Introduction of Rasa & Juji & GPT-2
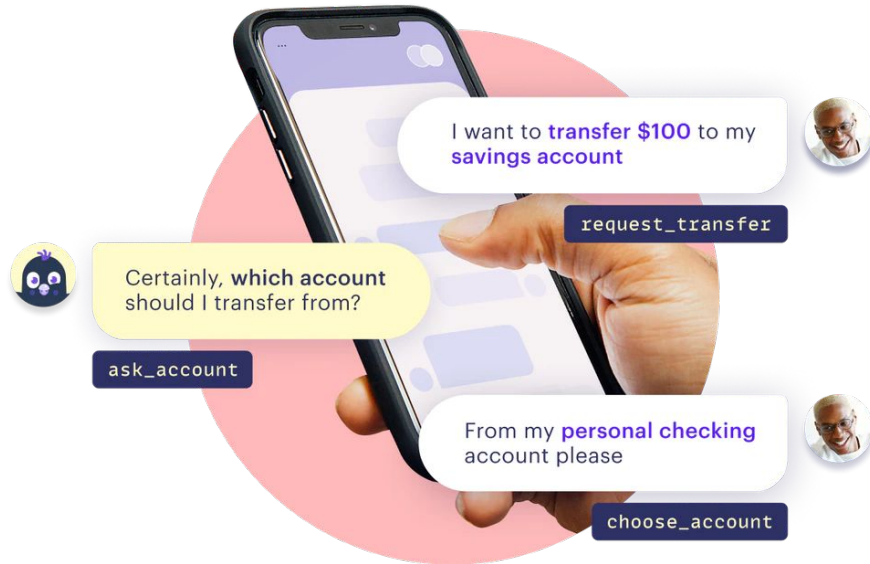
Jinfen Li

Commercial Chatbot

Rasa & Juji

Rasa

Rasa is an open source machine learning framework for automated text and voice-based conversations. Understand messages, hold conversations, and connect to messaging channels and APIs.

Mechanism

It is mainly composed of Rasa NLU (natural language understanding) and Rasa Core (mainly constructing conversations). Both can be used independently.

1. **Learning Language(s)**

   Provide a large training corpus. Rasa transform words into vectors.

2. **Learning Intents**

   Provide Rasa with NLU data. Intent represents a user's intention of a sentence. Each intent requires 12–20 training sentences.

   Rasa can identify entities with training data and NER algorithm. Entities could be detected from user inputs, e.g., [Mexican](cuisine).

3. **Learning Skills**

   Teach Rasa various actions in a file called domain to establishing a dialogue.

   Rasa's reaction/action can be defined by templates (default strings as responses), actions (any python script), etc. If Rasa knows that the user means to "greet," you can make it say 'hi', or make it look into your database according to entities extracted from users' statements (such as 'Mexican ' ~ 'cuisine').

Mechanism

4. Context-aware Conversations
   Rasa conversations, called "stories," to train Rasa's conversational skills.

   People's actual conversations can go in different directions according to the previous contexts, we can teach Rasa to deal with various situations through storytelling.

   We don't need to provide a story for all possible scenes, Rasa can deal with unseen conversations by combining and recreating old stories.

Rasa has many other convenient features, such as interactive learning (no need to write stories in advance), fallback (when the confidence of intent recognition is lower than a threshold you set, you can make it perform a default action such as using a seq2seq model).

Pros & Cons

Pros

- Sophisticated NLU engine
- Open source community
- Protect process sensitive consumer data: include the option to build and run your bot on-premises, which could be a deal-breaker for heavily regulated industries.

Cons

- Rasa isn't necessarily for beginners: Rasa's machine learning tools might strike you as a bit intimidating
- Computational consuming: spaCy, one of Rasa's default pipelines for processing user inputs, uses a lot of memory. Depending on your server configuration, spaCy could slow down other processes.

Applications

Health Care Assistant

- Tia: a Female healthcare chatbot that allows users to ask health questions
- nib: nib provides health and medical and travel insurance. nibby is a digital assistant embedded in the nib online member portal and mobile app, and available after hours via voice.
- Dialogue: Dialogue's mission is to improve humanity's well-being by reducing barriers to quality care. Dialogue set out to automate the patient intake process by developing a conversational AI assistant capable of gathering information about patients' medical conditions through a conversational evaluation before introducing them to a member of the care team.

Financial Assistant

- Raiffeisen: a leading Swiss retail bank, wanted to appeal to younger customers. Using Rasa, Raiffeisen developed an AI assistant to handle longer conversations and train non-technical team members.

Customer Service Assistant

- ERGO: a leading European insurance company, wanted to expand their customer service operations to provide 24/7 coverage—all while reducing costs. ERGO deployed Rasa Open Source and Rasa X, laying the framework to automate over 30% of customer requests before continuing to full automation.

More details please check https://rasa.com/customers/.

Research Paper

Implementing Rasa in Other Domain

➢ **Healthcare:**
- Quy Tran, B., Van Nguyen, T., Duc Phung, T., Tan Nguyen, V., Duy Tran, D., & Tung Ngo, S. (2021, February). FU Covid-19 AI Agent built on Attention algorithm using a combination of Transformer, ALBERT model, and RASA framework. In *2021 10th International Conference on Software and Computer Applications* (pp. 22-31).

➢ **Education:**
- Meshram, S., Naik, N., Megha, V. R., More, T., & Kharche, S. (2021, August). College Enquiry Chatbot using Rasa Framework. In *2021 Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1-8). IEEE.
- Windiatmoko, Y., Rahmadi, R., & Hidayatullah, A. F. (2021, February). Developing Facebook Chatbot Based on Deep Learning Using RASA Framework for University Enquiries. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1077, No. 1, p. 012060). IOP Publishing.

➢ **Library**
- Bagchi, M. (2020). Conceptualising a Library Chatbot using Open Source Conversational Artificial Intelligence. *DESIDOC Journal of Library & Information Technology*, *40*(6).

Enhancing Rasa NLU model

- Gamage, B., Pushpananda, R., & Weerasinghe, R. (2020, November). The impact of using pre-trained word embeddings in Sinhala chatbots. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)* (pp. 161-165). IEEE.
- Jiao, A. (2020, March). An intelligent chatbot system based on entity extraction using RASA NLU and neural network. In *Journal of Physics: Conference Series* (Vol. 1487, No. 1, p. 012014). IOP Publishing.

Juji

Juji cognitive AI assistants are powered by an advanced, humanized version of Artificial Intelligence. These AI assistants work wonders with instilled empathy and responsibility, and work well with others—humans and other AI alike.
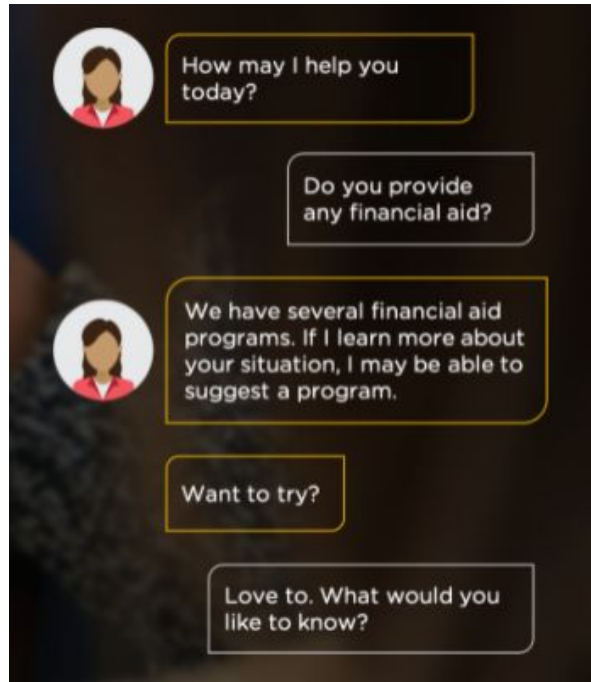
Mechanism

1. Define a main chat flow

Define a custom workflow of your AI assistant. Use Juji chatbot templates and pre-built AI dialogs without defining every intent or entity from scratch.

2. Upload Q&A pairs

Define Q&A pairs in a CSV file or in a GUI table. Upload the file or submit the entries to enable your AI assistant to answer user questions during a chat.

3. Auto-generate an AI

Juji automatically generates a custom AI assistant and manages its conversation context, enabling out-of-the-box cognitive intelligence.

Interface

Interface

MAIN CHAT FLOW   **Q&A BOARD**   **CHATBOT SETTINGS**

Set up chatbot to answer user questions/comments

You can add or edit user questions and corresponding chatbot responses in the table below and then submit them. Alternatively, you can download a CSV template, add/edit Q&A pairs in the CSV, and then upload the file. Less ⌃

New Q&A   Search Q&A   More Options

Submit

Manage Q&As I entered	Manage Q&As from users	↻
User Questions/Comments	Chatbot Responses	

Here displays user-asked questions that either weren't answered or answered with a low confidence by the chatbot. Verify or edit chatbot responses to improve the chatbot.

currently, no such questions exist yet.

Download CSV

Upload CSV

Interface

neutral stock persona

Albert   Alberta   Alicia   Ava   Clara

Collin   Jason   John   Jordan   Juji   Kai

cheerful stock persona   neutral stock persona

Kaya   Matt   Vita   Custom

Display Control

Display control by number of words: 50   words

Enable chat progress bar:

Enable user feedback during chat:

Conversation Tempo

Turn pace: 0   second(s)

Message pace: 1500   millisecond(s)

Response Control

General Settings

Minimum response length: 5   word(s)

Handling Unknown User Input

Default chatbot response to unknown user input: *Click to enter a custom response. Otherwise system default response will be used.*

Ask users to opt-in for an email notification when an answer to their question becomes available

Fallback Handling

During a conversation, users often digress from the main chat flow, such as asking a question or changing the chat flow. To handle highly diverse and complex user digressions, Juji has pre-built a rich set of fallback topics as listed below. You can use them all or just some of them.

Pros & Cons

Pros:

- No coding or technical skills required
- Easy, straightforward, simple UI
- Customizable logic
- Automatic chat flow management saves time and energy.

Cons:

- Resources is little
- Option is few

Use Cases

Customer Service - Helpdesk Automation

- During the COVID-19 pandemic, a group of volunteers has built Jennifer, a chatbot on Juji to combat misinformation and answer public's questions regarding COVID-19.

Education - Personalize Online Education

- **Personal Learning Assistant**: to understand a student's unique learning style and learning needs, provide proactive, personalized guidance, such as recommending suitable learning method or materials.

- **Teaching Assistant**: to aid an instructor in helping his/her students. My tasks include answering student questions about a particular course, sending instructor reminders, and collecting student feedback

User Research - Conversational Survey

- Juji helped build a chatbot to interview about 300 gamers and elicit their opinions about game trailers.

Healthcare - Personalize Telehealth

- **Personal Wellness Assistant**: to check in on users regularly (e.g., daily), monitoring their well-being and guiding them through wellness routines, such as writing a reflective journaling for maintaining mental well-being.

- **Care Information Assistant**: to engage with users 24x7 to provide them with healthcare or wellness information on demand. I can interpret natural language inquiries and retrieve requested information directly, relieving users from wading through multiple websites or web pages to find such information.

- **Care Triage Assistant**: to engage with users and triage care requests. My tasks include gathering critical data, answering care questions, as well as routing care requests based on gathered data.

Third-Party Publications Using Juji

Misinformation:

- Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li and Haiqi Zhang. Jennifer for COVID-19: An NLP-Powered Chatbot Built for the People and by the People to Combat Misinformation. ACL 2020 Workshop NLP-COVID

Education:

- Yu Chen and Timothy Hill. (2020) Teaching Business Students Chatbots: First Forays and Lessons Learned Americas Conference on Information Systems (AMCIS 2020).
- Sambhav Gupta, Krithika Jagannath, Nitin Aggarwal, Ramamurti Sridar, Shawn Wilde, and Yu Chen (corresponding author). (2019) Artificially Intelligent (AI) Tutors in the Classroom: A Need Assessment Study of Designing Chatbots to Support Student Learning. Pacific Asia Conference on Information Systems (PACIS 2019).
- Terri Lee, Krithika Jagannath, Nitin Aggarwal, Ramamurti Sridar, Shawn Wilde, Timothy Hill, Yu Chen (corresponding author). (2019) Intelligent Career Advisers in Your Pocket? A Need Assessment Study of Chatbots for Student Career Advising. Americas Conference on Information Systems (AMCIS 2019).

Ethical relevant Issue:

- Sarah Theres Völkel, Renate Haeuslschmid, Anna Werner, Heinrich Hussmann, Andreas Butz. (2020) How to Trick AI: Users' Strategies for Protecting Themselves from Automatic Personality Assessment. ACM CHI' 2020.
- Xu Han Am I Asking It Properly?: Designing and Evaluating Interview Chatbots to Improve Elicitation in an Ethical Way. ACM IUI' 2020.

Transformer-based Chatbot

GPT2, DialoGPT

Transformer

Introduction: The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. Transformer, connect the encoder and decoder solely through an attention mechanism, dispensing with recurrence and convolutions entirely. Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Illustrated Guide

Paper: Attention is all you need

Transformer Architecture
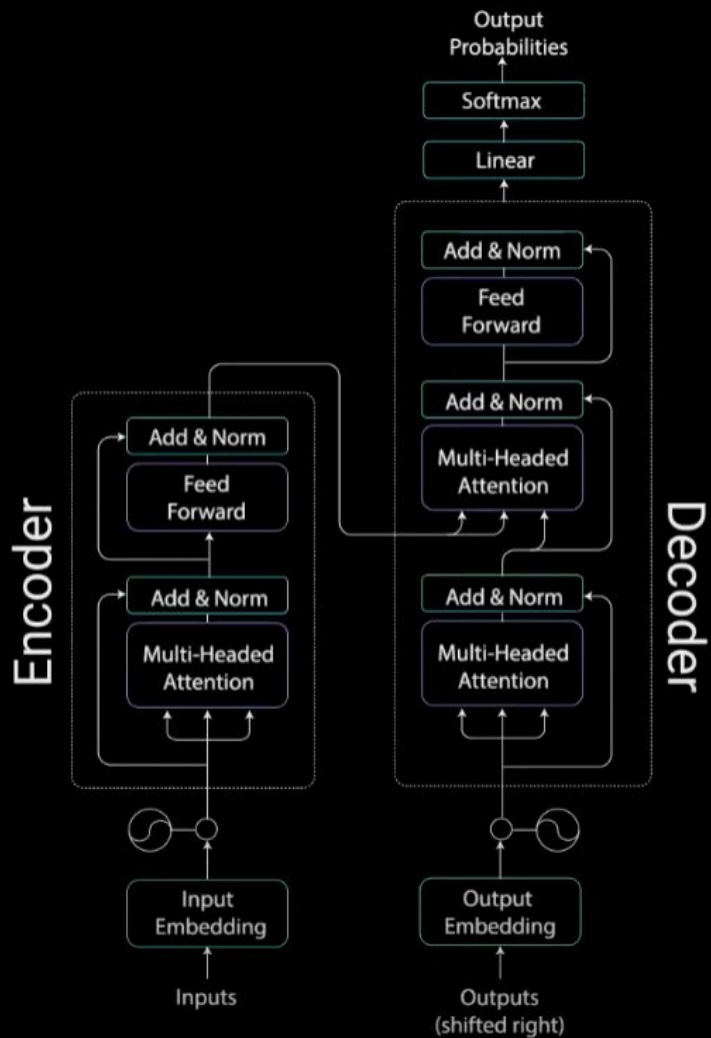
GPT-2

- GPT-2 is a language model using transformer decoder blocks, and is trained on a 40GB dataset called WebText.

- Adopt the generic transformer language model and leverages a stack of masked multi-head self-attention layers.

- GPT2, a 12-to-48 layer transformer with layer normalization and byte pair encoding for the tokenizer.

Pros & Cons

Pros:

- Allow diversity of tasks such as Reading Comprehension, Summarization, Translation and Question Answering
- Be able to perform in a zero-shot setting
- No needs for explicit supervision

Cons:

- **Privacy** – These sorts of models are fundamentally not at all privacy preserving when it comes to the training data. They are explicitly designed to reproduce the data they see. This is a huge problem in many domains (and for us).
- **Compute cost** – A model this big has significant compute costs, even just to run in prediction mode. This can be a real issue for practical deployments of this technology.
- **Bias** – might be an issue for some use cases, especially where there is no human in the loop.
- **It's probably not actually doing sophisticated reasoning**, and might fail in rare or unique cases.

Models/Applications Based on GPT-2

- **DIALOGPT** : a conversational model released by Microsoft, which achieves state-of-the-art performance for generating a relevant and consistent response in a dialog system setting. DIALOGPT is a "tunable gigaword-scale neural network model for generation of conversational responses, trained on Reddit data", which is an extension of GPT-2 specifically for neural response generation.

- **SubSimulatorGPT2**: was created in which a variety of GPT-2 instances trained on different subreddits made posts and replied to each other's comments

- **AI Dungeon:** a free-to-play single-player and multiplayer text adventure game which uses artificial intelligence to generate content. It also allows players to create and share their own custom adventure settings.

Mechanism of DIALOGPT

- Train on the basis of GPT-2.

- Adopt the generic transformer language model and leverages a stack of masked multi-head self-attention layers.

- Inherit from GPT-2, a 12-to-48 layer transformer with layer normalization and byte pair encoding for the tokenizer.

- Model a multi-turn dialogue session as a long text and frame the generation task as language modeling.

Pros & Cons of DIALOGPT

Pros:

- Sentences generated by DIALOGPT are diverse and contain information specific to the source prompt.
- Generate more relevant, contentful and context-consistent responses than strong baseline systems.

Cons:

- Retain the potential to generate output that may trigger offense.
- Output may reflect gender and other historical biases implicit in the data.
- Responses generated using this model may exhibit a propensity to express agreement with propositions that are unethical, biased or offensive (or the reverse, disagreeing with otherwise ethical statements).

Research Paper

Text generation

- Lee, J. S., & Hsiang, J. (2020). Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Information*, *62*, 101983.

Data summarization

- Kieuvongngam, V., Tan, B., & Niu, Y. (2020). Automatic text summarization of covid-19 medical research articles using bert and gpt-2. *arXiv preprint arXiv:2006.01997*.

Question Answering

- Klein, T., & Nabi, M. (2019). Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds. *arXiv preprint arXiv:1911.02365*.

Reading Comprehension

- Cheng, P., & Erk, K. (2020, April). Attending to entities for better text understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 7554-7561).
- Smirnov, D. (2019, September). Neural Network-Based Models with Commonsense Knowledge for Machine Reading Comprehension. In *Proceedings of the Student Research Workshop Associated with RANLP 2019* (pp. 90-94).

Machine Translation

- de Vries, W., & Nissim, M. (2020). As good as new. How to successfully recycle English GPT-2 to make models for other languages. *arXiv preprint arXiv:2012.05628*.

Summary

	Prons	Cons	Scenario
Rasa	- Sophisticated NLU engine - Open source community - Protect process sensitive consumer data	- Rasa isn't necessarily for beginners - Computational consuming(spaCy)	Professional and integrate with ML
Juji	- Easy, straightforward, simple UI - Automatic chat flow management saves time and energy.	- Resources is little - Option is few	Easy to operate, simple scenario that does not require large knowledge base
GPT-2	- is able to perform in a zero-shot setting - learn how to perform a surprising amount of tasks without the need for explicit supervision	not understanding real-world sensibility and coherence	- Text generation - Data summarization - Question Answering - Reading Comprehension - Machine Translation