# Freeing the Wizard of Oz: A Highly Configurable Platform for Behavioural Experiments with Conversational User Interfaces


Emily Shiu*
University of Waterloo
Waterloo, Ontario, Canada
emily.shiu@uwaterloo.ca

Anthony Tieu
University of Waterloo
Waterloo, Ontario, Canada
anthony.tieu@uwaterloo.ca

Joanna Xiong
University of Waterloo
Waterloo, Ontario, Canada
joanna.xiong@uwaterloo.ca

Leo Qi
University of Waterloo
Waterloo, Ontario, Canada
leo.qi@uwaterloo.ca

Cosmin Munteanu
University of Waterloo
Waterloo, Ontario, Canada
cosmin.munteanu@uwaterloo.ca

Katherine White
University of Waterloo
Waterloo, Ontario, Canada
white@uwaterloo.ca



## Abstract

This interactive exhibit demonstrates LIDLbot, an all-in-one experiment platform designed to be used for psychological or linguistic experiments on conversational user interface interactions that require per-trial manipulations or measurements. LIDLbot features customizable modules and built-in voice agent and chatbot simulators to help researchers build streamlined experiments quickly. It offers benefits similar to Wizard-of-Oz setups, while not requiring live researcher administration of stimuli or responses. In this demo, we demonstrate the features of the platform, how to use it to set up experiments, and its use from a Participant's perspective.



## CCS Concepts

• **Human-centered computing → HCI design and evaluation methods**; *Interface design prototyping*; *User interface design.*

## Keywords

conversational agents, wizard of oz, voice interfaces, chat interfaces, psychology experiments

**ACM Reference Format:**
Emily Shiu, Anthony Tieu, Joanna Xiong, Leo Qi, Cosmin Munteanu, and Katherine White. 2026. Freeing the Wizard of Oz: A Highly Configurable Platform for Behavioural Experiments with Conversational User Interfaces. In *Proceedings of Demo at Graphics Interface 2026 (GI '26).* ACM, New York, NY, USA, 3 pages.


## 1 Introduction

Psychological and linguistic experiments often require trial-by-trial customizations, where specific factors are manipulated, or participant responses are probed on a per-trial basis. When translating these methods into the realm of conversational user interface (CUI) interactions, such rigorous investigations are often resource-intensive or inaccessible due to limitations of existing tools and methods, which include commercially available products, Wizard of Oz tools, proprietary APIs or toolkits, and simulators. However, these tools are not optimal for a variety of reasons as discussed below, and their limitations impede the conduct of highly rigorous psychological and linguistic CUI experiments.

Many studies use Wizard-of-Oz (WoZ) methods [9, 10, 12] – where a non-functional prototype is manually controlled by an experimenter in another room. These methods are well-suited for early low-fidelity prototyping or exploratory investigations of user interface interactions. However, the necessity of having a human experimenter to manually control the experiment in the background leads to challenges, especially in maintaining consistent system behaviour across trials and participants at scale [14]. Additionally, WoZ tools are generally programmed with specific experiments in mind, and a great deal of resources may be needed to repurpose them for other studies [14]. Finally, WoZ methods may raise ethical concerns around issues such as the use of deceit, especially with vulnerable populations or with participants who cannot fully provide informed consent [13].

As for other methods, some studies use commercially available products, such as a consumer smart speaker device [2, 5, 11, 18]. Although studies can be implemented relatively quickly with this type of setup, it is difficult to isolate and control (experimentally manipulate) specific factors that drive participant behaviours and experiences. Additionally, updates could alter the product's functionality while the study was ongoing, compromising validity and reproducibility. Other methods include using proprietary APIs or specialized toolkits [1, 6, 16, 17], such as the Alexa Skills Kit, Google Dialogflow, Microsoft Bot Framework, or large language models (LLMs). However, use of these tools can present accessibility challenges, requiring significant financial or technological resources to access, or specialized expertise to implement [17]. Finally, some studies use CUI simulators that accept simple text entries and return predetermined text or audio outputs [3, 4, 7, 8, 15]. Simulators allow for many custom configurations, but as with WoZ tools, they are usually designed for specific experiments, and would require adaptation to accommodate other designs.

## 2 LIDLbot

Currently, there is no tool that offers a high degree of control, consistency, flexibility, and accessibility in creating CUI interaction experiments. To address the various challenges, we present LIDLbot – named after our *Lab for Infant Development and Language* (LIDL; pronounced like "little"). LIDLbot is a modular experimental platform with synchronized stimulus presentation that allows behavioural scientists to experimentally study CUI-related questions. As a platform, LIDLbot is an all-in-one package that includes


*Demo at Graphics Interface 2026 (GI '26), Waterloo, ON, Canada*

customizable components to build streamlined experiments. Such components include built-in voice or chat CUI simulators, text or image stimuli displays, and templates for embedding web surveys. Additionally, unlike WoZ tools where the participant interacts with the prototype device while the researcher operates a controller device, LIDLbot is designed to have two machines used solely by the participant. The platform is intended for use with task-based experiments, where participants interact with stimuli or a CUI system on a per-trial basis.

LIDLbot has three major advantages: First, its modular flexibility means that researchers do not need to build functional CUI prototypes from scratch for every new experiment. Second, LIDLbot allows for considerable control over experimental conditions. And, third, it allows for this control despite the fact that researcher operation is not required during the experimental session (beyond the initial setup).

## 3 Overview

As an integrated platform, LIDLbot facilitates the setup and control of experiments through automating both the administration of stimuli and CUI system responses. A **Researcher** designs specifies an order for the built-in customizable modules, which will then automatically transition seamlessly from one to the next. The **Participant** then uses both computers, and can progress through the study themselves.

The application has two running modes – dubbed the **Assistant Instance** and **Study Instance** – initialized on the two independent computers. (See Fig. 1 for visualization.) The **Study Instance** is used to display content such as instructions, stimuli prompts, and surveys. Meanwhile, the **Assistant Instance** is used to display the built-in voice agent or chatbot. Synchronization between the devices is achieved by using a common seed value generated based on a Participant ID entered by the Researcher at the start of an experiment session. Because of this method, content can be synchronized between the two machines without the need for server port configuration, or a wired/wireless connection.

In a typical experiment using LIDLbot, the Participant uses the CUI prototype on one of LIDLbot's computers (in either voice or chat mode) to complete tasks involving stimuli displayed on its other computer. Following the Participant's verbal or written input, a predetermined response will be given by the CUI for that trial, as customized by the Researcher.

### 3.1 Top-level Structure

LIDLbot follows a structured hierarchy to customize experiment design: **Set** > **Sequence** > **Module** (Figure 2).

A **Set** defines the overall flow of an experiment. Multiple Sets can be created and used as presets for different experiment configurations, such as when multiple counterbalance orders are needed. Each Set consists of two **Sequences**: one for the Assistant Instance and one for the Study Instance. The Sequences then determine the order of **Modules** shown on each Instance. Finally, **Modules** are used to display content defined by the Researcher. Types of Modules currently available include Web Survey, Content Viewer, Web Content Display, Session Code Entry, Prompter, Voicebot, and Chatbot.

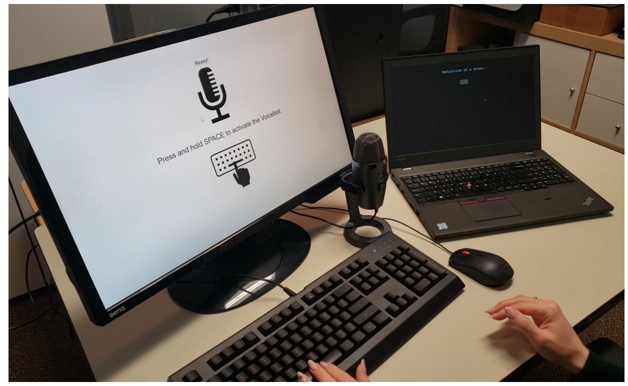

**Figure 1: A study setup using LIDLbot's built-in voice agent. Two computers are used by a Participant side-by-side; one running the Assistant Instance, and the other running the Study Instance.**

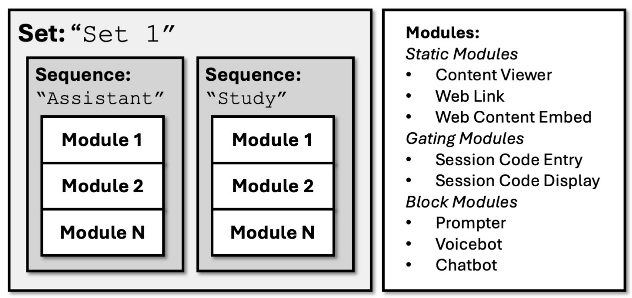


**Figure 2: Diagram depicting the relationship of experiment components as implemented in LIDLbot, and a list of available Modules.**

Using these Modules, a variety of content types are available to support various psychological and linguistic experiment setups.

Configuration can be performed using an Experiment Configurator or by directly editing the platform's `.json` files. Further details about customization options can be found at the demo table.

### 3.2 Demo

The demo will consist of LIDLbot running on two laptops on a table. The attendee will first be given a brief introduction and overview about the system. They will then get the chance to experience LIDLbot from the perspective of a Participant. The attendee will see the different Modules as implemented in a sample experiment, and interact with the Voicebot and Chatbot Modules. Attendees will give voice and text inputs to LIDLbot and experience how convincing LIDLbot feels from a Participant's perspective. Afterwards, the attendee will receive a quick tour of how to use the Experiment Configurator to set up a study.

Additional details on LIDLbot's implemention and configuration will be available at the demo table. The codebase for the LIDLbot Experiment Platform is open source and available under the MIT License at: https://github.com/Lab-for-Infant-Development-and-Language/LIDLbot-Experiment-Platform

## Acknowledgments

This work has been made possible through funding from The Games Institute Seed Grant. The authors would like to thank our developers: Devin Li, Alex Ly, Nehemiah Negussie, and Jean Wang; and our research assistants: Mihira Anand, Curtis Fung, Kashish Kahlon, Cindy Lo, Katie Martin, Leah Wong.

Generative AI was utilized to assist in building the system's codebase, and with formatting and syntax in LaTeX.

## References

[1] Theo Araujo. 2020. Conversational Agent Research Toolkit: An alternative for creating and managing chatbots for experimental research1. *Computational Communication Research* 2, 1 (Feb. 2020), 35–51. doi:10.5117/ccr2020.1.002.arau

[2] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300473

[3] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9 (Sept. 2010), 2355–2368. doi:10.1016/j.pragma.2009.12.012

[4] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121, 1 (Oct. 2011), 41–57. doi:10.1016/j.cognition.2011.05.011

[5] Janghee Cho. 2018. Mental Models and Home Virtual Assistants (HVAs). In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–6. doi:10.1145/3170427.3180286

[6] Michelle Cohn, Santiago Barreda, Katharine Graf Estes, Zhou Yu, and Georgia Zellou. 2024. Children and adults produce distinct technology- and human-directed speech. *Scientific Reports* 14, 1 (July 2024), 15611. doi:10.1038/s41598-024-66313-5

[7] Michelle Cohn, Bruno Ferenc Segedin, and Georgia Zellou. 2022. Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics* 90 (Jan. 2022), 101123. doi:10.1016/j.wocn.2021.101123

[8] Michelle Cohn and Georgia Zellou. 2021. Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments. *Frontiers in Communication* 6 (July 2021). doi:10.3389/fcomm.2021.675704

[9] Maria J. Galvez Trigo, Martin Porcheron, Joy Egede, Joel E. Fischer, Adrian Hazzard, Chris Greenhalgh, Edgar Bodiaj, and Michel Valstar. 2021. ALTCAI: Enabling the Use of Embodied Conversational Agents to Deliver Informal Health Advice during Wizard of Oz Studies. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*. ACM, Bilbao (online) Spain, 1–5. doi:10.1145/3469595.3469621

[10] Jieun Kim, Woochan Kim, Jungwoo Nam, and Hayeon Song. 2020. "I Can Feel Your Empathic Voice": Effects of Nonverbal Vocal Cues in Voice User Interface. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3334480.3383075

[11] Souheila Moussalli and Walcir Cardoso. 2020. Intelligent personal assistants: can they understand and be understood by accented L2 learners? *Computer Assisted Language Learning* 33, 8 (Nov. 2020), 865–890. doi:10.1080/09588221.2019.1595664

[12] Martin Porcheron, Joel E. Fischer, and Michel Valstar. 2020. NottReal: A Tool for Voice-based Wizard of Oz studies. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, 1–3. doi:10.1145/3405755.3406168

[13] Janet C. Read. 2018. *Wizard of Oz with Children – Ethical dilemmas & experiences from the field*. https://fase.it.lut.fi/wp-content/uploads/2018/08/woz.pdf Accessed: 2026-03.

[14] Stephan Schlögl, Gavin Doherty, and Saturnino Luz. 2014. Wizard of Oz Experimentation for Language Technology Applications: Challenges and Tools. *Interacting with Computers* 27, 6 (May 2014), 592–615. doi:10.1093/iwc/iwu016

[15] Huiyang Shen and Min Wang. 2023. Effects of social skills on lexical alignment in human-human interaction and human-computer interaction. *Computers in Human Behavior* 143 (June 2023), 107718. doi:10.1016/j.chb.2023.107718

[16] Laura Spillner and Nina Wenig. 2021. Talk to Me on My Level – Linguistic Alignment for Chatbots. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (Toulouse & Virtual, France) *(MobileHCI '21)*. Association for Computing Machinery, New York, NY, USA, Article 45, 12 pages. doi:10.1145/3447526.3472050

[17] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–35. doi:10.1145/3637364

[18] Yunhan Wu, Justin Edwards, Orla Cooney, Anna Bleakley, Philip R. Doyle, Leigh Clark, Daniel Rough, and Benjamin R. Cowan. 2020. Mental Workload and Language Production in Non-Native Speaker IPA Interaction. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3405755.3406118

Received 6 May 2026