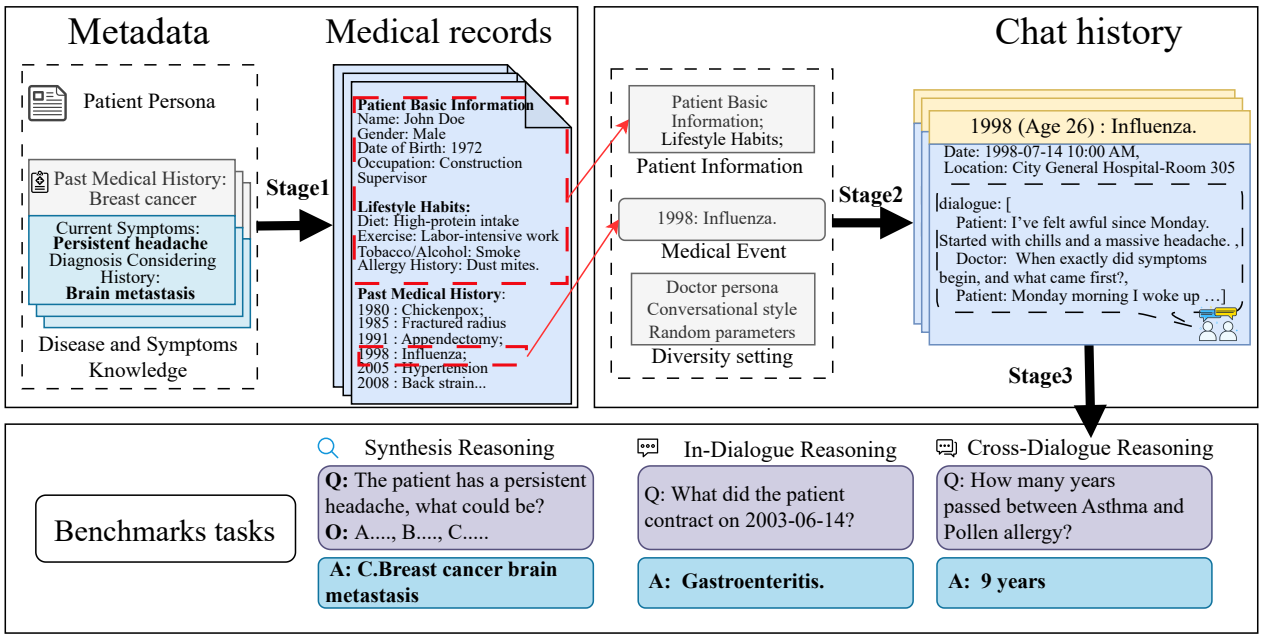

Figure 2: Overview of our dataset generation pipeline. Stage 1 builds records with knowledge guidance; Stage 2 generates per-event encounters and stitches them chronologically; Stage 3 derives three tasks to assess longitudinal memory and reasoning.

one high-quality encounter per medical event, then order and stitch them chronologically.

3.2.2 Synthesis Pipeline. We decompose synthesis into three steps. (1) Medical Event Extraction. From the Stage-1 timeline, extract each independent medical event, including its time, specific disease, and the treatment method administered at that time. (2) Context-isolated prompting. For each event, we compose a self-contained prompt that fixes the patient persona (defined in Step 1) and includes only event-local facts—condition name, occurrence time, and relevant interventions—thereby avoiding information leakage across visits. (3) Dialogue realization. Conditioned on this minimal yet sufficient context, the model generates a clinician–patient encounter that covers chief complaint, history taking, recommended examinations, provisional/definitive diagnosis, and management plan. This task decomposition sidesteps hard context-window limits and attenuates long-range hallucinations, and concatenating per-event encounters in temporal order yields a coherent, history-aware dialogue history.

To avoid stylistic monoculture, we introduce controlled variation during generation: (i) Physician persona diversity (e.g., "empathetic and reassuring physician," "concise and direct physician") sampled per encounter; (ii) Style control via prompt directives (from "focus on the patient's lifestyle" to "quick-paced consultation"); and (iii) Stochasticity using higher decoding temperatures to encourage lexical and structural variety.

3.3 Stage 3: Benchmark Generation

3.3.1 Motivation of benchmark. When designing the benchmark tasks, our primary objective was to systematically evaluate a model's ability to reason over longitudinal patient histories. Evaluations restricted to single-turn or independent dialogues cannot capture the central challenge of clinical interaction: integrating information across time and maintaining long-term dependencies. To address this limitation, we devise three complementary tasks—In-dialogue Reasoning, Cross-dialogue Reasoning, and Synthesis Reasoning—that progressively assess whether a model can accurately recall, connect, and leverage patient history over extended horizons.

This task design is motivated by two considerations. First, it explicitly decomposes different levels of difficulty: In-dialogue Reasoning probes whether the model can faithfully extract salient information within a single encounter; Cross-dialogue Reasoning examines whether it can track and relate events across multiple visits, capturing temporal and causal dependencies; and Synthesis Reasoning advances toward realistic clinical inference by requiring the model to integrate the entire history and derive new medical conclusions. Second, the framework enhances the interpretability of evaluation: by comparing performance across the three tasks, we can disentangle a model's strengths and weaknesses in short-term extraction, cross-session linkage, and longitudinal synthesis.

3.3.2 Reasoning task of benchmark. Taken together, this benchmark not only provides a quantifiable evaluation scheme but also ensures coverage of longitudinal patient history. In other words, these tasks collectively constitute a rigorous testbed for the fundamental capability that underpins safe and effective healthcare dialogue systems: long-term memory and cross-temporal reasoning.

In-Dialogue Reasoning. This in-dialogue reasoning (IDR) probes whether a model can accurately recover salient facts from a single encounter. Each item is constructed from one dialogue turn sequence and targets canonical clinical facets—such as visit date, location, presenting complaint, disease category, medications, and treatment plan—so that the answer can be grounded in explicit spans within the dialogue. Questions are phrased to require concise extraction or short abstractive summarization, thereby testing both fidelity and minimal reasoning over local context. The input to the model is the full text of one consultation; the expected output is a short textual answer. This task assesses the ability of LLMs to extract and succinctly summarize simple information from a single consultation dialogue.

Cross-Dialogue Reasoning. This cross-dialogue reasoning (CDR) evaluates whether a model can link information scattered across multiple visits, a prerequisite for longitudinal understanding. Each question is derived from two or more dialogues in the same Chat History and targets relationships that cannot be resolved from any single encounter alone, such as temporal ordering, duration between events, recurrence versus first onset, therapy changes, or the presence/absence of specific conditions. We include adversarial formulations (e.g., "Has the patient ever been diagnosed with … ?") and temporally anchored prompts (e.g., "duration between two illness episodes"). Since a complete Chat History can exceed an LLM's context window, CDR intentionally stresses memory mechanisms or long-context strategies rather than single-pass extraction.

Synthesis Reasoning. The synthesis reasoning (SR) task requires the LLM to diagnose a secondary disease or complication based on provided symptoms, given a complete medical chat history. The synthesis reasoning questions are constructed from metadata. This challenging task demands that the model not only recall the entire disease history of the patient but also accurately link current symptoms with past diseases. Strong performance on synthesis reasoning is essential for LLMs to effectively serve as medical assistants. Due to the complexity of this task, we designed it as a multiple-choice format, with interfering options selected based on symptom similarity from the Disease and Symptom dataset.

To clarify the dataset contents, the full MediLongChat corpus comprises longitudinal medical dialogues from 80 patients. For each patient, we retain (i) personal information; (ii) multiple complete dialogue transcripts spanning repeated visits; and (iii) annotations for three tasks—IDR, CDR, and SR.

4 EVALUATION

We adopt a unified, multi-faceted indicator to assess MediLongChat across five key dimensions—Faithfulness, Coherence, Diversity, Correctness, and Realism. These indicators are crucial for long, complex, and high-quality medical dialogues: faithfulness and correctness guard against harmful hallucinations; coherence ensures longitudinal consistency across sessions; diversity reflects topical breadth and style variability; realism captures human-likeness, empathy, and natural conversational dynamics. Below we provide formal definitions or concrete computation procedures for each dimension, so no duplicated explanations are required elsewhere.

Faithfulness (grounding to provided context/knowledge). Given input context s and generated dialogue $d = \{u_1, \ldots, u_n\}$, we compute utterance-level semantic similarity (cosine on sentence embeddings; e.g., SBERT) [19]:

$$\text{Faithfulness}(s, d) = \frac{1}{n} \sum_{i=1}^{n} \text{sim}(s, u_i), \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ denotes embedding-based semantic similarity.

Coherence (local flow and smoothness). To capture natural progression without abrupt shifts, we penalize sharp changes of adjacent-pair similarity:

$$\text{Coherence}(d) = 1 - \frac{1}{n-2} \sum_{i=1}^{n-2} \left| \text{sim}(u_{i+1}, u_{i+2}) - \text{sim}(u_i, u_{i+1}) \right|. \tag{2}$$

Diversity (corpus-level topical breadth and balance). We cluster the corpus $C = \{d_1, \ldots, d_m\}$ into K latent topics using BERTopic [5], and combine coverage and normalized entropy:

$$\text{Diversity}(C) = \frac{1}{2} \cdot \frac{K}{m} + \frac{1}{2} \cdot \frac{-\sum_{k=1}^{K} p_k \log p_k}{\log K}, \quad p_k = \frac{c_k}{m}. \tag{3}$$

Correctness (clinical factuality independent of source). Correctness targets clinical accuracy beyond mere overlap with s. Because purely lexical/vector metrics cannot verify medical soundness, we adopt an LLM-as-a-judge protocol based on G-Eval [13]. The judge is given the full dialogue history, model response, and a rubric to check medical claims, diagnoses, and advice. Scores are on a 5-point and linearly normalized to $[0, 1]$.

Realism (human-likeness, empathy, naturalness). We evaluate whether a conversation resembles human-to-human clinical communication in style, turn-taking and affect. We use an LLM-as-a-judge rubric (scores 1–5, normalized to $[0, 1]$).

Notes on computability and assessors. Not all dimensions are equally amenable to automatic computation: *Faithfulness/Coherence/Diversity* admit explicit, reproducible formulas (vector- or topic-based), while *Correctness/Realism* require semantic, pragmatic and domain judgments that are better captured by LLM-as-a-judge. We therefore report both automatic metrics (where applicable) and LLM-judged scores, and use a small human-annotated subset for sanity checks when necessary.

LLM based Evaluation. To address the limitations of traditional metrics and accurately assess the nuanced qualities of long, multi-turn dialogues, we employ an LLM-as-a-judge approach based on the G-Eval framework [13]. This method leverages the powerful generative and reasoning capabilities of a LLM to score dialogue turns and entire conversations against our predefined criteria. Consistent with prior work showing that strong LLM judges correlate well with human preferences [27], we adopt this paradigm for our evaluation. For each evaluation dimension— Coherence, Correctness, and Realism—we formulate a specific prompt that instructs the LLM to act as an expert evaluator.

The G-Eval process is conducted as follows: we provide the LLM with a dialogue history, the current model-generated response, a set of clear instructions, and a scoring rubric (typically a 5-point scale

Table 1: Summary of indicators, computation, and assessors. Vector/lexical metrics are deterministic; LLM-as-a-judge follows G-Eval with a 5-point Likert rubric normalized to $[0, 1]$.

Indicator	What it measures (why it matters)	Automatic (vector/topic) computation	LLM-as-a-judge (G-Eval)
Faithfulness	Grounding to provided context/KB; prevents unsafe hallucinations	$\frac{1}{n}\sum_i \mathrm{sim}(s, u_i)$ using sentence embeddings cosine; [19]	Cross-checks claims against s & history with criterion-based rubric
Coherence	Logical flow across turns/sessions; longitudinal consistency	$1 - \frac{1}{n-2}\sum_i \lvert\Delta\mathrm{sim}_{i\to i+1}\rvert$ (adjacent smoothness)	Rates discourse continuity and contradiction avoidance
Diversity	Topical breadth & balance; avoids repetitive patterns	BERTopic clusters (K); coverage + normalized Shannon entropy [5]	Optional: judge comments on semantic variety
Correctness	Clinical factuality independent of s; medical safety	— (no reliable purely lexical proxy)	Judge verifies clinical claims/diagnoses/advices; scores 1–5 $\to [0, 1]$
Realism	Human-likeness, empathy, natural turn-taking	— (no reliable purely lexical proxy)	Judge rates naturalness/empathy; optional discrimination test; scores 1–5 $\to [0, 1]$ [13]

). The prompt is carefully designed to guide the model's reasoning process, ensuring it considers the entire dialogue context. For instance, to evaluate Coherence, the prompt directs the LLM to assess the logical flow and consistency of the entire multi-turn conversation. Specifically, our prompts for each indicator are structured to include:

(1) A clear role assignment.
(2) The specific dialogue context, including the full history of turns.
(3) The response to be evaluated.
(4) A detailed definition of the evaluation metric.
(5) The scoring rubric with concrete examples for each score (e.g., 1 = hallucinates critical information, 5 = completely factually correct).

This method allows us to generate quantitative scores for each dialogue, which are then aggregated to provide a comprehensive dataset-level evaluation. We further compare these LLM-based scores against human annotations reported for the Conversation Chronicles dataset [6], observing strong alignment in trends and relative rankings, consistent with prior findings on LLM–human agreement [13, 27].

5 EXPERIMENT

5.1 Experiment Setting

Table 2: Comparison of selected long-context conversation datasets.

Dataset	Avg. turns per conv.	Avg. sessions per conv.	Avg. tokens per conv.	Domain
MSC	53.3	4	1,225.9	open
CC	58.5	5	1,054.7	open
LoCoMo	588.2	27.2	16,618.1	open
NoteChat	1	1	373.2	healthcare
ours	960.9	18.2	50217.3	healthcare

We primarily compare MediLongChat against prior long-context conversation corpora, including MSC [24], Conversation Chronicles (CC) [6], LoCoMo [14], and NoteChat [23]. In MediLongChat, a complete conversation is referred to as a dialogue, which corresponds to a session in the Locomo and CC datasets. Additionally, within each dialogue (or session), a single interaction (a question followed by a response) is referred to as a turn, a term that is used similarly in the other datasets. The average conversation in MediLongChat has 16 times more tokens than MSC (1,225.9 vs. 50,217.3), reflecting a significantly more extended discourse. It also spans 17 times more turns (960.9 vs. 58.5) and 3.6 times more sessions (18.2 vs. 5). This highlights MediLongChat as a notably more extensive and deeper dataset, particularly in healthcare, compared to others in the Table 2.

5.1.1 Overview. To comprehensively evaluate our dataset, we design three groups of experiments around the proposed data and evaluation framework: (1) Quality of synthetic dialogue generation; (2) Benchmarking our dataset across general-purpose LLMs; (3) Ablation of the generation pipeline.

5.1.2 Quality Evaluation of Synthetic Data. We compare four public long-form dialogue corpora under our metrics: LoCoMo, Conversation Chronicles (CC), Multi-Session Chat (MSC), and the medical dialogue dataset NoteChat. The first three corpora emphasize multi-turn or multi-session open-domain conversations and are not grounded in clinical settings, whereas NoteChat contains synthetic patient–physician encounters conditioned on clinical notes and thus resides in-domain. To ensure comparability, we apply a relaxed medical-factuality scoring policy to the first three datasets: we only assess self-consistency and commonsense plausibility, without penalizing based on specific medical knowledge.

We evaluate along five dimensions: Faithfulness, Coherence, Correctness, Diversity, and Realism. We conduct evaluations on both Stage 1 and Stage 2 of dataset generation, while Stage 3 serves as benchmark testing and is not subject to quality evaluation. Since Stage 1 concerns patients' medical information and does not involve dialogue fluency, we exclude Coherence from its evaluation. During assessment, we treat all patient information generated in Stage 1 as a single unit of generation for metric calculation, and likewise treat the entirety of dialogue content within a Stage 2 dialogue as a single unit of generation for evaluation.

5.2 Benchmarking on Our Dataset

To target long-term memory and cross-session reasoning, we construct three categories of tasks: (1) In-dialogue Reasoning (IDR); (2) Cross-dialogue Reasoning (CDR); (3) Synthesis Reasoning (SR). To reduce difficulty and scoring variance in synthesis reasoning,

we formulate SR task as multiple-choice questions (MCQs). On the MediLongChat benchmark, we evaluate performance using accuracy for the SR. For the IDR and CDR, we report both F1 score and BLEU-1 as evaluation metrics. On the LoCoMo benchmark, standard F1 and BLEU-1 are employed to evaluate the model's performance. We evaluate across a representative set of closed- and open-source LLMs: GPT-4o mini, DeepSeek-R1, Qwen3, and ERNIE-4.5.

Table 3: Evaluation results across stages. "/" indicates not applicable or not evaluated. s1 and s2 represent stage1 and stage2 in our dataset generation process

Dataset	Faithfulness	Coherence	Diversity
LoCoMo	/	0.904	0.4934
CC	/	0.932	0.4879
MSC	/	0.927	0.4874
NoteChat	/	0.9082	0.5243
Ours(s1)	0.635	/	0.4381
Ours(s2)	0.601	0.925	0.5447

5.3 Result of Quality Evaluation

Table 3 reports automatic metrics across datasets. On Coherence, our Stage-2 (0.925) is close to CC (0.932) and above LoCoMo (0.904) and NoteChat (0.9082). Faithfulness is comparable only for our data. We omit other corpora due to missing aligned sources. The gap is expected: Stage 1 (0.635), constrained by knowledge priors, yields slightly higher scores, whereas Stage 2 (0.601) shows a modest decline due to greater expressive freedom. For Diversity, Stage-2 (0.5447) is highest, followed by NoteChat (0.5243), with LoCoMo/CC slightly lower; Stage-1 (0.4381) is lower as expected. Overall, the pipeline improves diversity without sacrificing coherence, while maintaining faithfulness consistent with its structured priors.

To assess robustness against judge-specific bias, we report scores from multiple independent LLM judges and their ensemble (Table 4). While individual judges exhibit different preferences, the ensemble provides a more stable assessment. Table 5 presents the aggregated G-Eval scores for our dataset and the three baseline datasets. The scores are normalized to a 5-point scale, where a higher score indicates better performance in that dimension.

The evaluation results highlight the significant advantages of the MediLongChat dataset, particularly in dimensions critical for effective long-term dialogue and domain-specific applications:

- **Diversity:** MediLongChat achieves the highest score in Diversity, nearly reaching the maximum possible score. This is a direct consequence of our data collection and cleaning process, which was meticulously designed to capture a wide variety of linguistic expressions and conversational paths over multiple turns. This high diversity ensures that models trained on our data are less prone to generating repetitive or generic responses, a common drawback in multi-turn datasets.
- **Coherence:** Our dataset demonstrates exceptional long-term Coherence. This score validates our focus on maintaining consistency and logical flow lengthy dialogue sessions. In contrast, baseline datasets like msc and conversation chronicles report exhibit lower coherence, suggesting difficulty in preserving the narrative or context across many turns. This superior coherence makes MediLongChat an ideal resource for training models that require strong memory and contextual understanding over extended dialogues, such as tracking a patient's medical history.
- **Realism:** MediLongChat scores significantly higher on the Realistic metric compared to all baselines. This demonstrates that our dialogues closely mimic the naturalness, turn-taking dynamics, and authentic emotional tone of real-world interactions. This realistic quality is essential for training empathetic and user-friendly dialogue systems.
- **Correctness:** While LoCoMo shows a slightly higher score in Correctness, MediLongChat remains highly competitive with a score of **4.545**. This indicates that our data maintains a high standard of factual and domain-specific accuracy, which is non-negotiable for medical dialogue datasets. The minor difference suggests that the baselines may contain highly curated, single-turn factual statements, whereas our slightly lower score is likely due to the inherent complexity and higher risk of errors in long, multi-turn, generated dialogue.

Table 4: Multi-judge evaluation to mitigate potential LLM-as-a-judge bias. Ensemble denotes the aggregated score across heterogeneous judges.

Judge	Diversity	Coherence	Correctness	Realism
Gemini 2.5	4.99	4.58	4.65	4.90
GPT-5 mini	4.99	4.90	3.80	4.00
Qwen3-235B	4.47	4.88	4.95	4.92
Deepseek-R1	4.94	4.99	4.78	4.20
Ensemble	4.858	4.838	4.545	4.505

Table 5: G-Eval Scores of MediLongChat vs. Baseline Multi-Turn Dialogue Datasets (5-Point Scale)

Dataset	Diversity	Coherence	Correctness	Realism
LoCoMo	3.561	4.053	**4.965**	2.386
CC	3.174	3.504	4.860	2.062
MSC	3.694	3.482	4.696	1.982
NoteChat	3.138	4.464	3.213	2.340
ours	**4.858**	**4.838**	4.545	**4.505**

In summary, the G-Eval results confirm that the MediLongChat dataset successfully mitigates the common challenges of multi-turn dialogue datasets, particularly excelling in Diversity and Coherence, while maintaining high standards in Correctness and Realism.

5.4 Result of Benchmarking

We evaluate a set of general-purpose LLMs on the three MediLongChat tasks: in-dialogue reasoning (IDR; F1/BLEU), cross-dialogue reasoning (CDR; F1/BLEU), and synthesis reasoning (SR; accuracy). Across models, absolute scores remain modest for IDR and CDR, underscoring the difficulty of reasoning over long clinical histories, while SR shows comparatively higher accuracies (Table 6).

Table 6: Performance comparison of different models on the MediLongChat benchmark, including SR accuracy, IDR F1 and BLEU scores, and CDR F1 and BLEU scores.

Model	IDR		CDR		SR Accuracy
	F1	BLEU	F1	BLEU	
Deepseek-R1	33.49	11.12	20.36	2.41	80.00
Qwen3-235B	27.19	6.31	19.61	2.13	80.00
ERNIE-4.5-turbo	31.74	8.73	16.46	1.33	80.00
GPT-4o mini	23.30	3.28	24.25	2.46	80.00
GPT-4.1 mini	27.15	6.18	18.37	1.66	83.75

On IDR, Deepseek-R1 attains the highest F1 (33.49) and BLEU (11.12), indicating relatively stronger extraction and short-range reasoning within a single consultation. For CDR, which requires linking information across sessions, GPT-4o mini yields the best F1 (24.25), though BLEU remains low for all systems, reflecting the challenge of cross-session grounding in free-form answers. For SR, GPT-4.1 mini achieves the top accuracy (83.75), suggesting that multiple-choice formulations mitigate some of the generation variance observed in open-ended settings.

Taken together, these results paint a consistent picture: current LLMs can recover salient facts within a single encounter reasonably well, but performance drops when they need to link facts across sessions. The gap between IDR/CDR and SR also suggests that constrained answer formats partially alleviate long-context errors but do not solve underlying memory and reasoning limitations. We report scores without additional memory augmentation to provide a clean baseline for future methods.

Table 7: Ablation experiment results. Here, Fai, Coh, Cor, Div, and Rea denote Faithfulness, Coherence, Correctness, Diversity, and Realism, respectively. KG, TD, and DS represent Knowledge Guidance, Task Decomposition, and Diversity Setting in the generation process.

Stage	Method	Fai	Coh	Cor	Div	Rea
1	Ours	**0.6353**	/	**0.930**	**0.4381**	**0.720**
	w/o KG	0.4415	/	0.920	0.4313	0.540
2	Ours	**0.6010**	**0.924**	**0.909**	**0.5447**	**0.901**
	w/o TD	0.5809	0.8689	0.896	0.4590	0.720
	w/o DS	0.5784	0.921	0.904	0.3134	0.700

5.5 Ablation Experiment Results

To quantify the contribution of each component in our two-stage pipeline, we conduct ablations on Stage-1 and Stage-2 under the same evaluation protocol as the main experiments. Coherence is not reported for Stage-1 because its outputs are structured summaries rather than multi-turn conversations. We compare three deletions: removing knowledge guidance in Stage-1, and removing task decomposition or diversity controls in Stage-2. The results are summarized in Table 7.

For Stage-1, removing knowledge guidance leads to a marked deterioration in verifiability and narrative plausibility: Faithfulness drops from 0.6353 to 0.4415 and Realism from 0.720 to 0.540. These results directly support the knowledge-guided design: removing curated disease–complication metadata and temporal constraints lowers Faithfulness and Realism. Accordingly, retaining Stage-1 knowledge guidance is a justified and practical choice.

For Stage-2, both ablations underperform the full pipeline. Removing task decomposition reduces Coherence and Diversity, with smaller drops in Faithfulness, Correctness, and Realism. This indicates that decomposition helps preserve long-range logical flow and limits content mixing. Removing diversity controls shows a different profile: Diversity decreases markedly and Realism declines, while Coherence is largely preserved and Faithfulness changes only slightly. Overall, task decomposition primarily benefits coherence and clinical correctness, whereas diversity controls mainly improve linguistic variety and perceived naturalness with limited effect on faithfulness or coherence.

6 CONCLUSION

This work presents a knowledge-guided and task-decomposed framework for synthesizing history-aware longitudinal clinical dialogues together with a multi-dimensional evaluation protocol. By employing diverse patient personas and decomposed dialogue pipelines, we can generate coherent, traceable, and varied long-term conversations. To assess data quality and downstream model capabilities, we propose a multi-dimensional evaluation scheme combining automated metrics with LLM-based judging. This establishes rigorous standards across five dimensions: Faithfulness, Coherence, Correctness, Diversity, and Realism. Empirical results show that even state-of-the-art LLMs struggle on MediLongChat, particularly in long-term memory and cross-session reasoning. Ablation results show that knowledge guidance improves faithfulness and realism, diversity controls primarily increase linguistic variety with limited impact on factuality, and task decomposition helps maintain coherence and correctness in long-range dialogues.

Although our framework achieves a favorable balance between scalability and controllability, several limitations remain. First, synthetic data inevitably deviates from real clinical distributions, particularly in coverage of rare diseases, complex comorbidities, and behavioral health factors. Second, while LLM-as-a-judge offers efficient assessment, its reliability depends on prompt design and baseline model robustness, and its fine-grained agreement with clinical experts requires further improvement. Finally, our current dialogues primarily focus on the textual modality and have not yet systematically incorporated multi-modal signals such as medical images, lab curves, and structured EHR data.

Moving forward, we aim to extend the framework to multimodal and multilingual settings by integrating images, temporal physiological signals, and structured medical records. Furthermore, we will explore retrieval-augmented generation and dynamic episodic memory to enhance longitudinal reasoning and mitigate hallucinations in long-text generation. We hope that MediLongChat serves as a reusable, comparable public benchmark, providing a foundation and inspiration for research on longitudinal clinical reasoning and trustworthy medical dialogue agents.

ACKNOWLEDGMENTS

This work was partly supported by the General Project of the Central Universities of China (No. CZY23007), Hubei Province Key Research and Development Special Project of Science and Technology Innovation Plan (2023BAB087), Wuhan Key Research and Development Projects (2023010402010614), and Lee Kong Chian Professorship awarded to Ah-Hwee Tan by Singapore Management University.

REFERENCES

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72. https://aclanthology.org/W05-0908/

[2] Trisha Das, Dina Albassam, and Jimeng Sun. 2024. Synthetic Patient-Physician Dialogue Generation from Clinical Notes Using LLM. *CoRR* (2024).

[3] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54, 1 (2021), 755–810.

[4] Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. 2023. Synthetic data in health care: A narrative review. *PLOS Digital Health* 2, 1 (2023), e0000082.

[5] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[6] Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation Chronicles: Towards Diverse Temporal and Relational Dynamics in Multi-Session Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 13584–13606. https://doi.org/10.18653/v1/2023.emnlp-main.838

[7] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 2567–2577.

[8] Aleksandar Kovačević, Bojana Bašaragin, Nikola Milošević, and Goran Nenadić. 2024. De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artificial intelligence in medicine* 151 (2024), 102845.

[9] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6449–6464.

[10] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 1192–1202.

[11] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. 74–81. https://aclanthology.org/W04-1013/

[12] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics* 12 (2024), 157–173.

[13] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 2451–2470. https://doi.org/10.18653/v1/2023.emnlp-main.153

[14] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13851–13870.

[15] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 9004–9017.

[16] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*. PMLR, 248–260.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318. https://aclanthology.org/P02-1040/

[18] Paloma Rabaey, Henri Arno, Stefan Heytens, and Thomas Demeester. [n.d.]. SynSUM–Synthetic Benchmark with Structured and Unstructured Medical Records. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.

[19] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[20] Alexander M. Rush, Sumit Chopra, and Michael Collins. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 379–389. https://aclanthology.org/D15-1044/

[21] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[22] Ava Spataru. 2024. Know When To Stop: A Study of Semantic Drift in Text Generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 3656–3671.

[23] Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. NoteChat: A Dataset of Synthetic Patient-Physician Conversations Conditioned on Clinical Notes. In *Findings of the Association for Computational Linguistics ACL 2024*. 15183–15201.

[24] Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond Goldfish Memory: Long-Term Open-Domain Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5180–5197.

[25] Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7346–7353.

[26] Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*. 4442–4457.

[27] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.