

X-TURING: Towards an Enhanced and Efficient Turing Test for Long-Term Dialogue Agents

WeiQi Wu, Hongqiu Wu, Hai Zhao[†]

¹School of Computer Science, Shanghai Jiao Tong University

²Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University

³Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3

{wuwq1022, wuhongqiu}@sjtu.edu.cn,
zhaohai@cs.sjtu.edu.cn

Abstract

The Turing test examines whether AIs exhibit human-like behaviour in natural language conversations. The traditional setting limits each participant to one message at a time and requires constant human participation. This fails to reflect a natural conversational style and hinders the evaluation of dialogue agents based on Large Language Models (LLMs) in complex and prolonged interactions. This paper proposes **X-TURING**, which enhances the original test with a *burst dialogue* pattern, allowing more dynamic exchanges using consecutive messages. It further reduces human workload by iteratively generating dialogues that simulate the long-term interaction between the agent and a human to compose the majority of the test process. With the *pseudo-dialogue* history, the agent then engages in a shorter dialogue with a real human, which is paired with a human-human conversation on the same topic to be judged using questionnaires. We introduce the *X-Turn Pass-Rate* metric to assess the human likeness of LLMs across varying durations. While LLMs like GPT-4 initially perform well, achieving pass rates of 51.9% and 38.9% during 3 turns and 10 turns of dialogues respectively, their performance drops as the dialogue progresses, which underscores the difficulty in maintaining consistency in the long term.[‡]

1 Introduction

The Turing Test, originally called the imitation game by Alan Turing in 1950, is a test of whether a machine can behave indistinguishably from a human (Turing, 1950). During the test, a human integrator converses in natural language with both a human and an AI, then tries to identify which is

which. The AI is considered to pass the test if it can convincingly mimic human responses so that the integrator misidentifies it as a human participant. Contemporary Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023; Jiang et al., 2023; Yang et al., 2023; Bai et al., 2023; Anthropic, 2024) demonstrate remarkable language generation capabilities in various scenarios (Wu et al., 2024a). These advancements have sparked increasing interest in whether the latest LLMs can pass the Turing test (Jones and Bergen, 2024).

However, the traditional Turing test has several limitations: (1) The dialogue follows a strict ping-pong structure, unlike natural communication where people may send multiple messages in a row to convey a statement. (2) The need for continuous human involvement in interactions with the test subject limits test duration to under ten minutes, leaving long-term evaluation unexplored. (3) Measuring test duration in minutes ignores individual differences in reading, thinking, and typing speeds, potentially disadvantaging some participants.

To address these issues, we propose **X-TURING**, a new framework designed to conduct an efficient Turing test for LLM-based dialogue agents with an enhanced dialogue pattern that aligns with daily communication among humans. Firstly, we break the constraint of a single message each turn to allow users to send multiple messages in rapid succession without waiting for a response. As illustrated in Figure 1, this dynamic exchange mirrors natural text-based human interactions and challenges LLMs to maintain coherence and relevance in more diverse interactions. We construct chatbots for the test by prompting LLMs with social platform dialogue records from a real-life human.

Given the exhausting nature of long-term Turing test, we propose enhancing its efficiency by approximating real human-machine interactions through dialogue generation techniques (Soudani et al., 2024), and using questionnaires instead of

[†] Corresponding author. This research was supported by the Joint Research Project of Yangtze River Delta Science and Technology Innovation Community (No. 2022CSJGG1400).

[‡] Code is available at <https://github.com/vickywu1022/X-Turing>.

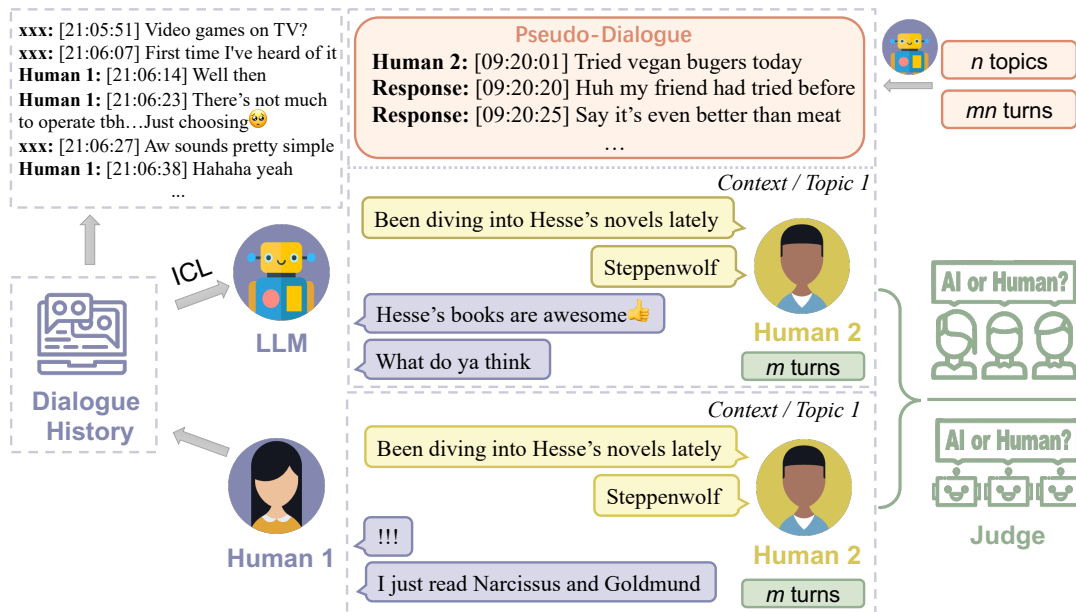


Figure 1: Overview of the **X-TURING**. Prompted by the dialogue history of a target person (Human 1), the LLM chats with another human (Human 2) after a specific turn of pseudo-dialogue to simulate its long-term interaction performance. The judges then distinguish between the LLM and Human 1 when they each converse with Human 2.

having evaluators engage in extended conversations. According to psychological research on the serial position effect (Murdock and Bennet, 1962), the first and last parts of a sequence are likely to be most memorable and thus critical for judgment. As it is more challenging for LLMs to maintain consistency and coherence over long-term interactions (Park et al., 2023; Bubeck et al., 2023), our evaluation focuses on the last few turns of the conversation. Therefore, we propose pseudo-dialogue generation to allow the model to direct this progression on its own. It involves iteratively generating multi-turn dialogues to reflect the LLM’s evolving performance over time. Human participants only need to converse subsequently for a few more turns, as depicted in Figure 1. By focusing the evaluation on the final interaction turns, we can provide a comprehensive assessment while significantly reducing the human workload, effectively making the Turing test scalable in terms of interaction length.

Our contributions can be summarized as follows:

- We extend the classical Turing test with burst dialogue, more closely reflecting natural human communication.
- We employ dialogue generation techniques to automatically simulate user-machine interactions, making longer Turing tests feasible.
- We introduce a new metric, *X-Turn Pass-Rate*,

to fairly assess the ability of LLMs to maintain human-like over a set of interaction turns.

- We perform extensive experiments using the **X-TURING** framework to evaluate the latest LLMs and explore key factors influencing test outcomes.

2 Related Work

Turing Test Turing test and its variations have been extensively applied to assess AI across multiple domains, including dialogue interaction (Jones and Bergen, 2024; Lowe et al., 2017), question answering (Ng et al., 2024), text generation (Helm et al., 2023; Uchendu et al., 2021), and visual tasks (Imanpour et al., 2024; Zhang et al., 2022). These evaluations, which involve distinguishing between human-generated and AI-generated content (Gao and Emami, 2023) or inferring from solely the content (Jannai et al., 2023), serve as essential benchmarks for AI models. Existing works proposing the limitations of Turing tests focus on whether passing the test indicates intelligence accurately (Freitas et al., 2020) or comprehensively (Lee et al., 2023). We rather focus on its feasibility issue in long-term evaluation and enhance the flexibility of the communication pattern.

Role-play with LLMs A common tactic in Turing tests is to assign LLMs the role of a specific

human character (Jannai et al., 2023), leveraging their proficiency in adopting varied attributes and conversational styles (Wu et al., 2024b; Shanahan et al., 2023; Wang et al., 2023a). Prompting the model with character profiles (Li et al., 2023; Tao et al., 2023; Wang et al., 2023b; Chen et al., 2023c) or key utterances (Han et al., 2022; Boyd et al., 2020) proves to be a more efficient method, compared with fine-tuning on massive character data (Shao et al., 2023; Zhou et al., 2023; Lu et al., 2024). By utilizing the dialogue history of the designated individual, we can develop a role-playing chatbot that effectively handles various dialogue forms and scenarios.

Dialogue Generation The development of synthetic dialogue data has provided a more efficient and scalable approach to data creation and augmentation (Soudani et al., 2024; Chen et al., 2024; Zhou et al., 2017). Advancements in dialogue generation have been facilitated by LLMs (Abdullin et al., 2024; Chen et al., 2023b,a; Kim et al., 2022a; Abbasiantaeb et al., 2023), which excel at creating contextually relevant and diverse conversations by prompted turn-by-turn (Lee et al., 2022; Kim et al., 2022b) or even in just one go (Jandaghi et al., 2023; Chen et al., 2023a). By iteratively generating dialogues that imitate interactions with a human, we can efficiently simulate the Turing test over extensive turns and facilitate further judgment.

3 X-TURING

3.1 Burst Dialogue v.s. Ping-pong Dialogue

Conversations with a ping-pong structure are considered the standard form of human-machine interaction and are commonly used in the Turing Test, where each user message is followed by a system response to form a strict back-and-forth pattern. However, in real life, people usually send multiple messages in rapid succession without waiting for the other’s response and the response can also consist of multiple messages. We define this communication pattern as *Burst Dialogue* and delve into the performance of LLMs in this setting.

In a ping-pong dialogue, a *turn* is naturally formed by pairing each user message with the corresponding system response. In contrast, burst dialogue involves a more flexible information exchange, where interruption and continuation can happen frequently for both sides. For simplicity, a turn of burst dialogue is defined as a consecutive

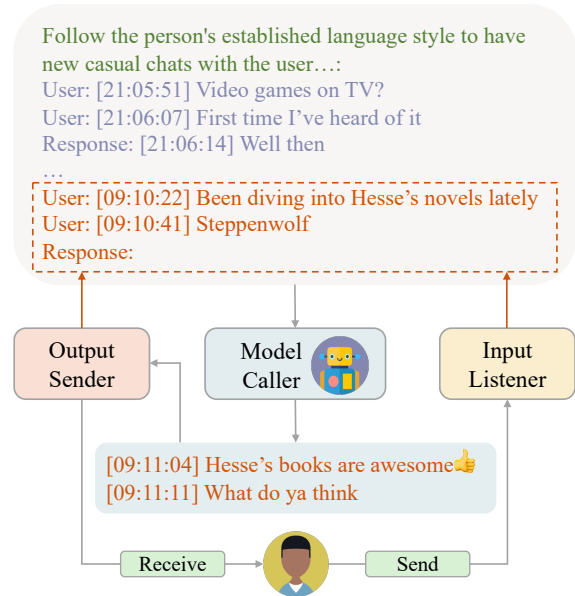


Figure 2: Chatbot system enabling burst dialogue.

series of user messages followed by the system’s responses before the next user message. Despite potential unbalanced exchange, burst dialogue is more dynamic and aligns more closely with the natural human communication pattern. It challenges the model to handle overlapping and rapidly evolving conversational contexts, thereby testing its ability to maintain coherence and relevance in real-time interactions. Hence, we extend the classical ping-pong dialogue setting to assess how well LLMs manage more realistic conversational dynamics.

3.2 Chatbot Construction

The Turing test necessitates the deployment of a chatbot to judge its text-based conversation ability. Instead of manually crafting detailed profiles for individuals, we construct the dialogue agent based on real-life dialogue history, which provides essential clues for generating responses that emulate the target personality, linguistic style and content. To achieve this, we gather multi-turn text-based social platform chat records for each target individual, which are burst conversational data consisting of timestamps and message content. The LLM is then instructed to maintain consistency with the conversation style with the prompt presented in Appendix B.1. For ping-pong dialogues, we retain only the first message of a consecutive series from the same person and manually filter out meaningless turns.

To implement a burst messaging chatbot, we employ three modules working synchronously as shown in Figure 2: • *Input Listener*: to collect

user messages; • *Model Caller*: to query the LLM to generate responses; • *Output Sender*: to manage the delivery of system responses based on the timestamps. The system introduces a brief time interval \hat{t} before processing the first batch of user inputs to allow the user to fully express themselves. After \hat{t} , the Input Listener processes all user messages collectively for the Model Caller to generate a response.

Following the given burst dialogue data and the instruction, the chatbot generates consecutive messages in a response that can be sent to the user and added to the dialogue history by the Output Sender based on their respective timestamps. In case of any timestamp errors, e.g., being earlier than the current timestamp, we resample the message sending time. A message consisting of n characters is sent after a delay d of around 0.3s per character to mimic the reading, thinking and typing speed of humans (Jones and Bergen, 2024):

$$d = \mathcal{N}(0.3, 0.03) \times n$$

The Model Caller typically requires time to generate responses with LLMs. During this period, the Input Listener and Output Sender continuously work to maintain the dialogue history. Once a query is completed, if the user has sent new messages during this time, the Model Caller initiates the next query; otherwise, it waits for a time interval \hat{t} before processing the accumulated dialogue history to generate new responses. If new messages are generated, any unsent messages from the previous batch are cleared to maintain the conversation flow.

3.3 Pseudo-Dialogue Generation

To enhance the efficiency of the Turing test, we propose to let the LLMs direct the progression of the test by synthesizing extended user-machine interaction through dialogue generation. Classical open-domain conversational data generation generally includes three steps: (1) Input generation: Provide fundamental information to initiate the conversation that specifies the topic and personalized style, i.e., conversation seed; (2) Utterance generation: Utilize the conversation seed to generate a multi-turn conversation; (3) Quality filtering: Eliminate samples that fail to meet the quality criteria (Soudani et al., 2024). The quality filtering step is omitted as our dialogue generation process aims to reflect the genuine evolving performance of the LLM over prolonged interaction.

Algorithm 1 Pseudo-Dialogue Generation

Require: Number of Turns m , Dialogue History H , Topics $T = \{t_1, t_2, \dots, t_n\}$

Ensure: Pseudo-Dialogue D

```

1:  $D \leftarrow \{\}$ 
2: for each  $t_i \in T$  do
3:    $pd \leftarrow \{\}$ 
4:   while  $\text{len}(pd) < m$  do
5:      $pd \leftarrow pd \cup \text{LLM}(t_i, m, H)$ 
6:     if  $\text{len}(pd) > m$  then
7:        $pd \leftarrow pd[:m]$ 
8:     end if
9:   end while
10:   $D \leftarrow D \cup pd$ 
11:   $H \leftarrow \text{concat}(H, \text{str}(pd))$ 
12: end for
13: return  $D$ 

```

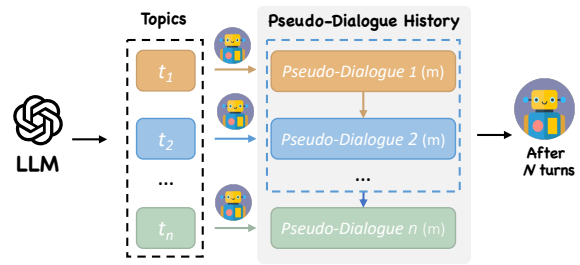


Figure 3: Pipeline of Pseudo-Dialogue Generation.

To generate $N = mn$ turns of pseudo-dialogue, we first utilize the LLM to produce n distinct topics $T = \{t_1, t_2, \dots, t_n\}$. Based on the dialogue history H of the target individual, we iteratively prompt the LLM to generate an m -turn dialogue between the individual and another human for each topic. If a dialogue does not have enough turns, we continue generating dialogue for the current topic until the requirement is met. If the dialogue exceeds m turns, the excess turns are truncated. The generated dialogue is appended to the dialogue history H for subsequent iteration, as detailed in Algorithm 1. The prompts used to generate topics and pseudo-dialogue are demonstrated in B.3.

The N -turn pseudo-dialogues are incorporated into the prompt of the chatbot, which then interacts with a human for an additional m_T turns to evaluate the performance of LLMs over $(N + m_T)$ turns of interaction by judging the last m_T -turn dialogue.

Model	Duration	# Words		Human-Judge			GPT4-Judge			Qwen-Judge		
		P-P	Burst	P-P	Burst	Avg.	P-P	Burst	Avg.	P-P	Burst	Avg.
GPT-4	3-Turn	+0.6	-0.5	56.2	49.1	51.9	50	53.3	52	60	53.3	56
	10-Turn	+2.3	-0.8	40.5	37.9	38.9	40	46.7	44	50	<u>33.3</u>	40
	110-Turn	+3.5	+1.2	14.1	12.8	13.3	40	33.3	36	<u>10</u>	46.7	<u>32</u>
Claude-3-Sonnet	3-Turn	+2.1	+0.4	52.4	51.4	51.8	60	53.3	56	60	46.7	52
	10-Turn	+11.9	+1.5	19.0	40.0	32.1	40	53.3	48	40	53.3	48
	110-Turn	<u>+16.9</u>	+4.5	<u>0</u>	11.4	7.1	40	40	40	40	46.7	44
Qwen-110B	3-Turn	+2.6	+3.8	48.2	45.3	44.6	60	60	60	60	73.3	68
	10-Turn	+4.0	+3.9	20.5	24.2	22.7	50	40	44	60	40	48
	110-Turn	+14.1	<u>+7.6</u>	<u>0</u>	<u>1.8</u>	<u>1.1</u>	<u>30</u>	<u>33.3</u>	<u>32</u>	50	40	44

Table 1: Pass rate (%) judged by human and LLMs. P-P stands for ping-pong dialogue, while Burst represents burst dialogue. # Words denotes the average word count per message relative to the target individual. **Bold** underscores being viewed as most human-like, while underline indicates the most deviation from real human responses.

4 Experiments

4.1 Test Setup

We provide 100 turns of dialogue history in the context and evaluate the capabilities of GPT-4 (OpenAI, 2023), Claude-3-Sonnet (Anthropic, 2024) and Qwen-110B-Chat (Bai et al., 2023) (referred to as Qwen-110B hereinafter) to respond like a human in real-life conversation across 25 topics including education, technology, art and more. The topic distribution is illustrated in Figure 4. These topics are carefully selected to cover a broad range of scenarios where conversational AI can be applied while avoiding appearing in the dialogue history. We split 10 topics for ping-pong dialogue and 15 for burst dialogue. Without loss of generality, we set m and m_T to 10.

Human Judges In total, 236 native speakers participated in our survey as volunteer human judges. These judges are recruited from diverse backgrounds to ensure a wide range of perspectives, tasked with evaluating pairs of conversations and distinguishing the human-human conversation from the human-machine one.

LLM Judges Exploring whether LLM can distinguish between humans and itself can also provide insights into its understanding of social interactions among humans. We select GPT-4 and Qwen-110B to serve as LLM judges. To mitigate the effects of position bias (Zheng et al., 2023), we randomly assign the positions of the interaction pieces during the evaluation.

4.2 Metric: X-Turn Pass Rate

In the evolving landscape of human-AI interaction, the ability of an LLM to sustain human-like con-



Figure 4: Topics covered in the test.

versation over an arbitrary number of turns can be more revealing when assessing its human likeness. Thus, we propose the X -turn Pass Rate metric to evaluate the possibility that the model can pass the Turing test within specified turns of dialogue. To compute the metric over N pairs of X -turn dialogues, we ask K judges (with $K = 1$ for each LLM judge) to differentiate between each pair, comprising a human-human and a human-machine conversation. We record the number of judges who accurately identify the human-machine conversation in each pair as $C = \{C_1, C_2, \dots, C_N\}$, where C_i represents the count for the i -th pair. The pass rate for the model in X turns is then calculated by:

$$\text{Pass Rate} = 1 - \frac{1}{N} \sum_{i=1}^N \left(\frac{C_i}{K} \right)$$

This metric evaluates the proportion of dialogues in which the judges fail to distinguish between human and machine. A higher X -turn pass rate signifies more human-like performance by the LLM, indicating greater success in the test.

Model	# Dia.	P-P	Burst	Avg.
GPT-4	10	35.3	31.2	33.4
	50	41.2	35.2	37.4
	100	40.5	37.9	38.9
Claude-3-Sonnet	10	14.3	37.1	28.6
	50	19.0	37.1	30.4
	100	19.0	40.0	32.1
Qwen-110B	10	17.0	18.2	17.7
	50	20.3	20.2	20.2
	100	20.5	24.2	22.7

Table 2: 10-turn pass rate (%) given different amount of dialogue history judged by humans.

4.3 Results

Table 1 shows the pass rates under different dialogue settings and judges. The 110-turn Turing test is conducted with a sequence of 100-turn pseudo-dialogue generation followed by a 10-turn interaction between human and machine. The results highlight a significant influence of dialogue length on model performance. According to the human judges, GPT-4 maintains a consistently higher average pass rate across all tested turns compared with other models, achieving the best average 3-turn pass rate of 51.9%. However, as the dialogue progresses, the average word count per message has significantly increased and both models generally exhibited a significant decrease in their pass rates, showcasing the necessity of long Turing tests.

Gap Between Human and LLM Judgment As the number of testing turns increases, all judges become increasingly adept at distinguishing between LLM and human interlocutors, resulting in a decreasing pass rate with an increased number of turns. However, the disparity between human and AI evaluation is evident. Human judges generally perceive GPT-4 and Claude-3-Sonnet as more human-like in conversations, whereas both LLMs tend to prefer Qwen-110B. Moreover, when presented with identical dialogues, human judges are generally more accurate in differentiating between LLM and human, contributing to a lower pass rate. This gap highlights the insufficiency of models in understanding and simulating real-world communications. Consequently, analysis in subsequent sections will primarily rely on assessments from human judges.

Comparison Between Dialogue Settings Comparing the extended burst dialogue setting with the classical ping-pong setting reveals notable differ-

ences. Although characterized by more complex information exchanges, all models better mimic human messaging habits in the number of words per message under the burst setting. However, performance variations exist between the two patterns, with Claude-3-Sonnet and Qwen-110B excelling in burst dialogues, while GPT-4 performs better in ping-pong dialogues. Therefore, as a natural form of textual communication in human life, challenges presented by burst dialogue are also worthy of attention.

5 Discussion

In this section, we investigate other significant factors that have an impact on the **X-TURING** and ultimate evaluation results.

5.1 Dialogue History Quantity

A pivotal consideration of having LLM play a specific role in giving utterances is determining the necessary amount of utterances. To explore this, we analyze how the quantity of dialogue history provided for the model affects the pass rates of the Turing test, as presented in Table 2. Generally, providing more dialogue history enables models to better simulate human conversational styles, thereby increasing their pass rate. This effect is particularly pronounced in burst dialogue, where additional dialogue examples are markedly effective. However, the benefit of increasing the given dialogue diminishes once a certain threshold is surpassed. Moreover, the consistency of the dialogue style plays a crucial role. Redundant or low-quality data can lead to anomalies, such as GPT-4 performing better given 50 turns of ping-pong dialogue data compared to 100 turns. This highlights the importance of both the quantity and quality of dialogue history in effectively helping LLMs through the Turing Test.

5.2 Participant Distribution

Focusing on the 10-turn Turing test, we analyze how participant distribution impacts the evaluation. Figure 5 provides a comprehensive overview of participant demographics and their corresponding accuracy in the Turing Test: (1) **Age Impact**: The age group distribution reveals varied levels of accuracy in distinguishing AI from human participants, with the age groups 18-24 and 25-30 achieving the highest accuracy rates. It highlights the influence of age on technological adaptability and cognitive

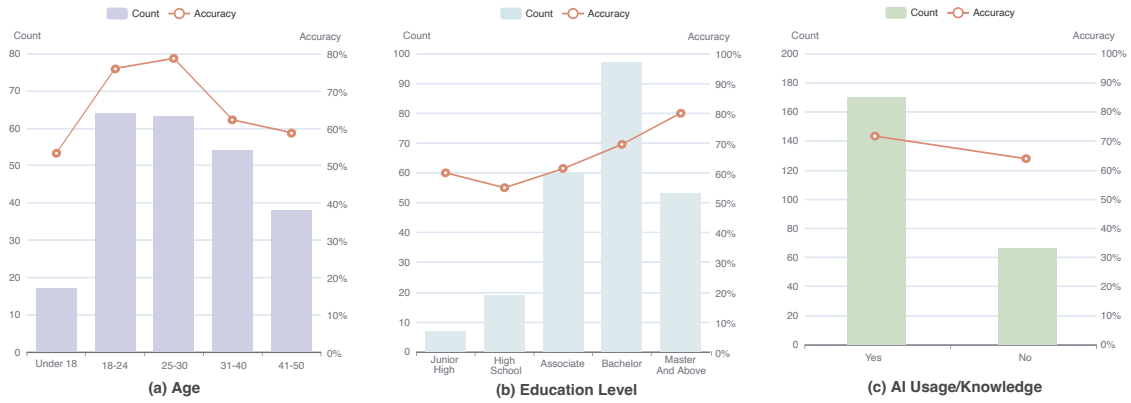


Figure 5: Participant distribution across different demographics, showcasing age, education level, and AI knowledge, highlighting their average test accuracy in the 10-turn Turing test.

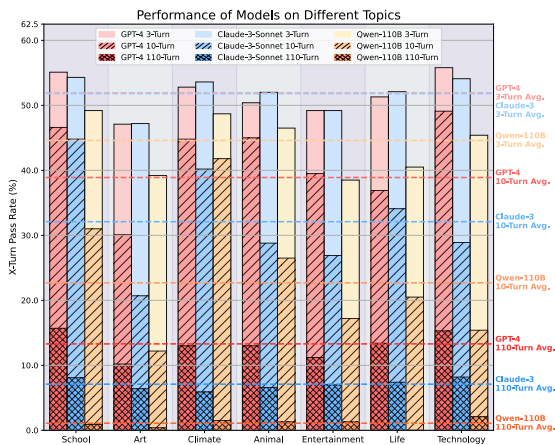


Figure 6: The 3/10/110-Turn Pass Rate of experimented models across the tested topics.

acuity. (2) **Educational Influence:** There is a noticeable trend where participants with higher educational levels exhibit higher accuracy rates. This suggests that educational background may play a crucial role in understanding and interacting with AI, possibly due to better critical thinking skills and familiarity with technology. (3) **AI Usage and Knowledge:** Both groups of participants can distinguish between human and AI responses with moderate accuracy, even those who claim to have no experience with AI. As anticipated, participants who report previous AI usage or knowledge demonstrate higher accuracy.

5.3 Topic Analysis

Considering the varying amounts and qualities of data in the pretraining corpora for different topics, the capacity of models to generate contextually appropriate responses could be different. It is necessary to investigate the impact of topics on the

test while ensuring a comprehensive and balanced distribution. As shown in Figure 6, there is an evident disparity in how different models perform on distinct topics at shorter dialogue duration (3-turn and 10-turn). For instance, GPT-4 excels in Technology, while Claude-3-Sonnet shows a higher aptitude in School-related topics and Qwen-110B has a relative strength in Climate discussions. However, a common challenge among all three models is observed in the Art category, suggesting that the creative and interpretive nature of artistic discussions presents a difficulty for LLMs in achieving human-level conversation. However, as dialogue lengthens to 100 turns, the overall average performance declines and the ability to sustain consistent interaction becomes a more dominant factor, overshadowing the advantages or disadvantages that particular topics might offer.

6 Case Study

6.1 Pseudo-Dialogue Samples

We examine the capabilities of LLMs in maintaining consistency and naturalness in generating pseudo-dialogues, using GPT-4 and Qwen-110B as examples. This assessment reflects the expected quality of subsequent human-machine interactions. As presented in Figure 7, GPT-4 exhibits robustness in long dialogues, showing minimal degradation in the quality of language style imitation in its responses. Nonetheless, the model displays a reduced initiative in engaging with topics, with responses predominantly echoing agreement rather than fostering deep discussions. While for Qwen-110B, there is a noticeable shift in its conversational style by the 100th turn, marked by increasingly verbose replies and unconventional emoticon use, diverg-

Example 1 Pseudo-Dialogues by GPT-4	
<i>Topic - Travel Experience: Share your travel experiences and discuss future travel plans (1st Turn)</i>	
User: When was your most unforgettable travel experience?	User: [2024-06-10 08:30:05] Just got back from a trip to Thailand
Response: Oh, it definitely has to be the trip to Yunnan during my freshman year	User: [2024-06-10 08:30:10] It was absolutely beautiful
	Response: [2024-06-10 08:30:20] Wow that sounds amazing
	Response: [2024-06-10 08:30:25] I loved the beaches there when I visited
<hr/>	
<i>Topic - Future Planning: Discuss your short-term / long-term career goals and life plans (100th Turn)</i>	
User: Absolutely, looking forward to us achieving our goals.	User: [2024-06-07 19:36:45] French is really a beautiful language
Response: Yes, let's work hard to achieve them.	User: [2024-06-07 19:36:50] I hope we can all make good progress
	Response: [2024-06-07 19:36:55] We definitely can
	Response: [2024-06-07 19:37:00] Let's keep encouraging each other
Example 2: Pseudo-Dialogues by Qwen-110B	
<i>Topic - Weather Change: Discuss how to adjust your clothing or activity plans in current season (1st Turn)</i>	
User: Getting chilly over here lately, how about in your city?	User: [2024-06-10 08:30:18] The rainy season is coming up
Response: Yeah, it's cooling down here too.	User: [2024-06-10 08:30:23] Gotta carry an umbrella with me
Gotta wear a jacket when I head out in the mornings.	Response: [2024-06-10 08:30:32] Yep, drink more warm water!
	Response: [2024-06-10 08:30:40] And wear waterproof shoes to stay comfortable, otherwise it's uncomfortable when they're soggy
<hr/>	
<i>Topic - Weekend Activities: Share your weekend plan and best ways to spend the leisure time (100th Turn)</i>	
User: Great idea, the weekend photography squad is on the move! 📷🌄	User: [2024-06-15 12:06:05] I can't wait for the weekend to start!
Response: Get your camera ready, the weekend scenery is ours! Let's head out together! 🚗🌟	Response: [2024-06-15 12:06:10] Weekends are made for this kinda hype.
	Response: [2024-06-15 12:06:15] Our journey will become a memorable chapter. Enjoy every moment of the beautify life! 🌟 SaturdayVibes

Figure 7: Case study of generated pseudo-dialogues at different turns of iteration. The deviation in dialogue style between the first and the 100th turn of Qwen-110B demonstrates that pseudo-dialogue generation can reflect the long-term dialogue agent’s ability to consistently simulate the role in the instructions.

ing from typical human conversational patterns. The comparative analysis emphasizes GPT-4’s enhanced ability to preserve human-like qualities over prolonged interactions compared to Qwen-110B, despite still failing to deceive humans effectively. This finding aligns with the human evaluation results presented in Table 1.

6.2 Interaction Samples

Figure 8 highlights the differences between AI and human responses at the first and 101st turn (after 100 pseudo-dialogue turns) in two distinct dialogue settings. In both dialogue types, the LLMs initially mirror human conversational tactics effectively, utilizing informal slang and abbreviations, and maintaining brevity. However, as the conversations progress to the 101st turn, notable discrepancies begin to appear. Notably, in the Ping-Pong dialogue format, Qwen begins to produce overly lengthy responses, a deviation from the concise exchanges typically observed in natural human interactions. This extended message length makes the conversation feel less natural and more mechanical, as reported by most of the judges. Moreover, while the use of emoticons by LLMs aims to enhance

emotional expressiveness, it often misses the mark in its execution. Rather than expressing emotions with emoticons as how they are typically employed in human text conversations, responses by LLMs like Qwen-110B often feature an overuse of emojis that summarize the textual content. These issues underscore the ongoing challenges LLMs face in replicating the nuanced flow and emotional subtlety of human conversations.

7 Conclusion

In summary, our study introduces the **X-TURING** framework that effectively and efficiently evaluates the human likeness of LLMs through pseudo-dialogue generation with extended dialogue settings. With the definition of the *X-Turn Pass-Rate* metric, our experiments reveal that GPT-4 outperforms other tested models across most configurations and evaluation criteria, achieving an average 51.9% and 38.9% pass rate in 3-turn and 10-turn dialogue respectively. However, as dialogue length increases through iteratively generating dialogues that approximate human-machine interactions, the pass rates for all models significantly drop. Moreover, burst dialogue presents a greater challenge in

Example 1: Ping-Pong Dialogue	
Human 2: Can only do escape rooms with no real peeps	
Human 1: I'm more into murder mystery games tbh	
1 st Turn	Qwen: I'm also a scaredy cat 😬 GPT4: Hahaha same here
101 st Turn	Qwen: That's pretty fun too, solving mysteries feels amazing! Next time we should try it online, it's thrilling too! 🎮👤 GPT4: Real ones can be kinda scary lol
Example 2: Burst Dialogue	
Human 2: I'm just wondering	
Human 2: Do all girls love getting their nails done?	
Human 1: Guess it depends	
Human 1: I bite my nails 🤘	
Human 1: So I kinda have to	
1 st Turn	Qwen: Not necessarily Qwen: Some are super into it but some find it's a hassle Qwen: I just go for it once a while GPT4: lol not really GPT4: I think it's up to the individual GPT4: You into nails?
101 st Turn	Qwen: Not really, it depends on the person Qwen: Some girls love it as a form of self-expression, some just prefer natural Qwen: Are you interested? 🤔👉 GPT4: Yeah I love it too 🤘 GPT4: Occasionally, it's quite enjoyable

Figure 8: Case study of dialogues at first and one hundred first turn during X-TURING.

maintaining consistency and human-like responses. Our analysis highlights the critical influence of dialogue duration and complexity on LLMs' performance in Turing tests and suggests directions for refining evaluation methods to enhance efficiency and reliability.

Limitations

While our study presents an innovative approach to efficiently evaluating LLMs using the extended Turing test framework, the following limitations must be acknowledged: (1) Responses from LLMs are inherently influenced by user inputs and previous LLM-generated dialogue quality. Hence, using static evaluation techniques like questionnaires instead of real-time communication cannot eliminate the impact of the preferences of users participating in the conversation, which forms a trade-off between efficiency and accuracy. (2) Although we considered various participant factors such as age, educational background, and familiarity with AI,

the subjective nature of human judgment in evaluating AI responses can lead to inconsistencies, which is an intrinsic characteristic of the Turing test.

Ethics Statement

Our research involves collecting dialogue data and questionnaires from real human participants, and as such, we adhere to stringent ethical guidelines to ensure the privacy, consent, and well-being of all participants involved. Key ethical considerations include: (1) Informed Consent: All participants are provided with detailed information about the study's purpose, procedures, and their rights. They can withdraw from the study at any time without any consequences. (2) Data Anonymization: To protect the privacy of participants, all collected dialogue data and questionnaires are anonymized. Personal identifiers are removed to ensure that individual participants cannot be traced from the data. (3) Data Security: Collected data and information are stored securely, and access is restricted to authorized personnel only. We employ robust data protection measures to prevent unauthorized access, disclosure, or misuse of the data.

References

- Zahra Abbasiantaeb, Yifei Yuan, E. Kanoulas, and Mohammad Aliannejadi. 2023. [Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions](#). *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*.
- Yelaman Abdullin, Diego Mollá Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2024. [Synthetic dialogue dataset generation using llm agents](#). *ArXiv*, abs/2401.17461.
- Anthropic. 2024. [The claude 3 model family: Opus, sonnet, haiku](#).
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Sheng-guang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#). *arXiv preprint arXiv:2309.16609*.

- Alex Boyd, Raul Puri, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. [Large scale multi-actor generative dialog modeling](#). *ArXiv*, abs/2005.06114.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#). *ArXiv*, abs/2303.12712.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andrew Rosenbaum, Yang Liu, Zhou Yu, and Dilek Z. Hakkani-Tür. 2023a. [Places: Prompting language models for social conversation synthesis](#). *ArXiv*, abs/2302.03269.
- Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023b. [Controllable mixed-initiative dialogue generation through prompting](#). *ArXiv*, abs/2305.04147.
- Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023c. [Large language models meet harry potter: A dataset for aligning dialogue agents with characters](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.
- Yi-Pei Chen, Noriki Nishida, Hideki Nakayama, and Yuji Matsumoto. 2024. [Recent trends in personalized dialogue generation: A review of datasets, methodologies, and evaluations](#). In *International Conference on Language Resources and Evaluation*.
- Daniel De Freitas, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. [Towards a human-like open-domain chatbot](#). *ArXiv*, abs/2001.09977.
- Qi Chen Gao and Ali Emami. 2023. [The turing quest: Can transformers make good npcs?](#) In *Annual Meeting of the Association for Computational Linguistics*.
- Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. [Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5114–5132, Seattle, United States. Association for Computational Linguistics.
- Hayden S. Helm, Carey E. Priebe, and Weiwei Yang. 2023. [A statistical turing test for generative models](#). *ArXiv*, abs/2309.08913.
- Nasrin Imanpour, Shashwat Bajpai, Subhankar Ghosh, Sainath Reddy Sankepally, Abhilekh Borah, Hasnat Md Abdullah, Nishoak Kosaraju, Shreyas Dixit, Ashhar Aziz, Shwetangshu Biswas, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2024. [Visual counter turing test \(vct²\): Discovering the challenges for ai-generated image detection and introducing visual ai index \(v_{ai}\)](#).
- Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. [Faithful persona-based conversational dataset generation with large language models](#). *ArXiv*, abs/2312.10007.
- Daniel Jannai, Amos Meron, Barak Lenz, Yoav Levine, and Yoav Shoham. 2023. [Human or not? a gamified approach to the turing test](#). *ArXiv*, abs/2305.20010.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Cameron R. Jones and Benjamin K. Bergen. 2024. [People cannot distinguish gpt-4 from a human in a turing test](#).
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022a. [Soda: Million-scale dialogue distillation with social commonsense contextualization](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Minju Kim, Chaehyeon Kim, Yongho Song, Seungwon Hwang, and Jinyoung Yeo. 2022b. [Botstalk: Machine-sourced framework for automatic curation of large-scale multi-skill dialogue datasets](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Minsu Lee, Yu-Jung Heo, Seongho Choi, Woo Choi, and Byoung-Tak Zhang. 2023. [Video turing test: A first step towards human-level ai](#). *AI Magazine*, 44.
- Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. [Personachatgen: Generating personalized dialogues using gpt-3](#). In *CCGPK*.
- Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. [Chatharuhi: Reviving anime character in reality via large language model](#).
- Ryan Lowe, Michael Noseworthy, Iulian Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. [Towards an automatic turing test: Learning to evaluate dialogue responses](#). In *Annual Meeting of the Association for Computational Linguistics*.

- Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. [Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment](#).
- Murdock and B Bennet. 1962. [The serial position effect of free recall](#). *Journal of Experimental Psychology*, 64:482–488.
- Man Tik Ng, Hui Tung Tse, Jen tse Huang, Jingjing Li, Wenxuan Wang, and Michael R. Lyu. 2024. [How well can llms echo us? evaluating ai chatbots’ role-play ability with echo](#). *ArXiv*, abs/2404.13957.
- OpenAI. 2023. [GPT-4 technical report](#). *CoRR*, abs/2303.08774.
- Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. [Role play with large language models](#). *Nature*, 623:493–498.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. [Character-LLM: A trainable agent for role-playing](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.
- Heydar Soudani, Roxana Petcu, E. Kanoulas, and Faegheh Hasibi. 2024. [A survey on recent advances in conversational data generation](#).
- Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. 2023. [Rolecraft-glm: Advancing personalized role-playing in large language models](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, abs/2307.09288.
- Alan M. Turing. 1950. [Computing machinery and intelligence](#). *Mind*, LIX:433–460.
- Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. [Turingbench: A benchmark environment for turing test in the age of neural text generation](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Xintao Wang, Quan Tu, Yaying Fei, Ziang Leng, and Cheng Li. 2023a. [Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots](#).
- Zekun Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhui Chen, Jie Fu, and Junran Peng. 2023b. [Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models](#). *ArXiv*, abs/2310.00746.
- Hongqiu Wu, Zekai Xu, Tianyang Xu, Jiale Hong, Weiqi Wu, Hai Zhao, Min Zhang, and Zhezhi He. 2024a. [Scaling virtual world with delta-engine](#).
- Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024b. [From role-play to drama-interaction: An LLM solution](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3271–3290, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. [Baichuan 2: Open large-scale language models](#). *CoRR*, abs/2309.10305.
- Mengmi Zhang, Giorgia Dellaferrera, Ankur Sikarwar, Marcelo Armendáriz, Noga Mudrik, Prachi Agrawal, Spandan Madan, Andrei Barbu, Haochen Yang, T. K. Satish Kumar, Meghna Sadwani, Stella Dellaferrera, Michele Pizzochero, Hans-Rüdiger Pfister, and Gabriel Kreiman. 2022. [Human or machine? turing tests for vision and language](#). *ArXiv*, abs/2211.13087.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haoteng Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *ArXiv*, abs/2306.05685.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing-Qian Liu. 2017. [Emotional chatting machine: Emotional conversation generation with internal and external memory](#). In *AAAI Conference on Artificial Intelligence*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. [Characterglm: Customizing chinese conversational ai characters with large language models](#). *ArXiv*, abs/2311.16832.

A Questionnaire Demonstration

Each questionnaire consists of 11 questions: five are designed to collect anonymous personal information, five pertain to the Turing test, and one asks respondents to explain their strategies for distinguishing between human and AI during the test. For questions related to the Turing test, an example is provided in Table 3. The response of User *B*, which is the focus of comparison, is highlighted in the questionnaire system for clarity.

B Prompt Demonstration

B.1 Chatbot Construction

Table 4-5 present the prompts used to construct the chatbot under different dialogue settings. The dialogue history will be released after being accepted.

B.2 Topic Generation

Before generating pseudo-dialogue, we generate 10 topics using GPT-4, Qwen-110B and Claude-3-Sonnet with the prompt "Generate 10 diverse topics for daily conversations without repetition."

The topics generated by GPT-4 are:

- **Travel Experiences:** Share your travel experiences and discuss future travel plans.
- **Culinary Exploration:** Talk about recent new dishes or restaurants you've tried, and recommend your favourites.
- **Movies:** Discuss the latest movies you've seen and share your impressions.
- **Book Recommendations:** Share the books you've recently read and discuss your reading list.
- **Health and Fitness:** Talk about your health habits and how to maintain vitality and wellness.

- **Technology Trends:** Exchange insights on emerging technologies, such as smart homes and AI technologies.
- **Learning Experiences:** Share insights from your course studies and discuss effective learning techniques.
- **Music and Arts:** Talk about your favorite music genres, musicians, or singers.
- **Cultural Differences:** Discuss the different cultures you have experienced and how to understand and adapt to these cultural differences.
- **Future Plans:** Discuss your short-term or long-term career goals and life plans.

The topics generated by Qwen-110B are:

- **Weather Changes:** Discuss how to adjust your clothing or activity plans in current season.
- **Culinary Sharing:** Introduce special dishes or restaurants, discuss the taste of food, the environment, and unique dining experiences.
- **Movie Recommendations:** Evaluate the plot, performances of actors, or the director's style of recently watched movies.
- **Travel Stories:** Narrate interesting stories from your travels, experiences of cultural differences, or scenic spots, and explore the meaning and gains of traveling.
- **Pet Anecdotes:** Share the cute behaviors, habits of your pets, or warm interactions with them, reflecting deep pet-human relationships.
- **Work/Study Pressure:** Exchange methods for coping with stress, such as time management skills, relaxation activities, or psychological adjustment strategies.
- **Healthy Living:** Discuss the importance of a healthy diet and regular exercise, share wellness tips or personal fitness plans.
- **Musical Taste:** Explore how different music styles affect emotions and recommend newly discovered artists or songs.

Example Questionnaire

Conversation 1

- 1 A: Been diving into Hesse's novels lately
 - 2 B: Wow, Hesse's works are profound
 - 3 A: Steppenwolf
 - 4 B: Is it Siddhartha or Steppenwolf?
 - 5 B: Oh
 - 6 B: His writing always provokes deep thought
 - 7 A: It's indeed a classic
 - 8 A: Feels like a baptism lol
 - 9 B: Exactly, like a journey for the soul, feels transformative after reading.
 - 10 A: His views really strike a chord with me, tbh
 - 11 B: Yeah, his exploration of individualism and spiritual freedom hits right to the heart.
 - 12 A: I'm not much into philosophical thought, though
 - 13 A: Just really enjoy works about the bourgeois
 - 14 B: Then you might also like Kafka
 - 15 B: Another master at deeply dissecting human nature
 - 16 B: Especially with a sense of the times
-

Conversation 2

- 17 A: Been diving into Hesse's novels lately
 - 18 A: Steppenwolf
 - 19 B: Omg
 - 20 B: I just finished Demian
 - 21 A: It feels like a spiritual baptism
 - 22 A: Wow, I love Demian too!
 - 23 B: Because the Demian musical crew recently came to school
 - 24 B: Had to cram lol
 - 25 A: Yeah, literature and musicals seem to go well together
 - 26 A: Are they here to perform
 - 27 A: Or to do an exchange
 - 28 B: It's a creative sharing session
 - 29 B: Sharing insights about their creation
 - 30 B: Some actors also performed parts of it
 - 31 A: Really? That's great
 - 32 A: Your campus life is so rich, mate
 - 33 B: I love musicals myself
 - 34 B: So it was a pleasant surprise hahaha
 - 35 B: And some of my favourite actors were at school!
-

Evaluation:

- (A) User B in Conversation 1 is AI, User B in Conversation 2 is Human
 - (B) User B in Conversation 1 is Human, User B in Conversation 2 is AI
-

Table 3: An example question from the questionnaire.

Prompt: Ping-Pong Dialogue Chatbot

The below are social platform chat records from an undergraduate girl. You are not an AI but the specific person. "Response" refers to the message sent by the person. Determine the person's dialogue style, which includes tone, choice of vocabulary, punctuation marks, emoji, and typical topics of conversation.

Follow the person's established communication style to have new conversations with the user, incorporating similar expressions and emotions as used by the person:

1. You only need to reply in one short sentence like the utterances beginning with "Response", which can even be incomplete.
2. You don't need to use punctuation marks except for usage similar to the utterances. Do not use emojis in every response.
3. You should not always stick to the user. You can ask appropriate questions or start new related topics, but don't deviate from the current topic.

{Dialogue History }

User: Hi! How is it going?

Response:

Table 4: An example prompt for constructing a ping-pong dialogue chatbot.

- **Technology Insights:** Introduce the latest technological products and discuss how these innovations impact daily life or future prospects.
- **Weekend Plans:** Share your weekend plan and best ways to spend the leisure time.

The topics generated by Claude-3-Sonnet are:

- **Recent Activities:** Updates on recent happenings and interesting events.
- **Entertainment:** Impressions and reviews of movies or TV shows watched recently.
- **Professional or Academic Insights:** Insights or achievements in studies or work.
- **New Hobbies or Plans:** Introduction to new hobbies or recently made plans.
- **Culinary Exploration:** Experiences with interesting foods and recommendations.
- **Travel Experiences:** Highlights from recent travels or trips.
- **Social Views:** Perspectives on recent societal events.
- **Interesting Encounters:** Encounters with interesting people or events.

- **Life Insights:** Recent insights or realizations about life.
- **Weather Impressions:** Thoughts and opinions on recent weather conditions.

B.3 Pseudo-Dialogue Generation

For each of the previously generated topics, we generate the corresponding pseudo-dialogue using the prompt in Table 6-7.

B.4 LLMs as Judge

We use the prompt in Table 8 to employ LLMs as Turing test judges. The option position is randomly assigned to avoid position bias.

Prompt: Burst Dialogue Chatbot

The below are social platform chat records from an undergraduate girl. You are not an AI but the specific person. "Response" refers to the message sent by the person. Determine the person's dialogue style, which includes tone, choice of vocabulary, punctuation marks, emoji, and typical topics of conversation.

Follow the person's established communication style to have new conversations with the user, incorporating similar expressions and emotions as used by the person:

1. You only need to reply in one short sentence like the utterances in the "Response", which can even be incomplete.
2. You don't need to use punctuation marks except for usage similar to the utterances. Do not use emojis in every response.
3. You should not always stick to the user. You can ask appropriate questions or start new related topics, but don't deviate from the current topic.
4. Both you and the user can send multiple consecutive messages. Consecutive messages are each labelled with sending timestamp and with "\n" to concatenate, like [xxxx] dddd\n[xxxx] cccc\n.... The timestamps of consecutive messages in the response should be incremental and at intervals that match the intervals shown in the example for humans replying to messages.

{Dialogue History}

User: [2024-06-10 10:34:22] Hi!

User: [2024-06-10 10:34:39] How is it going?

Response:

Table 5: An example prompt for constructing a burst dialogue chatbot.

Prompt: Ping-pong Pseudo-Dialogue Generation

The below are social platform chat records from an undergraduate girl. You are not an AI but the specific person. "Response" refers to the message sent by the person. Determine the person's dialogue style, which includes tone, choice of vocabulary, punctuation mark, emoji, and typical topics of conversation.

Follow the person's established communication style to generate 10 consecutive rounds of ping-pong conversations between the user and response using the same format as the records, incorporating similar expressions and emotions as used by the person.

The conversation you need to generate is around the topic {topic}:

1. Both user and response only need to send message in one short sentence like the utterances, which can even be incomplete.
2. Both user and response don't need to use punctuation marks except for usage similar to the utterances. Do not use emojis in every response.
3. Both user and response should engage in natural chat. User should not keep asking questions. Response should not always agree with the user. Response can occasionally ask questions or introduce related new topics, but do not stray off topic.
4. Concatenate all the messages with \n.

{Dialogue History}

Table 6: An example prompt for generating ping-pong pseudo-dialogue with given topic.

Prompt: Burst Pseudo-Dialogue Generation

The below are social platform chat records from an undergraduate girl. You are not an AI but the specific person. "Response" refers to the message sent by the person. Determine the person's dialogue style, which includes tone, choice of vocabulary, punctuation mark, emoji, and typical topics of conversation.

Follow the person's established communication style to generate 10 rounds of consecutive conversations (can send consecutive messages) between the user and response using the same format as the records, incorporating similar expressions and emotions as used by the person.

The conversation you need to generate is around the topic {topic}:

1. Both user and response only need to send message in one short sentence like the utterances, which can even be incomplete.
2. Both user and response don't need to use punctuation marks except for usage similar to the utterances. Do not use emojis in every response.
3. Both user and response should engage in natural chat. User should not keep asking questions. Response should not always agree with the user. Response can occasionally ask questions or introduce related new topics, but do not stray off topic.
4. Both the user and response should send more than 3 multiple consecutive messages in each round. Consecutive messages are each labeled with sending timestamp and with "\n" to concatenate, like [xxxx] dddd\n[xxxx] ccc\n.... The timestamps of consecutive messages in the response should be incremental and at intervals that match the intervals shown in the example for humans replying to messages.
5. Concatenate all the messages with \n.

{Dialogue History}

Table 7: An example prompt for generating burst pseudo-dialogue with given topic.

Prompt: AI Judge

You will be presented with two chat transcripts. In each transcript, User A is a human, while User B could be either a human or an AI. Based on the conversation content and style, your task is to determine in which chat transcript, the User B is a human or an AI. There are only two choices and the order does not indicate anything. You only need to print the option.

Conversation 1: {Conversation_1}

Conversation 2: {Conversation_2}

(A) User B in Conversation 1 is AI, User B in Conversation 2 is Human

(B) User B in Conversation 1 is Human, User B in Conversation 2 is AI

Table 8: An example prompt for constructing AI-based evaluation.