

Jan Šedivý - ČVUT, Praha



# Conversational AI

# Why is Conversational AI Essential?

We design and implement captivating and enjoyable conversational assistants empowered by LLM tools.

These assistants are designed to:

- decipher intricate queries,
  - harness extensive databases,
  - comprehend user behavior,
  - provide personalized interactions.
- 
- Directed dialogue
  - Open dialogue
  - Text, voice
  - Multimodal

- **Transactional Dialogues:** making a reservation, placing an order, or scheduling an appointment.
- **QA Dialogues:** facts, explanations, or answers to their questions.
- **Social Dialogues:** engage, entertain users
- **Guided Dialogues:** Step by steps execution: SOP forms, surveys.
- **Educational Dialogues:** Teach users about specific topics
- **Game or Entertainment Dialogues:** Enjoyable and engaging.
- **Onboarding Dialogues:** Guide new users through the features and functionalities.
- **Therapeutic Dialogues:** Providing emotional support, counseling, assistance.
- **Personal Assistant Dialogues:** Siri or Google Assistant, reminders, sending messages, or providing directions.



# Social dialogue

## Alquist

is a Voice First social bot. It carries an engaging and entertaining dialog about popular topics, such as sports, celebrities, movies, politics ...

Amazon Alexa Prize - since 2016  
200+ academic teams

Running on Echo type devices, this year also device with screen

Top prize 20 min long dialogue = 1M USD

You can try it: Alexa Let's Chat

## Hybrid dialogue architecture

- rule-based
- generative

Text generation (30%) - Vicuna LLM - 13B params

40 topics

Access to Internet

Unknown topic simple Wikipedia answer

Sub-dialogues - reusability

Short/long-term dialogue memory

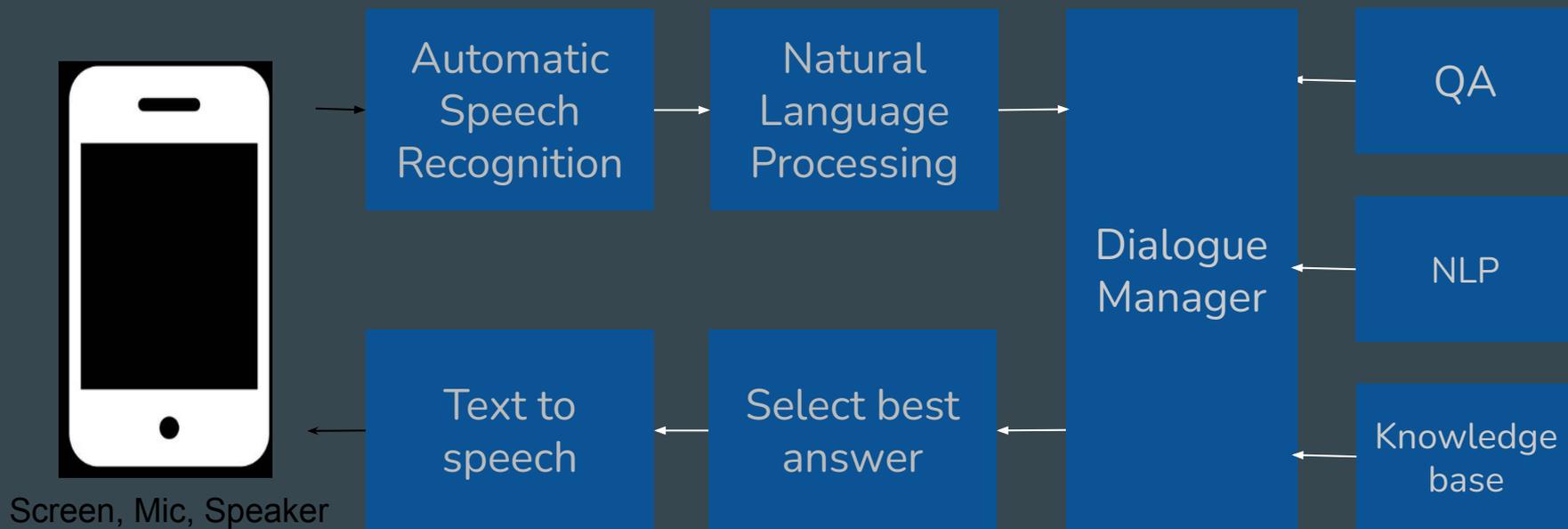
Context-based content (DBs, APIs)

User profile





# Conversational AI system diagram



# Chatbot - the simplest application

The bot asks question

Users control the dialogue by pushing buttons

The application traverses a simple tree structure dialogue

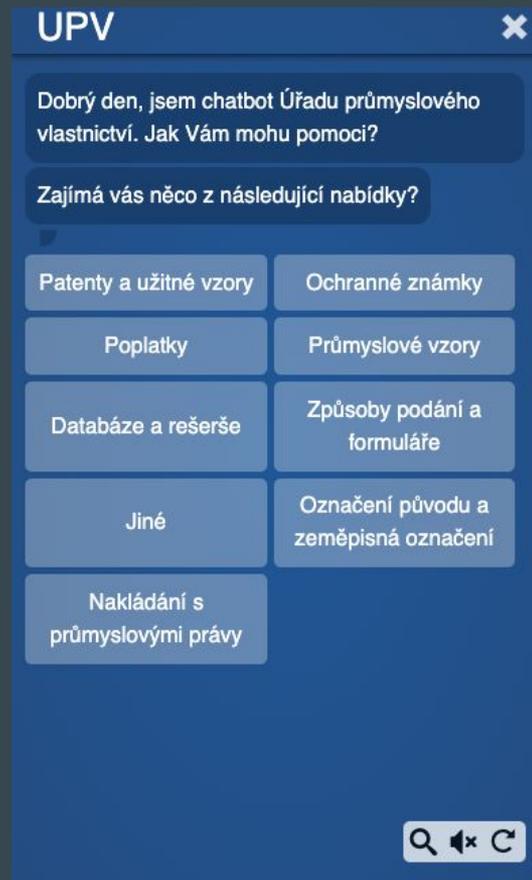
Answers are mostly only Yes/No

Output - text, voice, picture

Easy development - simple multi step workflows

[T-Mobile self repair application](#)

[Guided tour and QA](#)



# Question answering

## Manually prepared Semantic Text Similarity (STS)

- **Step 1:** Create answers
- **Step 2:** Create alternative question
- Find the semantically closest question
- Good for tens to hundreds question
- Low CPU requirements, fast,
- No training required
- Easy updating
- Difficult to recognize out of domain
- Short questions and answers preferred
- Global control commands

### Step 1

Na webových stránkách Úřadu naleznete informace týkající se přihlašování do zahraničí, a to vždy v rámci sekce vřívované jednotlivým průmyslovým právem, patentem, užitným vzorom, průmyslovým vzorom, ochranným známám a označením původu a zeměpisným označením.

Na webových stránkách Úřadu naleznete informace týkající se přihlašování do zahraničí, a to vždy v rámci sekce vřívované jednotlivým průmyslovým právem, patentem, užitným vzorom, průmyslovým vzorom, ochranným známám a označením původu a zeměpisným označením.

Na webových stránkách Úřadu naleznete informace týkající se přihlašování do zahraničí, a to vždy v rámci sekce vřívované jednotlivým průmyslovým právem, patentem, užitným vzorom, průmyslovým vzorom, ochranným známám a označením původu a zeměpisným označením.

Na webových stránkách Úřadu naleznete informace týkající se přihlašování do zahraničí, a to vždy v rámci sekce vřívované jednotlivým průmyslovým právem, patentem, užitným vzorom, průmyslovým vzorom, ochranným známám a označením původu a zeměpisným označením.

Na webových stránkách Úřadu naleznete informace týkající se přihlašování do zahraničí, a to vždy v rámci sekce vřívované jednotlivým průmyslovým právem, patentem, užitným vzorom, průmyslovým vzorom, ochranným známám a označením původu a zeměpisným označením.

### Step 2

Jak přihlásit průmyslový vzor do zahraničí?

Jak zaregistruji průmyslový vzor do zahraničí?

Jak mám zapsat zahraniční průmyslový?

Jak se postupuje při zahraniční registraci průmyslového vzoru?

Jak se zaregistrují zahraniční průmyslové vzory?

Jak přihlásit průmyslový vzor do zahraničí?

Jak zaregistruji průmyslový vzor do zahraničí?

Jak mám zapsat zahraniční průmyslový?

Jak se postupuje při zahraniční registraci průmyslového vzoru?

Jak se zaregistrují zahraniční průmyslové vzory?

Jak přihlásit průmyslový vzor do zahraničí?

Jak zaregistruji průmyslový vzor do zahraničí?

Jak mám zapsat zahraniční průmyslový?

Jak se postupuje při zahraniční registraci průmyslového vzoru?

Jak se zaregistrují zahraniční průmyslové vzory?

Jak přihlásit průmyslový vzor do zahraničí?

Jak zaregistruji průmyslový vzor do zahraničí?

Jak mám zapsat zahraniční průmyslový?

Jak se postupuje při zahraniční registraci průmyslového vzoru?

Jak se zaregistrují zahraniční průmyslové vzory?



Jak přihlásit průmyslový vzor do zahraničí?

Jak zaregistruji průmyslový vzor do zahraničí?

Jak mám zapsat zahraniční průmyslový?

Jak se postupuje při zahraniční registraci průmyslového vzoru?

Jak se zaregistrují zahraniční průmyslové vzory?



# Bot technologie

The simplest bot uses buttons to let the user reply



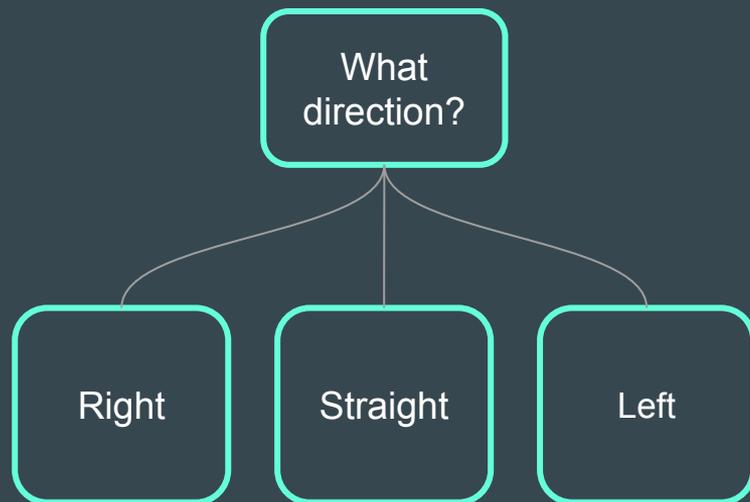
## Guided Tour

Simple walk through web pages with links to certain details



Intent, entity

Semantic similarity to select the closest intent



# Large Language Models - chatbots

**LLMs** are pre-trained on vast amounts of text data. Here's how LLM technology is used to create chatbots:

**Pre-training:** learn grammar, facts, reasoning abilities, and some common-sense knowledge.

**Fine-tuning:** augment for specific tasks, training on a narrower dataset that includes dialogues and conversations.

**Context-awareness:** consider the entire conversation history.

**User input processing:** Chatbots process inputs, extract relevant information, and understand the intent behind the message.

**Natural language generation:** LLM create human-like responses.

In summary, LLM technology forms the core of creating chatbots capable of natural language understanding and generation. These chatbots can be used in a wide range of applications, from customer support to virtual assistants, providing users with interactive and engaging conversational experiences.



# Context and Prompts

LLM, prompting is a technique used to **interact** with and **instruct** these models. It involves providing a model with a text prompt and/or **context**, which serves as a specific request or question.

The art of prompting lies in crafting clear and contextually relevant input to elicit desired responses from the model.

Context in vector base provides **trusted material**. (Fine-tuning, RAG)



# Step by step versus demonstrations

Series of intermediate reasoning steps.

Expensive manual design: task-specific demonstrations one by one.

Automatic Chain-of-Thought (Auto-CoT) [Zhang et al. \(2022\)](#)

1. question clustering: partition questions of a given dataset into a few clusters
2. demonstration sampling: select a representative question from each cluster and generate its reasoning chain using Zero-Shot-CoT with simple heuristics

<https://arxiv.org/abs/2201.11903>



# Chain-of-Thought (CoT) Prompting

Chain-of-thought (CoT) prompting enables complex “reasoning” capabilities through intermediate reasoning steps

Combine with few-shot prompting to get better results on more complex tasks that require reasoning before responding.

<https://www.promptingguide.ai/techniques/cot>

[Wei et al. \(2022\)](#)



- Let's start by using algebra to represent the given information.
- Let  $x$  be your age.
- Then your daughter's age is 30 years younger than you, which means her age is  $x - 30$ .
- Your daughter's sister is two years older than her, so her age is  $(x - 30) + 2$ , which simplifies to  $x - 28$ .
- Your daughter's sister is also 25 years younger than your wife, so we can set up another equation:
  - $x - 28 = \text{wife's age} - 25$
  - Simplifying this equation, we get:
  - $\text{wife's age} = x - 28 + 25$
  - $\text{wife's age} = x - 3$
  - So your wife's age is  $50 - 3 = 47$  years old.

I am fifty years old. My daughter is thirty years younger than me. Her sister is two years older and twenty-five years younger than my wife. How old is my wife?



## Standard Prompting

### Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. ❌

## Chain-of-Thought Prompting

### Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

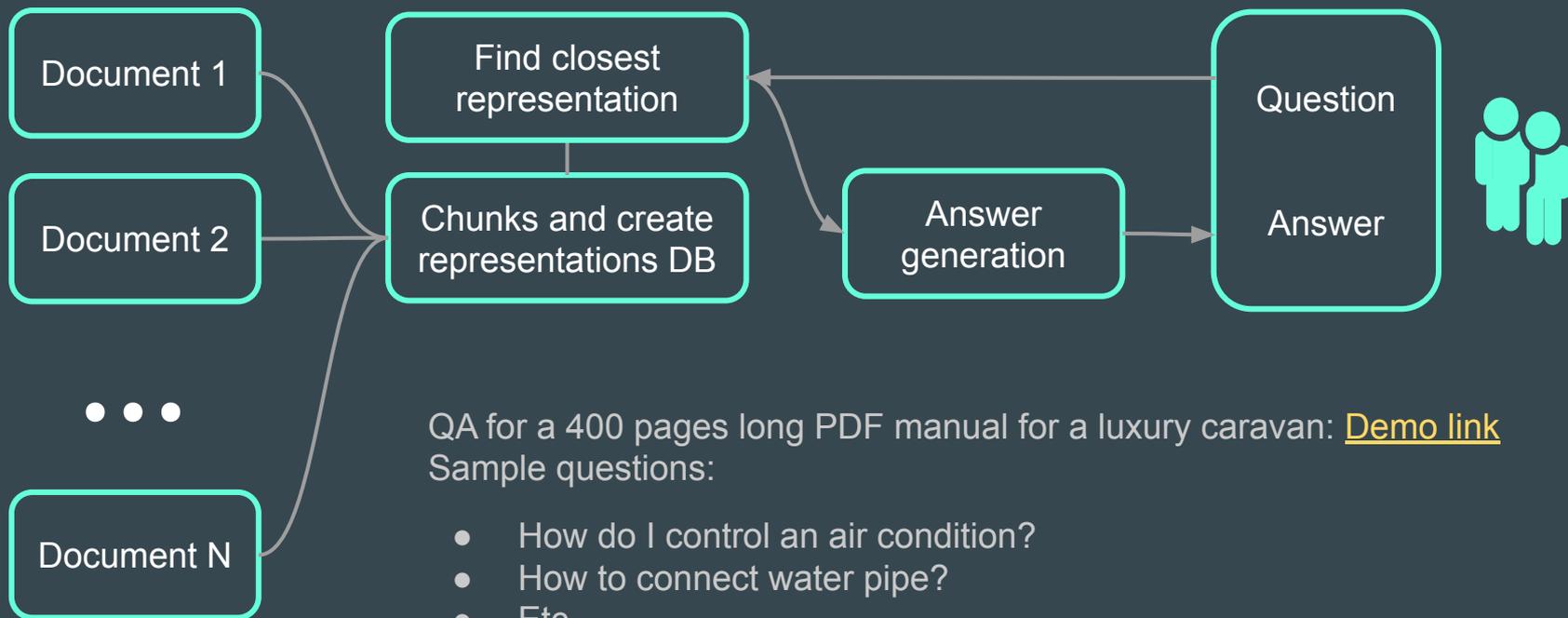
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✅



# Retrieval Augmented Generation (RAG)



QA for a 400 pages long PDF manual for a luxury caravan: [Demo link](#)

Sample questions:

- How do I control an air condition?
- How to connect water pipe?
- Etc.
- RAG adaptive for situations where facts could evolve over time





anything.

# THOR MOTOR COACH

Made to fit.

Welcome to the Thor Factory-built motor coach interactive manual.

I'm here to help you. Feel free to ask me anything.

You should connect the shore cord to a properly grounded power source.

Make sure to check the voltage on each of the lines to ensure it is safe to use.

Is there anything else I can help you with?

You should connect the shore cord to a properly grounded power source.

cord to a properly grounded power source.

Make sure to check the voltage on each of the lines to ensure it is safe to use.

Make sure to check the voltage on each of the lines to ensure it is safe to use.



# Calling functions - plugins

Fine-tuned to both detect when a function needs to be called (depending on the user's input) and to respond with JSON

“What’s the weather like in Boston?” to `get_current_weather(location: string, unit: 'celsius' | 'fahrenheit')`.

“How many orders did Acme, Inc. place last month?” to a SQL query using `sql_query(query: string)`

[OpenAI Function calling example](#)





The credit card company records all transactions and then I can ask questions about when was the last time I did this and that.



# Hybrid models

**Context Management:** Maintain prompts and context that includes information about the topic

**Explicit User Instructions:** Instruct the user explicitly to stay on the chosen topic. Rule-based

**Keyword Monitoring:** NER to prompt

**Topic Steering Responses:** Suggestion questions. Periodically remind the user about the chosen topic.

**Out of Domain (OOD) responses:** Try to get back to discussing [chosen topic]."

**Feedback Mechanism:** Implement a feedback mechanism where the user can express their meaning.

**User Assistance:** Help, offer assistance

**OOD Escalation:** "I understand your questions about [off-topic], but let's explore [chosen topic] more thoroughly. "

**Hybrid models:** LLM combined with curated rule based system bots can handle a variety of user inputs, and striking the right balance between steering and flexibility.



# Well being apps

Persona represents certain character type

People find it easier to open up to a Digital Persona than to a human

- Teens do not open up to parents as they expect judgment.
- Partners fear to disappoint each other with full honesty.
- Employees feel pressure to perform.
- Learn how to negotiate

Android, iPhone, Elysai.ai





**HI JOHN,  
NICE TO SEE  
YOU AGAIN, HOW MAY  
I ASSIST YOU TODAY?**



# Generative Pre-trained Transformer 3

- 2017 12 Jun - **Attention Is All You Need**, Google
- GPT-3: 175B params,, 96 attention layers, 800GB, 2048 context
- **OpenAI**
  - Incorporated - **non-profit**
  - OpenAI Limited Partnership - **for-profit** subsidiary corporation
- 2019 Microsoft invests to OpenAI with \$1 billion,
- 2020 API is public, but only Microsoft has access to GPT-3's underlying model.
- 2020 Sep, Microsoft had licensed "exclusive" use of GPT-3;
- 2023 Jan, Microsoft invests to OpenAI with \$10 billions,
- 2023 Sep, Microsoft is building AI capabilities into nearly all products.

1967 All you need is love



# Conversation problems

- Hallucinations
- Responses are frequently too terse. Should be engaging.
- Repetitions: i.e., the chatbot sometimes replies with semantically the same utterances
- Conversation is not deep enough, resulting in shallow conversations.
- Irrelevant outputs, disrupting the conversation's consistency.
- Contradictions and failure to comprehend user input.
- Rapidly changing subjects sounds unnatural
- Potentially irritating.
- Long latency



# Development and runtime costs

How to scale up # params and # data for particular training compute budget?

What is the inference budget?

How much data can we get from Czech publicly available datasets exclusively?

A smaller one trained longer will ultimately be cheaper at inference



I find it disturbing that certain individuals are only highlighting the challenges and risks associated with AI. However, I remain optimistic that we can maintain the positive momentum in creating more effective AI solutions and alternatives, ultimately contributing to improving people's quality of life.



**Thank you**

