

Artificial intelligence (AI) technologies are emerging and affecting our lives in many aspects. However, the majority of individuals are merely *users* of AI with little capability to adapt AI to their own heterogeneous needs. Most software developers also encounter significant barriers to developing AI-powered applications. **In my research, I design, build, and study interactive systems to empower end-users and non-AI-expert developers to configure, extend, and create AI-powered applications.**

**My work approaches AI system research from a human-centered perspective**, where I design and build new AI systems and techniques based on the insights from my studies of the capabilities, contexts, preferences, and needs of different user groups. As a human-computer interaction (HCI) researcher with a background in natural language processing (NLP) and end-user software engineering, I apply HCI design, user research, and evaluation methods in designing new interfaces and interaction techniques to make AI techniques more accessible and more useful for end users and developers, while at the same time, developing new AI techniques that enable novel interactions.

**My interdisciplinary research has resulted in contributions across premier academic venues in HCI, NLP, programming tools, and system research** (e.g., CHI, UIST, CSCW, ACL, MobiSys, VL/HCC), producing 15 major peer-reviewed publications and book chapters, with 10 first-authored. My first-authored work has been awarded two Best Paper [3,6] and a Best Paper Honorable Mention [1] awards and resulted in sustained collaborations with academic and industry partners (UCLA; UCSD; Yale; Microsoft Research; Verizon; J.P. Morgan). Over the last 5 years, I helped raise over \$1 million in research grants to support my research.

Below, I summarize two areas in my work. The first focuses on the end-users, where I developed and deployed a multi-modal interactive-task-learning agent that allows them to teach new tasks and concepts using a combination of natural language instructions and demonstrations on the graphical user interfaces (GUIs) of existing apps [1–8]. The second focuses on non-expert software developers, lowering the barrier for them to build AI-powered systems [9,11,13,14]. My future research work will continue this trajectory of democratizing AI to empower individuals with two areas of emphasis: the future of work with AI and the natural interaction modalities in AI development tools.

## **SUGILITE: HUMAN-AI COLLABORATION IN INTERACTIVE TASK LEARNING**

Enabling end-users to automate their tasks using intelligent agents has been a long-standing objective in both the HCI and the AI communities. Instead of letting the users learn programming tools and languages, I developed a new intelligent agent named SUGILITE [1] that learns the human’s natural way of instructing tasks and concepts.

**SUGILITE uses a new multi-modal interactive task learning approach that enables end users to teach new tasks and concepts using combined natural language instructions and demonstrations on existing app GUIs.** This project presents a new human-AI interaction paradigm for task learning, where the existing third-party app GUIs are used as a medium for users to communicate their intents with an AI agent instead of as the interfaces for interacting with the underlying computing services. Compared to prior systems, SUGILITE made significant progress in usability, applicability, generalizability, expressiveness, robustness, and shareability. SUGILITE has been open-sourced on GitHub, deployed in-situ through Google Play Store, and evaluated in several user studies.

**On the HCI side, SUGILITE features novel intelligent interfaces and interaction techniques that facilitate the collaboration between the user and the AI system.** The multi-modal approach allows the user to first describe their intended task in natural language, and demonstrate task procedures and concept on existing app GUIs. I created a collaborative approach for users to disambiguate the intents of their actions [4]. After an ambiguous demonstrated action, the agent asks the user to clarify through natural language conversations. The system provides an interactive visualization that highlights the agent’s understanding of the user’s utterances and prompts the user to explain specific aspects (Figure 1). My work extends and contributes to frameworks in designing human-AI interaction such as *mixed-initiative interaction* and *mutual disambiguation* in multi-modal interfaces.

My work in SUGILITE introduced a top-down lazy-evaluation dialog framework [8] for supporting concept and conditional instructions. Informed by results from a formative study, this new framework allows the user to start with saying an automation rule at a high level with ambiguous and vague concepts, procedures, and conditions. The agent then recursively resolves the ambiguities and vagueness with the user through multi-turn conversations, seeks clarifications from the user, and guides the user to define new concepts and procedures if needed (Figure 2).

I am particularly interested in the issue of error handling in human-AI interaction, which is often overlooked in prior research. In SUGILITE, I designed a new multi-modal interface [3] that helps users discover and recover from

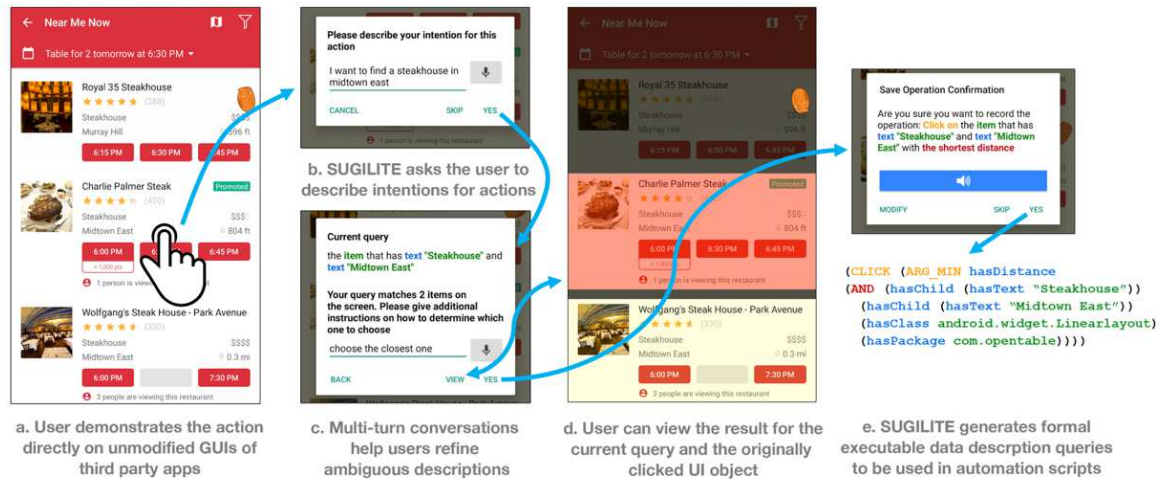


Figure 2. The screenshots of SUGILITE’s demonstration mechanism and its multi-modal mixed-initiative intent clarification process for the demonstrated actions.

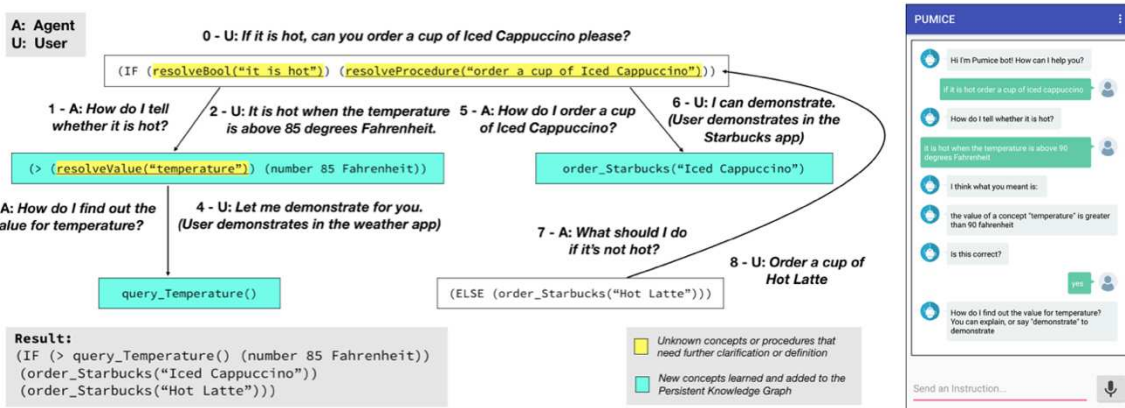


Figure 1. An example dialog structure when SUGILITE learns a new task that contains a conditional and new concepts. The numbers indicate the sequence of the utterances. The screenshot on the right shows the corresponding conversational interface.

conversational breakdowns based on the grounding theory in communication. This interface displays the system’s state of understanding through GUI screenshots, allows the user to refer to third-party apps and their GUI screens in conversations as inputs for intent disambiguation, and enables the user to repair breakdowns using direct manipulation on these screenshots.

**On the applied AI/NLP side, I created new AI-powered system capabilities.** For example, a key challenge in multi-modal task learning is to ground the user’s natural language explanations to the corresponding GUI elements. I developed a new technique in SUGILITE that represents each screen as a *UI Snapshot Graph* [4,7]. A floating semantic parser parses the user’s natural language instruction into graph queries on this graph in a simple and flexible LISP-like functional query language. An iterative human-in-the-loop pipeline executes the query on the graph, validates if the result matches the entity that the user originally demonstrated, and seeks additional input from the user if the result does not match, or matches more than one entity (Figure 3). The goal of this process is to generate a query that uniquely matches the target UI element and also reflects the user’s underlying intent.

This GUI grounding approach allows SUGILITE to learn generalized procedures (e.g., to order *any* kind of beverage from Starbucks) from a demonstration of a specific instance of the task (e.g., ordering an iced cappuccino) [1]. SUGILITE achieves this by grounding the utterance (e.g., “order a cup of iced cappuccino”) to the target UI elements (e.g., click on the menu item that has the text “Iced Cappuccino”) and the arguments of the demonstrated actions. It analyzes the hierarchical structure of GUI to extract the potential values of the parameter. The GUI-grounding approach is also used for disambiguating the procedures to invoke [1] and for learning generalized concepts [8].

Another major barrier to the adoption of GUI-based interactive task learning agents has been the privacy concern in sharing the user-taught scripts. To address this, I developed a new mechanism [2] for the privacy-preserving sharing of GUI-based demonstrational scripts. This mechanism can identify and obfuscate personal information at the time of sharing with minimal user intervention. It collects and aggregates hashed texts from apps on users' phones in everyday use. This aggregated data allows the system to identify potentially personal fields in scripts, and obfuscate their values. At runtime, the obfuscated fields can be rebuilt locally, which helps preserve the shared scripts' transparency, readability, robustness, and generalizability.

I also developed a new technique for computationally modeling GUIs and GUI interactions, generating semantic embeddings for use in neural networks [10]. With the rise of data-driven methods for modeling GUI interactions, GUIs have become not only interfaces for users to interact with computing services, but also valuable data sources that encode the task flow, interactions, and design patterns of the corresponding apps. My new self-supervised SCREEN2VEC technique generates vector representations for GUI Screens and components using their textual contents, visual design and layout patterns, and app-specific contexts without requiring human annotation efforts, addressing an important gap in the prior literature. The result vectors are shown to be effective and useful in many downstream tasks for modeling user tasks, querying GUI screens with natural language commands, and embedding composability.

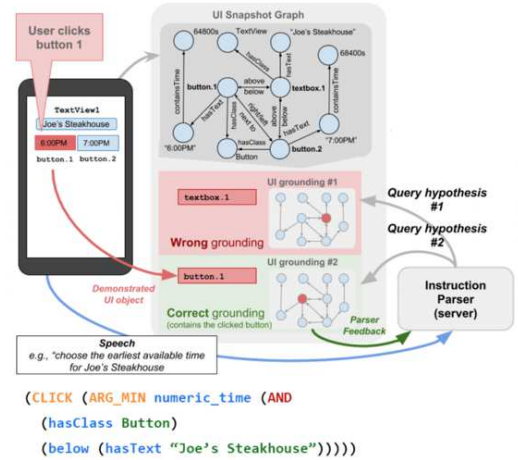


Figure 3. The UI Snapshot Graph for grounding natural language instructions to GUIs

## LOWERING THE BARRIER TO DEVELOPING AI-POWERED APPLICATIONS

I also created tools to support non-expert developers to build AI-powered applications. There are more than 27 millions software developers in the world. Most of them do not have specialized expertise in AI domains, which prevents them from adopting the recent AI advances in their own application. My work seeks to lower the barrier to developing AI-powered applications through new interactive developer tools.

KITE [9] is a practical system I created that enables developers to bootstrap task-oriented conversational bots from existing mobile apps. Using traces of users using an app, KITE can automatically derive a task model, a graph of actions, and the associated inputs representing different task execution paths. On top of them, KITE generates a natural language interface for the task model with questions and answers using a Transformer-based neural network (Figure 4). The developer can then iteratively test and revise the auto-generated bot in KITE's interactive interface (Figure 5). KITE was not only evaluated in a lab study and published at MobiSys 2018, but also patented and is in the process of being incorporated into commercial products.

In collaboration with other teams of researchers, we developed three additional developer tools for AI-powered applications: GENO, PRIVACYSTREAM, and WIKIBRAIN. GENO [13] is a tool for adding the support for coordinated multi-modal voice inputs (e.g., moving *this event* to the next Monday) to existing web apps without requiring NLP and multi-modal interaction expertise. PRIVACYSTREAM [11] introduces a new Android developer framework for

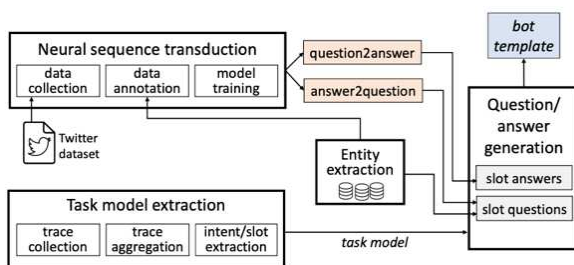


Figure 4. The processing pipeline of KITE's approach

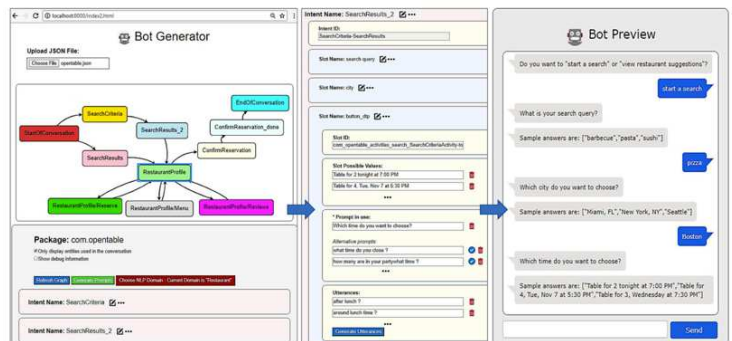


Figure 5. The interactive bot developing and testing interfaces of KITE

accessing and processing personal data (e.g., sensory inputs, location, contacts, photos) as a stream in a functional programming model, making it easier for developers to use personal data while simultaneously making it easier to analyze how that personal data is processed. WIKIBRAIN [14] democratizes computation with Wikipedia knowledge for developers by providing simple access to the Wikipedia data and the state-of-art Wikipedia-based algorithms. All these tools have been open-sourced on GitHub and used by the communities of developers and researchers.

## FUTURE RESEARCH AGENDA

In the future, I plan to continue my line of research in **democratizing AI to empower individuals**. In this section, I outline two research themes that I am the most excited about.

### End-User-Programmable AI Systems for the Future of Work

SUGILITE presents an example of supporting end user programming for task automation. For the next step, I am particularly interested in developing end-user-programmable AI systems to specifically address the challenges faced by the workforce in the era of “the fourth industrial revolution”. While the adoption of AI technologies has significantly increased the overall productivity of society, the gains have mostly gone to a small group at the top while many workers have been stagnated in wages or displaced by AI. I attribute much of this phenomenon to the *top-down* approach used in developing and deploying AI. While many workers, especially the rapidly expanding group of “gig workers”, participate in the corporate-provided AI infrastructures, they often act as interchangeable and replaceable components (e.g., Uber drivers, Instacart shoppers, data annotators) in large algorithm-directed workflows with little agency. The AI systems are often developed from a corporate’s perspective with its best interest in mind, with little support for individual worker’s personal preferences, motivation, and creativity.

**I propose to explore a *bottom-up* approach that helps individual workers automate and augment their work with AI systems.** Instead of replacing workers with AI, this approach can emancipate workers from mundane repetitive tasks, enabling them to focus on the creative and social aspects of work they love that AI systems cannot feasibly replace in near future. In our formative study with SUGILITE, many users stated they would love to automate their work tasks. The current SUGILITE is already able to learn work-related tasks that involve basic lookups, simple conditional rules, and well-defined actions through app GUIs. In the future, I plan to conduct contextual user research to investigate the automation needs of workers in different disciplines, study the societal impacts of end-user-programmable AI systems for the future-of-work, and develop more effective and more personalized interactive AI systems to assist individual users with their work-related tasks in their best interests.

A significant barrier to the bottom-up approach is the data ownership problem. An effective AI system usually requires a large amount of data to train, which is often only feasible for large organizations. I consider the recent advances in general-purpose language (e.g., BERT, GPT-3), reasoning (e.g., COMET, Atomic), and task completion (e.g., Almond) models an important step towards end-user programmable AI. These models are great fits for *few-shot learning* and *transfer learning*, allowing effective incorporation of general knowledge with small amounts of data and instructions from end users. For my future research, I plan to focus on designing the interaction between the user and the AI system in these scenarios, bridging the transparency and explainability of AI systems with the user’s natural cognitive models for the task, and enabling the user to provide useful inputs that boost the generalizability, expressiveness, robustness of the AI system through effective human-AI collaboration.

### Natural Interaction Modalities in Developer Tools

In my past work of designing different interactive developer tools, I have used several distinct input modalities and interaction methods for program specifications, such as conventional source code, natural language spoken instructions and conversations, program synthesis with demonstrated examples, and direct manipulation in visual programming and GUIs. For the next step, **I plan to further explore the design space of interactive tools for diverse types of programming activities in the full end-to-end lifecycle of developing AI systems.** such as data labeling, data wrangling, exploratory analysis, model tuning, error analysis, bias mitigation, system deployment and monitoring, system extension, and performance optimization.

The advances in fields such as natural language processing, sensing for gesture and gaze tracking, interactive data visualization, program synthesis, and computational modeling of programming activities through software repository mining have laid the technical foundation for my next breakthrough in multi-modal interaction for software development. In my design process, I will specifically focus on the needs, capabilities, and routines of different types of developers, creating dynamic developer interfaces that are adaptive to these changing factors.

## REFERENCES

- [1] **Toby Jia-Jun Li**, Amos Azaria, and Brad A. Myers. SUGILITE: Creating Multimodal Smartphone Automation by Demonstration. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, ACM. **Best Paper Honorable Mention**.
- [2] **Toby Jia-Jun Li**, Jingya Chen, Brandon Canfield, and Brad A. Myers. Privacy-Preserving Script Sharing in GUI-Based Programming-by-Demonstration Systems. In *Proceedings of the ACM on Human-Computer Interaction 4*, CSCW1 (May 2020), ACM.
- [3] **Toby Jia-Jun Li**, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*, ACM. **Best Paper**.
- [4] **Toby Jia-Jun Li**, Igor Labutov, Xiaohan Nancy Li, Xiaoyi Zhang, Wenze Shi, Wanling Ding, Tom M. Mitchell, and Brad A. Myers. APPINITE: A Multi-Modal Interface for Specifying Data Descriptions in Programming by Demonstration Using Natural Language Instructions. In *2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC '18)*, IEEE.
- [5] **Toby Jia-Jun Li**, Igor Labutov, Brad A. Myers, Amos Azaria, Alexander I. Rudnicky, and Tom M. Mitchell. Teaching Agents When They Fail: End User Development in Goal-oriented Conversational Agents. Chapter of *Studies in Conversational UX Design*, Springer, 2018.
- [6] **Toby Jia-Jun Li**, Yuanchun Li, Fanglin Chen, and Brad A. Myers. Programming IoT Devices by Demonstration Using Mobile Apps. In *End-User Development, Proceedings of 6th International Symposium on End User Development (IS-EUD '17)*, Springer. **Best Paper**.
- [7] **Toby Jia-Jun Li**, Tom Mitchell, and Brad Myers. Interactive Task Learning from GUI-Grounded Natural Language Instructions and Demonstrations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20): System Demonstrations*.
- [8] **Toby Jia-Jun Li**, Marissa Radensky, Justin Jia, Kirielle Singarajah, Tom M. Mitchell, and Brad A. Myers. PUMICE: A Multi-Modal Agent that Learns Concepts and Conditionals from Natural Language and Demonstrations. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST '19)*, ACM.
- [9] **Toby Jia-Jun Li** and Oriana Riva. KITE: Building conversational bots from mobile apps. In *Proceedings of the 16th ACM International Conference on Mobile Systems, Applications, and Services (MobiSys '18)*, ACM.
- [10] **Toby Jia-Jun Li**, Lindsay Popowski, Tom M. Mitchell, and Brad A. Myers. Screen2Vec: Semantic Embeddings of GUI Screens and GUI Components. *In Submission*.
- [11] Yuanchun Li, Fanglin Chen, **Toby Jia-Jun Li**, Yao Guo, Gang Huang, Matthew Fredrikson, Yuvraj Agarwal, and Jason I. Hong. PrivacyStreams: Enabling Transparency in Personal Data Processing for Mobile Apps. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 3 (September 2017), 76:1–76:26, ACM.
- [12] Brad A. Myers, Andrew J. Ko, Chris Scaffidi, Stephen Oney, YoungSeok Yoon, Kerry Chang, Mary Beth Kery, and **Toby Jia-Jun Li**. Making End User Development More Natural. Chapter of *New Perspectives in End-User Development*, Springer, 2017.
- [13] Ritam Sarmah, Yunpeng Ding, Di Wang, Cheuk Yin Phipson Lee, **Toby Jia-Jun Li**, and Xiang “Anthony” Chen. Geno: A Developer Tool for Authoring Multimodal Interaction on Existing Web Applications. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST '20)*, ACM.
- [14] Shilad Sen, **Toby Jia-Jun Li**, WikiBrain Team, and Brent Hecht. WikiBrain: Democratizing computation on Wikipedia. In *Proceedings of the 10th International Symposium on Open Collaboration (WikiSym + OpenSym 2014)*, ACM.