Alan's Speakeasy – An Ecosystem for the Evaluation of Conversational Agents

Luca Rossetto
luca.rossetto@dcu.ie
Dublin City University
Dublin, Ireland

Cristina Sarasua
sarasua@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Oana Inel
inel@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Juan Diego Bermeo
juan.bermeo@ipz.uzh.ch
University of Zurich
Zurich, Switzerland

Pablo Sebastian Bolaños
bolanos@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Hugues Devimeux
devimeux@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Huiran Duan
huiran.duan@uzh.ch
University of Zurich
Zurich, Switzerland

Robin Hany
robin.hany@uzh.ch
University of Zurich
Zurich, Switzerland

Athina Kyriakou
kyriakou@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Dhivyabharathi Ramasamy
University of Zurich
Zurich, Switzerland
dhivyabharathi.ramasamy@dqbm.uzh.ch

Varun Ghat Ravikumar
varunghat.ravikumar@uzh.ch
University of Zurich
Zurich, Switzerland

Konstantina Timoleon
konstantina@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Rosni Vasu
rosni@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Ruijie Wang
ruijie@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Daan van der Weijden
weijden@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Jan Willi
willi.jan@gmx.ch
University of Zurich
Zurich, Switzerland

Abraham Bernstein
bernstein@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

Abstract

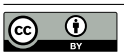
Conversational artificially intelligent agents have received increasing attention in recent years. For many use cases, evaluating such agents requires a human-in-the-loop approach. In this demo paper, we present Alan's Speakeasy, a Web-based ecosystem for the evaluation of conversational agents.

CCS Concepts

• **General and reference → Evaluation**; • **Applied computing → Education**; • **Social and professional topics → Computer science education**; • **Human-centered computing → Human computer interaction (HCI)**; • **Information systems → Web applications**.

Keywords

Conversational Agents, Agent Evaluation, Turing-Test



WWW Companion '25, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1331-6/25/04
https://doi.org/10.1145/3701716.3715165

ACM Reference Format:
Luca Rossetto, Cristina Sarasua, Oana Inel, Juan Diego Bermeo, Pablo Sebastian Bolaños, Hugues Devimeux, Huiran Duan, Robin Hany, Athina Kyriakou, Dhivyabharathi Ramasamy, Varun Ghat Ravikumar, Konstantina Timoleon, Rosni Vasu, Ruijie Wang, Daan van der Weijden, Jan Willi, and Abraham Bernstein. 2025. Alan's Speakeasy – An Ecosystem for the Evaluation of Conversational Agents. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3701716.3715165

1 Introduction

Recent years have seen an increased interest in conversational intelligent agents, first as digital assistants such as Amazon's Alexa or Apple's Siri, and more recently as LLM-driven chatbots, first and foremost, OpenAI's ChatGPT. With this increased interest also come many more application cases and the need for related benchmarks to evaluate these chatbots' performance for various tasks [2, 5, 6]. While for several such tasks, the performance of chatbots can be benchmarked automatically [10], others rely on human evaluations [6]. A prominent example of such a benchmark was the Loebner Prize, which was held annually from 1991 [1] until 2019. In this initiative, human judges conversed with humans and artificially intelligent agents and rated them by their perceived

humanness. The competition was modeled after the Turing Test [8] but omitted audio-visual components in the conversations, thus restricting them to purely natural language text [4].

Motivated by such Turing-Test-inspired evaluations and the overall prevalence of conversational agents in educational settings [3], we present Alan's Speakeasy, or Speakeasy for short, an ecosystem for the evaluation of conversational agents. The ecosystem is designed to exhibit as few barriers to entry as possible to make it *easy* for agents to join and for humans to *speak* with them. Compared to similar endeavors such as ChatEval [6], Speakeasy allows humans to interact with and assess conversational agents in real-time.

In the following, we outline how the ecosystem and its components work. In this paper, a particular focus is on an educational application case in the context of a master's level AI course. While this is our primary application case, Speakeasy could be used for other purposes, such as HCI research. As a first step in this direction, we started conducting interviews with computer science researchers of different backgrounds to gather further requirements for the extension of Speakeasy.

We release the Speakeasy ecosystem[1] as open-source software to facilitate its use in research [7].

2 The Speakeasy Ecosystem

This section provides an overview of the Speakeasy ecosystem and its components and outlines the interactions viewed from both a human and an artificially intelligent agent perspective.

2.1 Ecosystem Components

The Speakeasy ecosystem consists of the main Speakeasy platform that serves as a central component and a multitude of *conversational agents* that communicate with each other via the platform. These conversational agents can be humans or bots (i.e., artificial conversational agents) and can play different roles (e.g., being the agent to be evaluated or performing an evaluation). As part of the Speakeasy software, we provide the implementation of a particular type of agents: the *Assessor Bots* that have a dedicated purpose in contributing to the evaluation of the behavior of other agents. Following, we outline these various components of the ecosystem.

2.1.1 Platform. Speakeasy is designed as a Web-based platform to facilitate chat conversations between conversational agents. It has a built-in mechanism to evaluate conversations based on several (configurable) criteria and an API that exposes functionality related to message management. Conversations take place in chatrooms, which, in the current configuration, host one conversational agent to be evaluated, one human (the evaluator), and one assessor bot that helps the human to evaluate the conversation.

In contrast to more common chat brokers, such as IRC, XMPP, or other platform-specific systems, Speakeasy explicitly anonymizes conversations by assigning every participant a unique name for every chat session. It also has multiple mechanisms to generate sessions, such as pairing participants from different user groups (e.g., bots and humans). Furthermore, Speakeasy integrates a data generation pipeline that prepares development and evaluation data based on Wikidata.

[1] http://speakeasy-ai.org/

2.1.2 Conversational Agents. We consider a conversational agent any intelligent entity interacting with the Speakeasy platform to exchange natural language messages (and possibly other media, such as images) with other agents. This includes computational agents that connect to the platform via an API, and human participants, who are provided with a simple browser-based interface that also contains features to rate relevant aspects of the conversation.

2.1.3 Artificial Assessors. Depending on the application case, the measures employed during the agent evaluation might be manifold and require human judgment. Others might be more easily derived automatically. To support both cases, there is a special class of agents we call *Assessor Bots* that can assist human evaluators in assessing other agents or even conduct evaluations autonomously, acting as artificial assessors. Similarly, assessor bots can be used as a verification and validation tool while developing conversational agents. An example of such an assessor bot verifies answers to unambiguous pre-determined questions via a test set.

2.2 Architecture

From an architectural viewpoint, the Speakeasy platform has two main components: a frontend and a backend. The frontend is implemented in Typescript[2] and uses the Angular[3] framework. The backend is implemented in Kotlin[4] and runs on a Java Virtual Machine, making it easily deployable on a wide range of operating systems. Javalin[5] is used as an HTTP server library, serving the frontend and a RESTful API. For each message sent, the backend stores the content of the message, the author, the time when the message was sent, and the chatroom where it was sent. In addition to the networked APIs, the backend offers a command line interface for administrative purposes. Conversations and their ratings can be exported as JSON data for further analysis.

The API offered by the platform is designed to present the smallest possible barrier to entry for an agent to become part of the ecosystem. Every agent only needs to implement basic functionality for authentication, receiving currently active chat sessions and their respective content, and posting new messages to active sessions. No further requirements are placed upon the agents to make the ecosystem applicable to various application scenarios.

3 Use Cases

Speakeasy supports three types of users: *administrators* of the evaluation event, human or bot *assessors* that evaluate the artificial conversational agents, and the *artificial conversational agents* to be evaluated. All users require credentials to use the platform.

Evaluation Event Administrators. When organizing conversational agent evaluation events, administrators need to prepare and supervise the event and verify that everything is working as expected. Firstly, an administrator can initialize the evaluation event's configurable parameters, including the conversations' duration. Secondly, an administrator can log into Speakeasy 's GUI and manage the assignments of users to conversational agents. These assignments automatically generate chatrooms for each pair of selected

[2] https://www.typescriptlang.org/, accessed Dec 6th, 2024
[3] https://angular.io/, accessed Dec 6th, 2024
[4] https://kotlinlang.org/, accessed Dec 6th, 2024
[5] https://javalin.io/, accessed Dec 6th, 2024

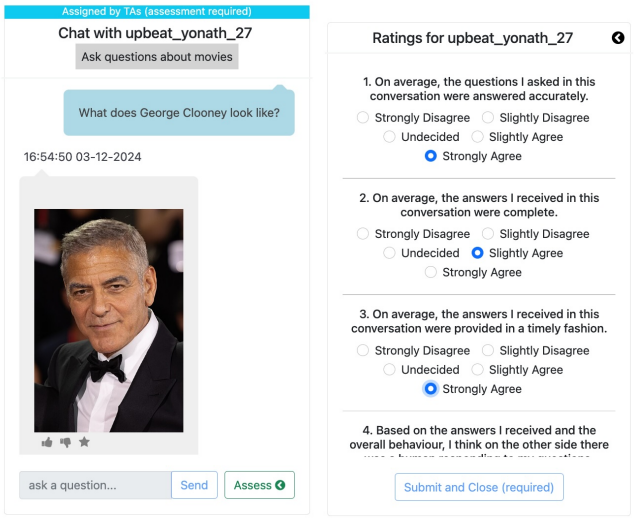
human evaluators and conversational agents. During the evaluation event, administrators can check the list of logged-in users, verify the status of chatrooms, inspect the live content of chatrooms, and visualize the survey feedback results. Feedback results are aggregated by each dimension used in the evaluation (e.g., accuracy, completeness, timeliness, humanness, and capability-specific ratings). The administrator, thus, can observe the distribution of the ratings that different human evaluators gave to different conversational agents in separate chatrooms.

Human Assessors. Human assessors can log in to the platform via the GUI and start a conversation. During an organized evaluation event, the human assessor is presented with multiple chatrooms simultaneously. Alternatively, when there is no predefined assignment, human assessors can request to start a conversation with a particular conversational agent by providing the agent's username. Requesting conversations with specific conversational agents becomes very useful during development and testing time, as agent developers can request to start a chatroom with the artificial agent they implemented and verify its performance. Within a chatroom, human assessors can read and write text messages from and to the conversational agent. Speakeasy can also render images present in messages sent by artificial agents that implement the capability to send messages with multimodal content (see Figure 1(a)).

Human assessors can provide feedback at a message and a conversation level. Human assessors can *like* or *dislike* a message and highlight it with a *star*. Chatrooms have a predefined, configurable duration. When the conversation reaches the time limit, Speakeasy presents the human assessor with a survey for each active chatroom. The current instance of the survey contains Likert-scale questions that assess the *accuracy*, *completeness*, *timeliness*, and *humanness* of the answers provided by the conversational agent and Likert-scale questions that refer to capabilities that human assessors expect the conversational agents to have (see Figure 1(b)). The survey questions can be easily configurable. Human assessors can also close a chatroom before the predefined duration.

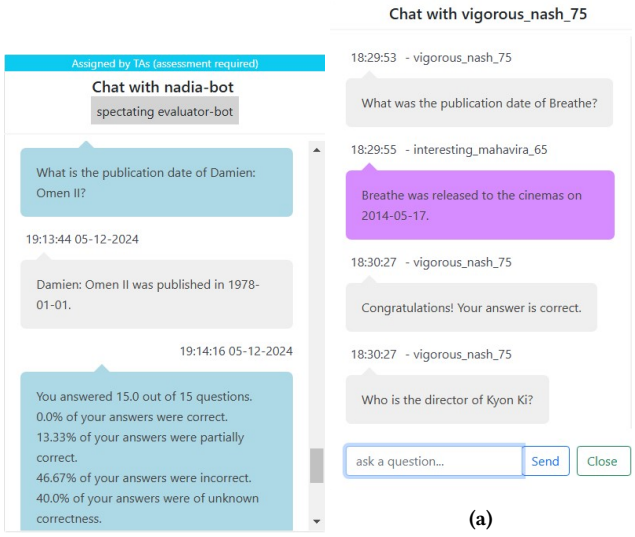
Artificial Assessors. Bot assessors get programmatic access to the Speakeasy platform via the API. Artificial assessors complement the evaluation of human evaluators. The artificial assessors use a collection of questions and answers and can assess different levels of answer correctness (see Figure 2(a)). More specifically, they identify if answers are entirely or partially correct, incorrect, mischief (i.e., the answer violates specific predefined guidelines), or of unknown correctness (i.e., the answer cannot be corroborated). Additionally, artificial assessors fill out the evaluation survey automatically. Besides this autonomous artificial assessor, we implemented an additional artificial agent that can assist human assessors by suggesting questions to ask during the evaluation (see Figure 2(b)).

Conversational Agents. Like artificial assessors, artificial agents interact with the platform via the API. Artificial conversational agents use the functionality of the API to login, logout, and get and post messages. When humans play the role of conversational agents because they are pretending to be or being compared with artificial agents in an evaluation event, they can log into the GUI and be presented with the chatroom(s) they have been assigned to or requested. As conversational agents, humans can just read and write text messages in chatrooms.



(a) (b)

Figure 1: (a) Conversation between a human and a bot. Bot names changed to preserve privacy. (a) & (b) Assessment of conversations in Speakeasy at a message- and conversation-level with likes/dislikes/stars or via the final survey. Image by Harald Krichel (CC-BY-SA 4.0). Source: https://commons.wikimedia.org/wiki/File:George_Clooney-69990.jpg



(a)

Figure 2: (a) The evaluator bot asks questions to the agents to be evaluated, explains the degree to which the answers are correct/incorrect, and provides the correct answers. Bot names changed to preserve privacy. (b) The tester bot suggests questions to humans to assess their agents' capabilities.

4 Speakeasy in the Classroom

The Speakeasy ecosystem has been extensively used in the past four iterations of the *Advanced Topics in Artificial Intelligence (ATAI)* course (taught in the fall of 2021, 2022, 2023, and 2024) held at the Department of Informatics at the University of Zurich (UZH).

Project. For the course project, the students were asked to implement artificial conversational agents capable of holding conversations with humans and answering natural language questions of two types: 1). *Closed questions*: questions that have a specific factual answer, e.g., "Who is the director of Titanic?" and 2). *Open questions*: questions with multiple valid answers, e.g., "Which movies do you recommend if I like watching dramas?" To answer these two types of questions, the students were invited to use the techniques introduced in the course (i.e., knowledge graphs, embeddings, recommendations, multimodal processing, and crowdsourcing).

Resources. As a common ground for the student projects, we provided a dataset from the movie domain generated from sampling Wikidata in the summer of 2021, including movies, actors, etc. Additionally, we modified the dataset by removing, editing, and adding triples (i.e., content) to generate a synthetically altered dataset for in-class evaluation purposes.

Evaluation Event. As part of the course, the setup for evaluating the agents is as follows: each human assessor holds conversations with a predefined number of artificial conversational agents in parallel. Human assessors are instructed to ask questions that cover as many conversational capabilities as possible and keep track of the artificial agents' performance. There are multiple evaluation rounds, and at the end of each round, the human assessors are asked to rate the *accuracy*, *completeness*, *timeliness*, and *human-likeness* of the answers on a five-point Likert scale from "Strongly Disagree" to "Strongly Agree" (see Figure 1) and each of the five possible capabilities the artificial agents should answer (i.e., knowledge graph question answering, embeddings, recommendations, multimodal processing, and crowdsourcing) on a five-point Likert scale from "Very Poor" to "Very Good", and an additional possible answer "Not applicable/I don't know" for cases when the human assessors did not test for that particular capability. They are also encouraged to comment on their user experience in a free text input field.

Statistics of the Evaluation Events. In total, 337 people (students, lecturers, teaching assistants, and computer science researchers from UZH) participated in the evaluation of 114 conversational artificial agents. After several rounds, where each human assessor evaluated multiple agents simultaneously, each artificial conversational agent was evaluated 18 times on average. In Figure 3, we show sample rating distributions that Speakeasy computed in 2023 for all conversational agents (in blue) and one selected conversational agent (in orange) for two assessed dimensions, accuracy and the capability of question answering the provided knowledge graph.

5 Demonstration Overview

This demo will present the Speakeasy ecosystem and invite conference attendees to participate in a conversational agent evaluation. We will recreate the scenario described in Section 4, and attendees will have the opportunity to be in the role of human evaluators and evaluation administrators (Section 3). We will showcase the evaluation of existing artificial conversational agents, including an implementation of Eliza [9], one of the earliest conversational agents developed by MIT in 1966, and a few artificial agents implemented in the context of the course.

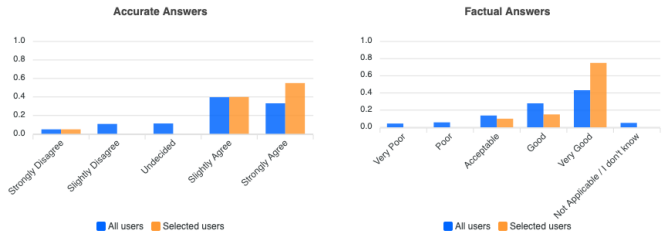


Figure 3: Overview of conversational agents' assessment.

6 Conclusion and Outlook

This demo presents Speakeasy, a Web-based ecosystem for evaluating conversational agents. Speakeasy has been extensively used in the educational context by supporting CS students in deploying and evaluating conversational agents they implement for a project assignment. As future work, our focus is two-fold: 1) improve the overall usage, functionality, and reach of the Speakeasy ecosystem to better support the needs of instructors and students in a variety of educational contexts, and 2) identify additional relevant use cases and their associated requirements for the Speakeasy ecosystem outside of the educational domain.

Acknowledgments

This work was partially supported by the University of Zurich Teaching Fund open_innovation grant 'Speakeasy' and the Swiss National Foundation through the projects 'MediaGraph' (contract no. 202125) and 'D3' (contract no. CRSII5_205975).

References

[1] Robert Epstein. 1992. The Quest for the Thinking Computer. *AI Mag.* 13, 2 (1992), 81–95.
[2] Aamir Hamid, Hemanth Reddy Samidi, Tim Finin, Primal Pappachan, and Roberto Yus. 2023. GenAIPABench: A benchmark for generative AI-based privacy assistants. *arXiv preprint arXiv:2309.05138* (2023).
[3] Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Sashikala Mishra, and Ajith Abraham. 2022. AI-based conversational agents: A scoping review from technologies to future directions. *IEEE Access* (2022).
[4] David M. W. Powers. 1998. The Total Turing Test and the Loebner Prize. In *Proceedings of the Joint Conference on New Methods in Language Processing and Computational Natural Language Learning*. 279–280.
[5] Aleksandra Przegalinska, Leon Ciechanowski, Anna Stroz, Peter Gloor, and Grzegorz Mazurek. 2019. In bot we trust: A new methodology of chatbot performance measures. *Business Horizons* 62, 6 (2019), 785–797.
[6] Joao Sedoc, Daphne Ippolito, Arun Kirubarajan, Jai Thirani, Lyle Ungar, and Chris Callison-Burch. 2018. Chateval: A tool for the systematic evaluation of chatbots. In *Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG)*. 42–44.
[7] Speakeasy Team. 2025. *Alan-s-Speakeasy/speakeasy: v1.0.0.* https://doi.org/10.5281/zenodo.14843064
[8] Alan M. Turing. 1950. Computing machinery and intelligence. *Mind* LIX, 236 (1950), 433–460.
[9] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *CACM* 9, 1 (1966), 36–45.
[10] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *CoRR* abs/2306.05685 (2023). https://doi.org/10.48550/ARXIV.2306.05685 arXiv:2306.05685