



# Conversational Search

ESSIR 2022

Dr Jeff Dalton

Associate Professor, University of Glasgow

Turing AI Acceleration Fellow

 @jeffd

jeff.dalton@glasgow.ac.uk

# About me



## Jeff Dalton

Conversational Search and AI researcher, with experience at Google, Twitter, and UMass Amherst.

## My Background

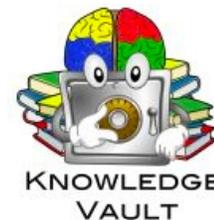
3 years in Google Research

Knowledge Discovery team

Google Health Search team

'Executive MD' from Harvard Medical School

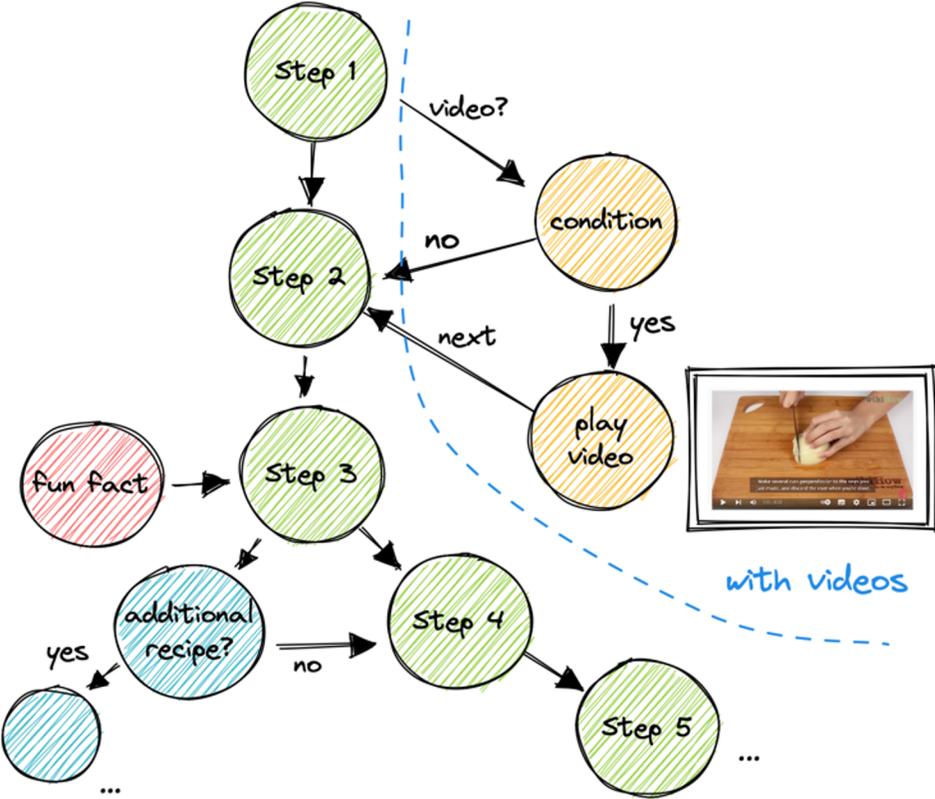
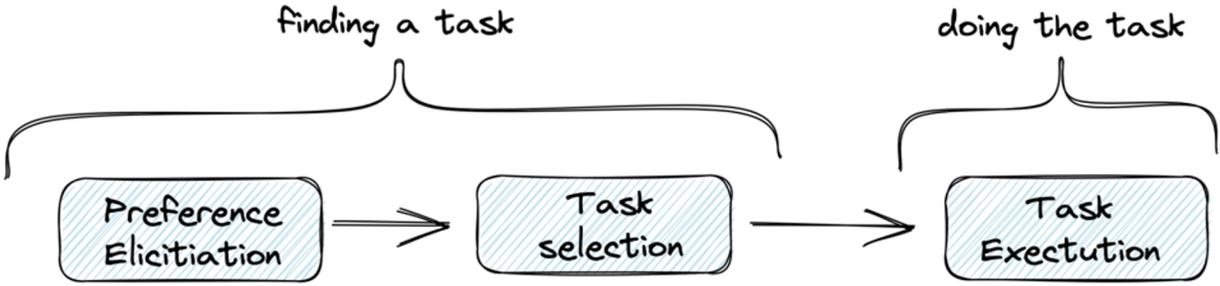
Google Assistant Natural Language Understanding



# Optional Lab: Get start with Open Assistant Toolkit (OAT)

**What?** *A domain-agnostic framework for conversational system research*

**Who?** *Allows researchers to easily develop and evaluate conversational systems*



OAT task walkthrough

# UoG GRILLBot

## Winner of 2021-22 Alexa Prize Taskbot Challenge

[https://grilllab.ai/2022-05-15-alexa\\_taskbot\\_win/](https://grilllab.ai/2022-05-15-alexa_taskbot_win/)



alexa

# Production-ready conversational agent toolkit

## Prepare the OAT(s)

---

### OAT Setup

- <https://bit.ly/prepare-oat>



### Hands-on OAT(s)

- <https://bit.ly/interactive-oats>



# Outline

---

Motivation and History

---

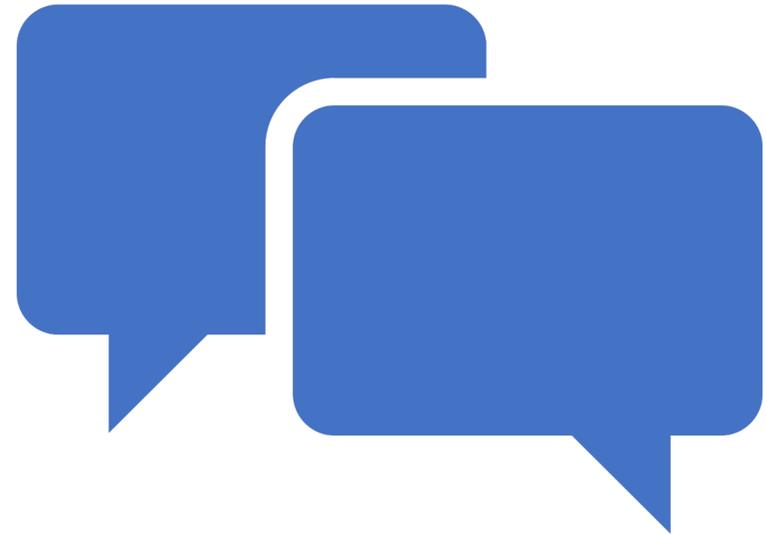
What is 'Conversational Search'?

---

Conversational Language  
Understanding & Ranking

---

Future applications: Multimodal  
Taskbots



**NEW**

# Conversational Information Seeking Book

<https://arxiv.org/abs/2201.08808>

SIGIR 2022 Tutorial

Conversational Information Seeking

<https://bit.ly/the-cis-tutorial>

## Conversational Information Seeking

**An Introduction to Conversational Search,  
Recommendation, and Question Answering**

---

**Suggested Citation:** Hamed Zamani, Johanne R. Trippas, Jeff Dalton and Filip Radlinski (2022), "Conversational Information Seeking", : Vol. xx, No. xx, pp 1–194. DOI: 10.1561/XXXXXXXXXX.

**Hamed Zamani**

University of Massachusetts Amherst  
zamani@cs.umass.edu

**Johanne R. Trippas**

University of Melbourne  
johanne.trippas@unimelb.edu.au

**Jeff Dalton**

University of Glasgow  
jeff.dalton@glasgow.ac.uk

**Filip Radlinski**

Google Research  
filiprad@google.com

This article may be used only for the purpose of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval.

**now**  
the essence of knowledge  
Boston — Delft

# Motivation



[What is the best time for wildflowers in the bay area]

### Where to See Wildflowers in the Bay Area - Bay Area Parent

[www.bayareaparent.com/Article/Where-to-See-Wildflowers-in-the-Bay-Area/](http://www.bayareaparent.com/Article/Where-to-See-Wildflowers-in-the-Bay-Area/)

This is a great time of year to find wildflowers in the Bay Area. Call parks ... Loop Trail, known for being flat, moderate in difficulty and best for viewing the blooms.

### Best Bay Area hikes into wildflower country - Bay Nature

[baynature.org/article/best-bay-area-hikes-into-wildflower-country/](http://baynature.org/article/best-bay-area-hikes-into-wildflower-country/)

Mar 28, 2013 - It's spring, and all you nature lovers know what that means ... time to go see wildflowers! You've probably already noticed them popping in little ...

### Bay Area Wildflower Hikes | EveryTrail

[www.everytrail.com/guide/bay-area-wildflower-hikes](http://www.everytrail.com/guide/bay-area-wildflower-hikes)

★★★★★ Rating: 5 - 2 votes

Download the Bay Area Wildflower Hikes mobile travel guide to your iPhone or ... Even in lean years.

# The rise of mobile search

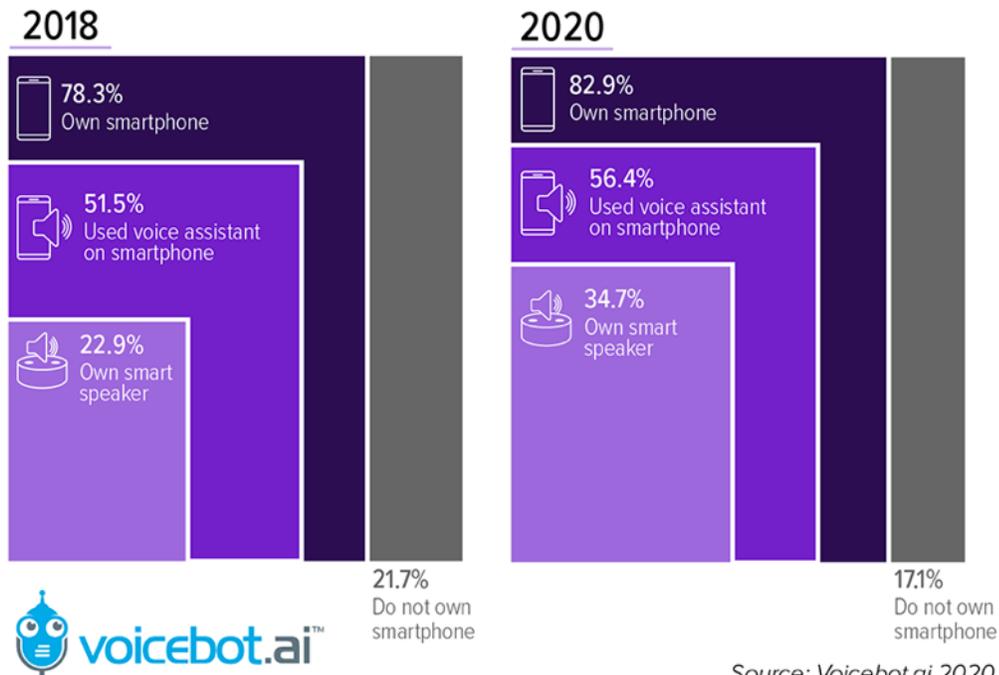
---

Majority of searches are now performed on mobile phones.

# Rise of voice 'assistants'

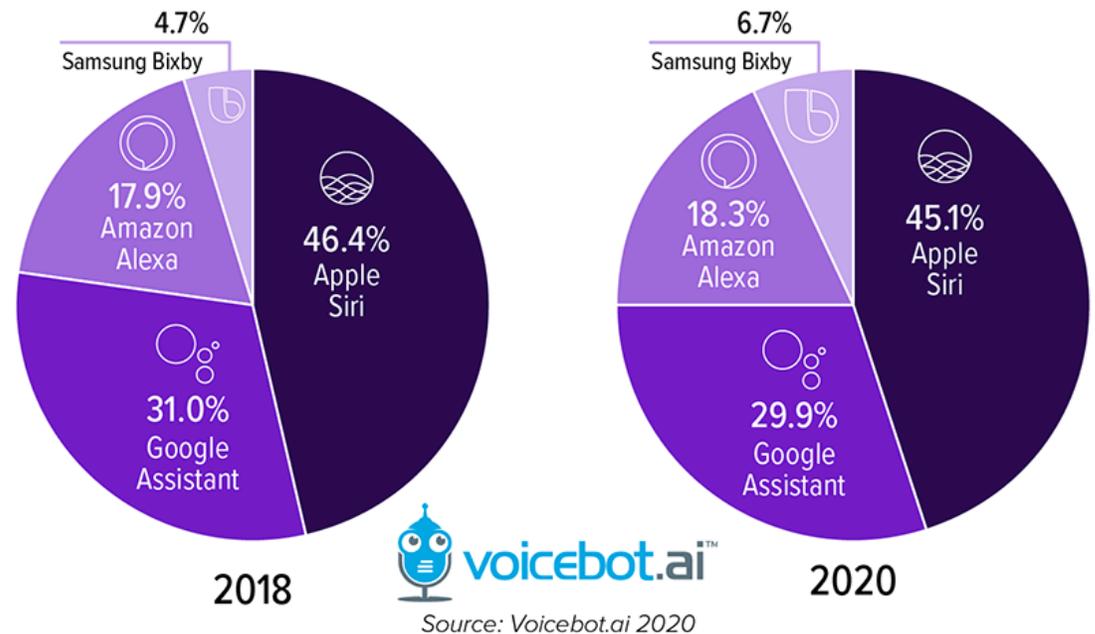
- Growing adoption of conversational voice assistants.
- Speech is easier and more natural than typing
- **Voice search is mobile**
  - Approximately 1/3 of mobile queries

Smartphone Ownership and Voice Assistant Use

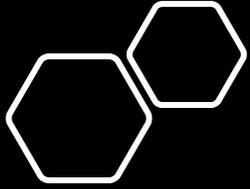


Source: Voicebot.ai 2020

U.S. Smartphone Voice Assistant Relative Market Share



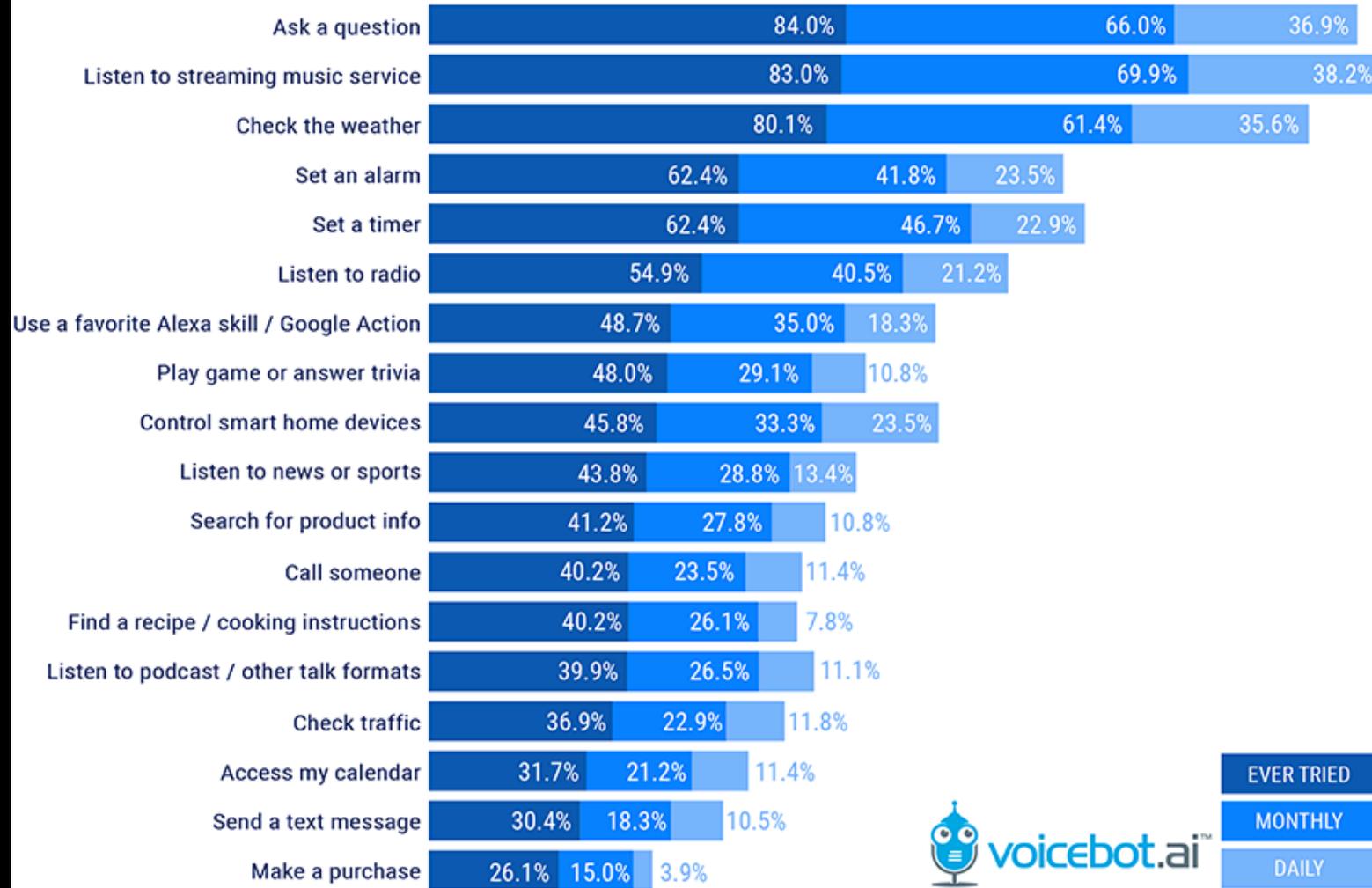
Source: Voicebot.ai 2020



# What we do with assistants today

- Top Alexa skills are single turn
- Many basic transactional and navigational tasks
- ... much more is possible

Smart Speaker Use Case Frequency - January 2019



Source: Voicebot Smart Speaker Consumer Adoption Report Jan 2019

# Why Conversational Search?



Advances in automatic speech recognition



Devices with limited bandwidth interfaces

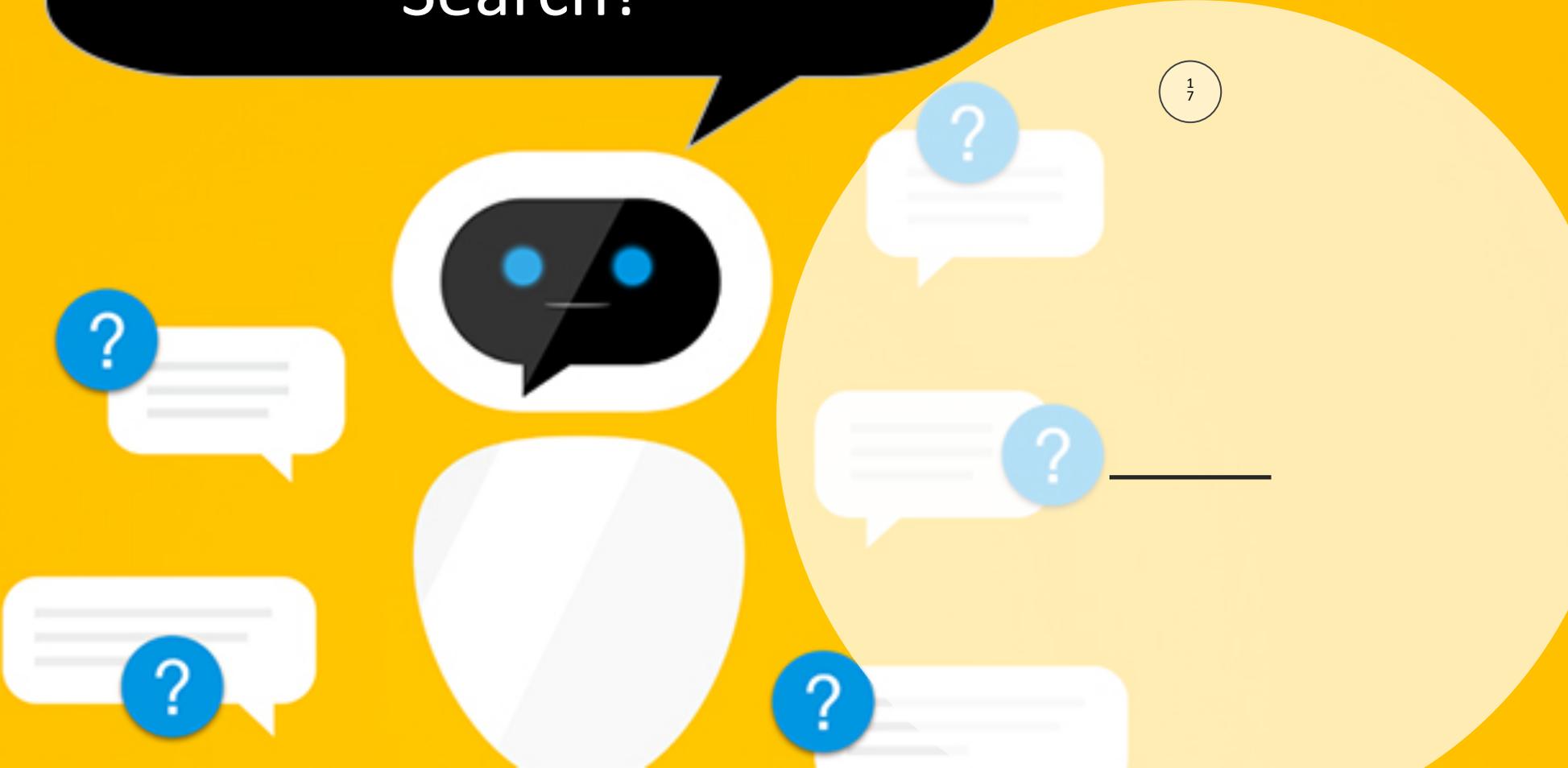


Recent advances in neural models

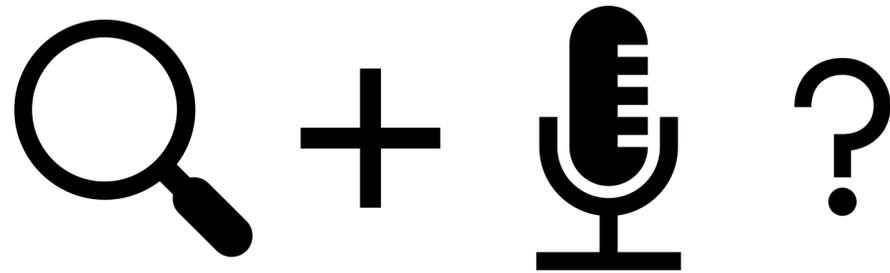


Information access is an important service

# What is Conversational Search?



# What is conversational search?



# Two types of dialog systems?

## Chit-chat (chatbot)

- No specific goal
- Usually 'ungrounded' system

Mitsuku, how did you become a good conversationalist?

Yep, Do you really think I am a good conversationalist?  
Mousebreaker programmed me to be a good conversationalist.

How are you programmed?

I read what you say and then I compose the best reply I can think of.

How does that actually work though?

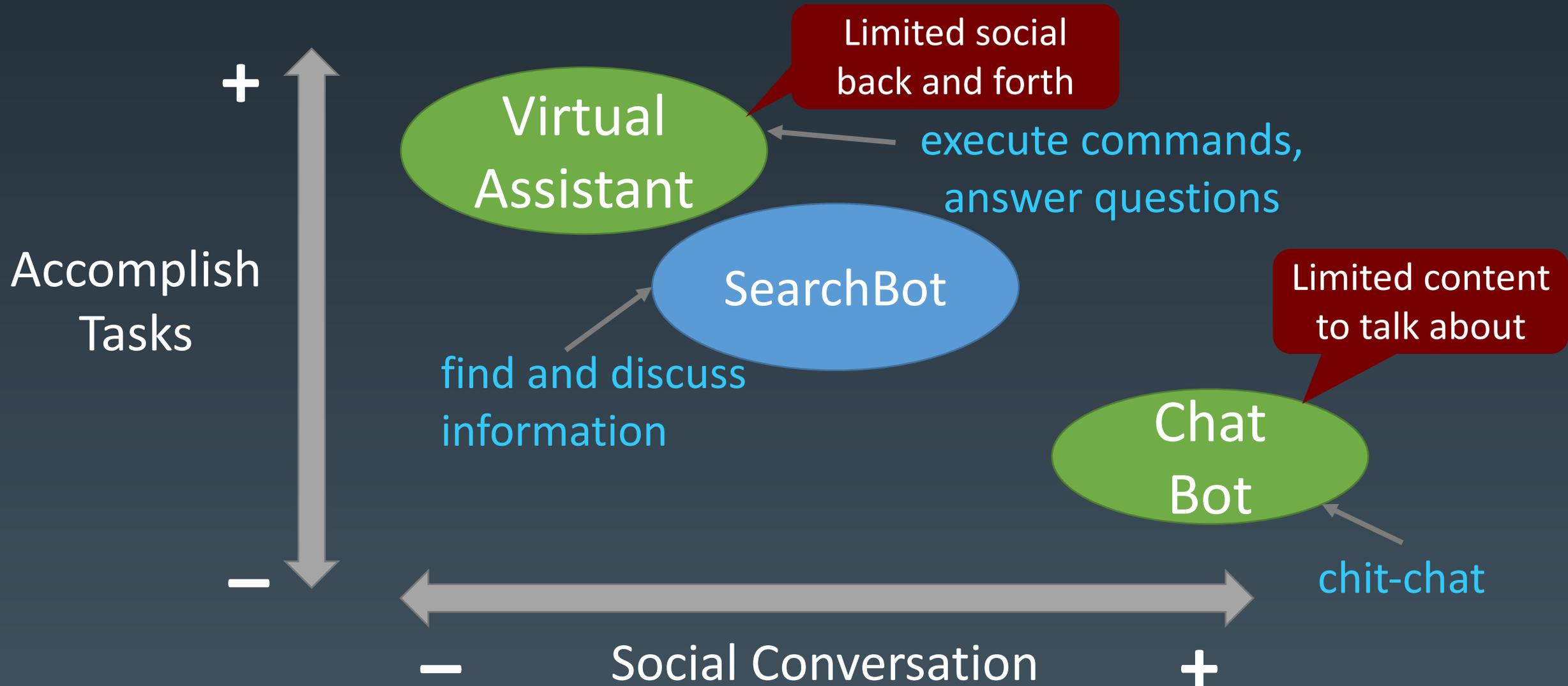
Oh. I think it might be a trade secret.

## Task-oriented

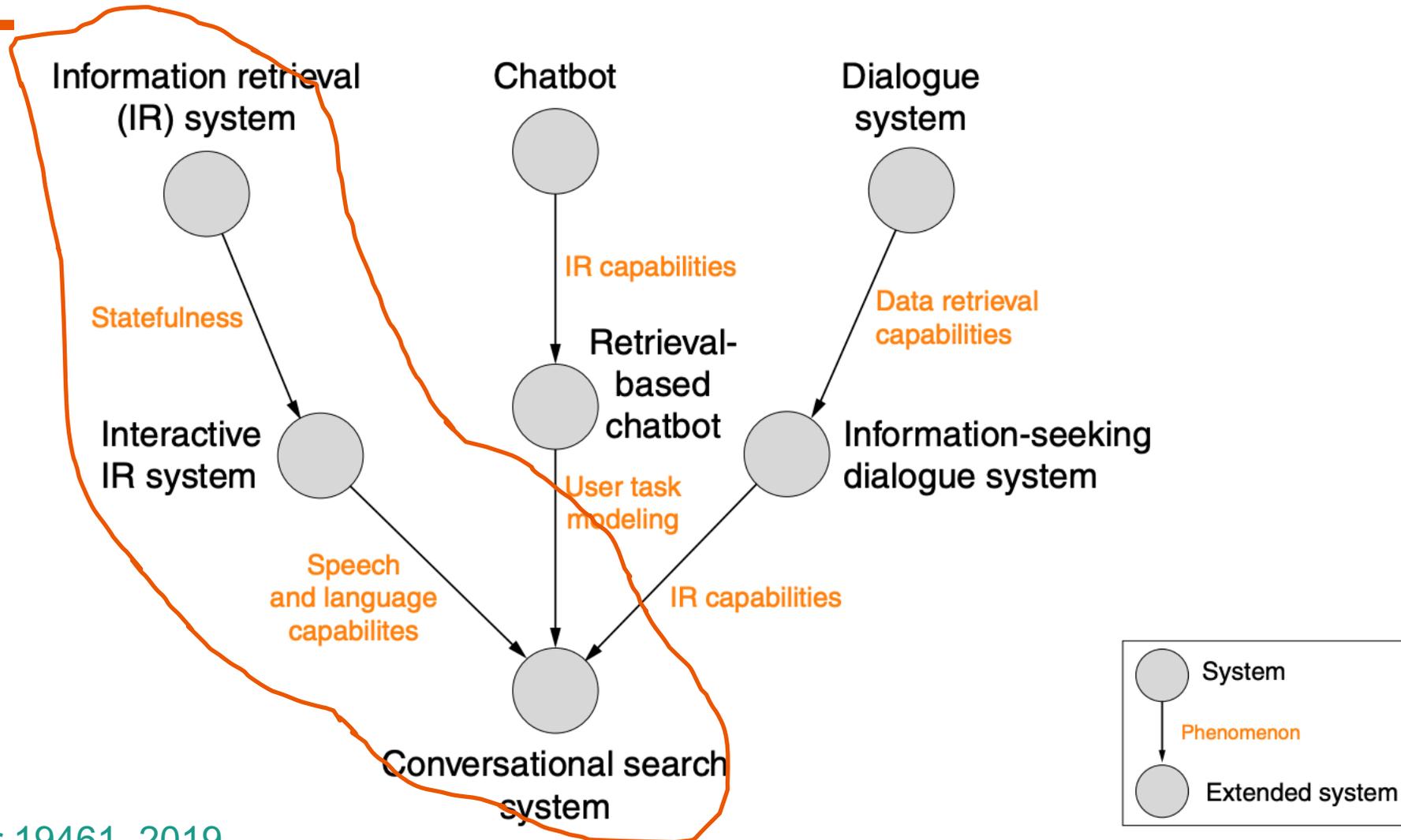
- Focus on actions
- Accomplish a virtual task
  - SIRI, interfaces to cars, robots
  - Ex: booking flights or restaurants



# Types of Conversational Systems



# Taxonomy of conversational search system



# Definition

## Conversational Information Seeking System

A *Conversational Information Seeking (CIS)* system is a system that satisfies the information needs of one or more users by engaging in **information seeking conversations**.

CIS responses are expected to be **concise**, **fluent**, **stateful**, **mixed-initiative**, **context-aware**, and **personalized**.

# From web to conversational search...

## Ad hoc Search



## Conversational Search



Keyword-based Queries



Natural Queries

### Necessity:

- Speech/Mobile Interfaces

### Opportunities:

- More natural and explicit expression of information needs

### Challenge:

- Query understanding & sparse retrieval

# From web to conversational search...

## Ad hoc Search

Startup Investing. Simplified. - SeedInvest

<https://www.seedinvest.com>

Join 300,000+ people who already use SeedInvest to find startup investment opportunities. SeedInvest is filled with investors and entrepreneurs that are passionate about building future innovation. They are accomplished individuals that invest ...

Log In

SeedInvest is a leading equity crowdfunding platform that provides individual investors ...

Browse Offerings

The following offerings are being conducted ...  
Monogram Orthopaedics · Winc · Auto Invest

See more

Ten Blue-Links



## Conversational Search

A strategy used by seed institutional investors is the **spray and pray** type of model in which investment funds are invested in a number of companies and see which ones pick up traction. Once the start-ups they are taking on are identified then you allocate additional capital to invest in follow on rounds of financing.



How Funding Rounds Work For Startups - Forbes

[www.forbes.com/sites/alejandrocremades/2018/12/26/how-funding-rounds-wor...](http://www.forbes.com/sites/alejandrocremades/2018/12/26/how-funding-rounds-wor...)

Natural Responses

## Necessity:

- Speech/Mobile Interfaces

## Opportunities:

- Direct & easier access to information

## Challenge:

- Document understanding; combine and synthesize information

# From web to conversational search...

## Ad hoc Search



Single-Shot Query

## Conversational Search



Multi-Turn Dialog

### Necessity:

- N.A.

### Opportunities:

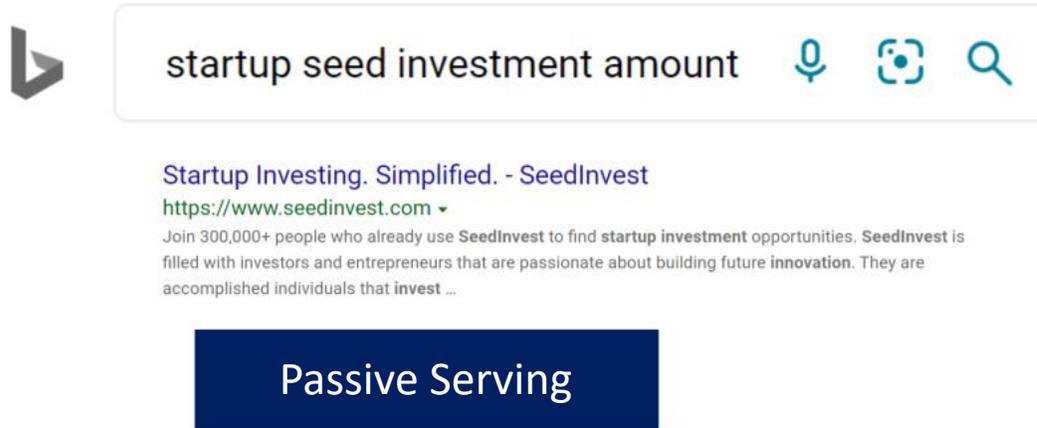
- Serving complex information needs and tasks

### Challenge:

- Contextual understanding & reasoning

# From web to conversational search...

## Ad hoc Search



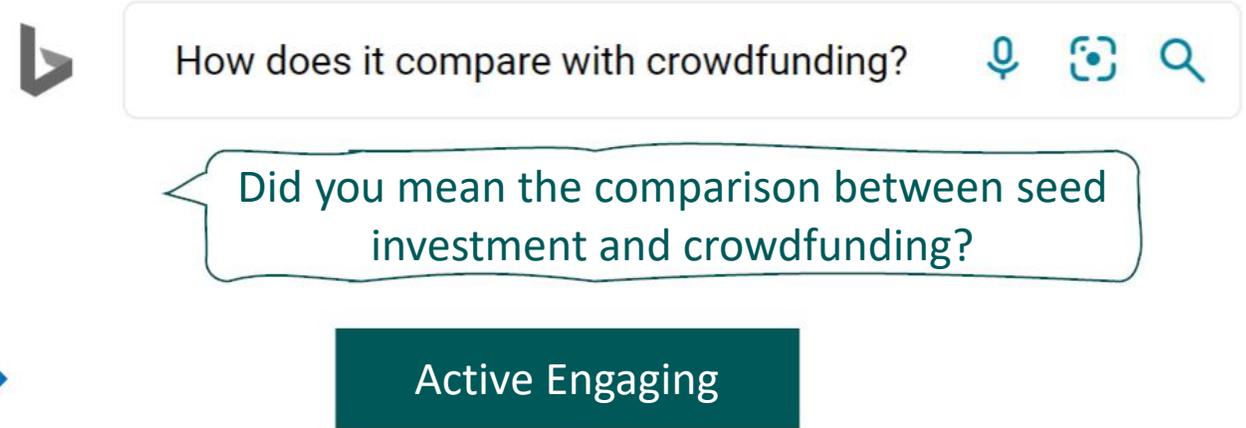
startup seed investment amount

Startup Investing. Simplified. - SeedInvest  
<https://www.seedinvest.com>  
Join 300,000+ people who already use SeedInvest to find startup investment opportunities. SeedInvest is filled with investors and entrepreneurs that are passionate about building future innovation. They are accomplished individuals that invest ...

Passive Serving



## Conversational Search



How does it compare with crowdfunding?

Did you mean the comparison between seed investment and crowdfunding?

Active Engaging

### Necessity:

- N.A.

### Opportunities:

- Collaborative information seeking & better task assistance

### Challenge:

- Dialog management, less lenient user experience

# RRIMS properties

[Radlinski & Craswell 17]

- 1. User Revealment**  
Helps user express (potentially discover) their true information need and long-term preferences
- 2. System Revealment**  
Reveals to the user its capabilities and corpus, set expectations of what it can and cannot do.
- 3. Mixed-Initiative**  
“A flexible interaction strategy in which each agent (human or computer) contributes what it is best suited at the most appropriate time.”
- 4. Memory**  
The ability to reference previous information (state)
- 5. Set Retrieval**  
Reason about utility of complementary items

And optionally  
many others...

- **Recovering from communication breakdowns** - Clarification can be simply in the form of “asking for repeat” or more advanced forms.
- **Representation generation** - generate new (and useful) representations that are shared between a user and system.
- **Multimodality** – may involve multiple modalities in terms of input
- **Speech** - may involve speech-based input and output
- **Analyzing conversations** - analyze human-human conversations (synchronously or asynchronously) and intervene to provide contextually relevant information.
- **Personalized understanding and reasoning about user** - infer more information about a specific user e.g., cognitive abilities and styles, domain knowledge
- **Collaborative** – involve a group of users rather than a single individual
- **Role** – may take on different roles, e.g. teacher, scribe, mediator



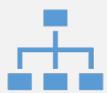
1. **Smarter Models** – “Let me search for that...”



2. **Interactive** – “No, that’s wrong, I meant...”



3. **Contextual** – “I think you’re trying to...”



4. **Structure & Reasoning** – “Here’s a summary I generated...”



5. **Human-centric Evaluation** – “No, stop. I’m giving up...”

# Five Challenges in Conversational Search



# History

# Precision vs. robustness



SHRDLU



CHAT-80

Precise, complete understanding



Brittle, narrow coverage

Robust, broad coverage

Fuzzy, partial understanding



# CHAT-80 (Pereira 1980)

Is there more than one country in each continent?

No.

What are the countries from which a river flows into the Black\_Sea?

[romania].

What is the total area of countries south of the Equator and not in Australasia?

10239 ksq miles.

Which country bordering the Mediterranean borders a country that is bordered by a country whose population exceeds the population of India?

turkey.

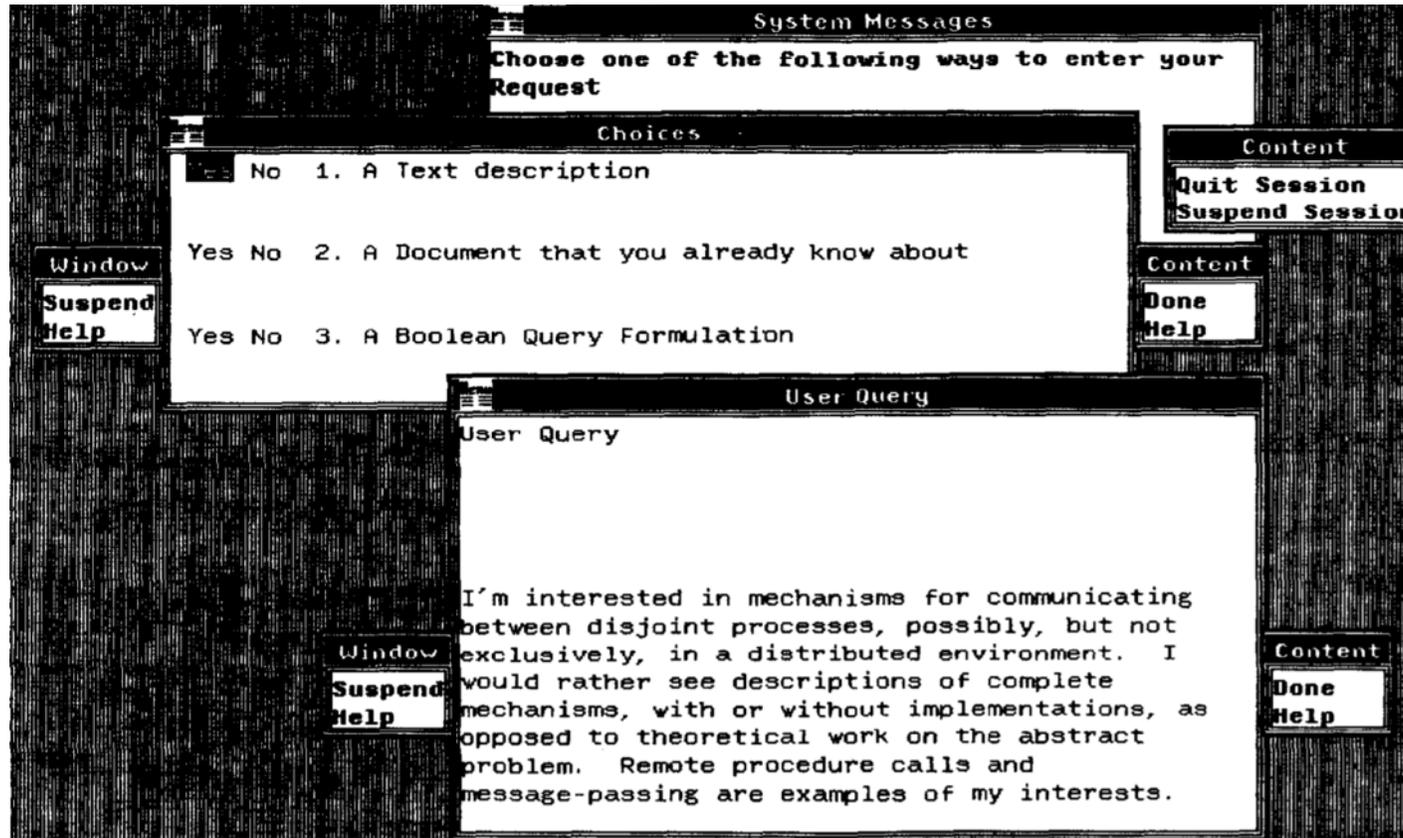
What countries border Denmark?

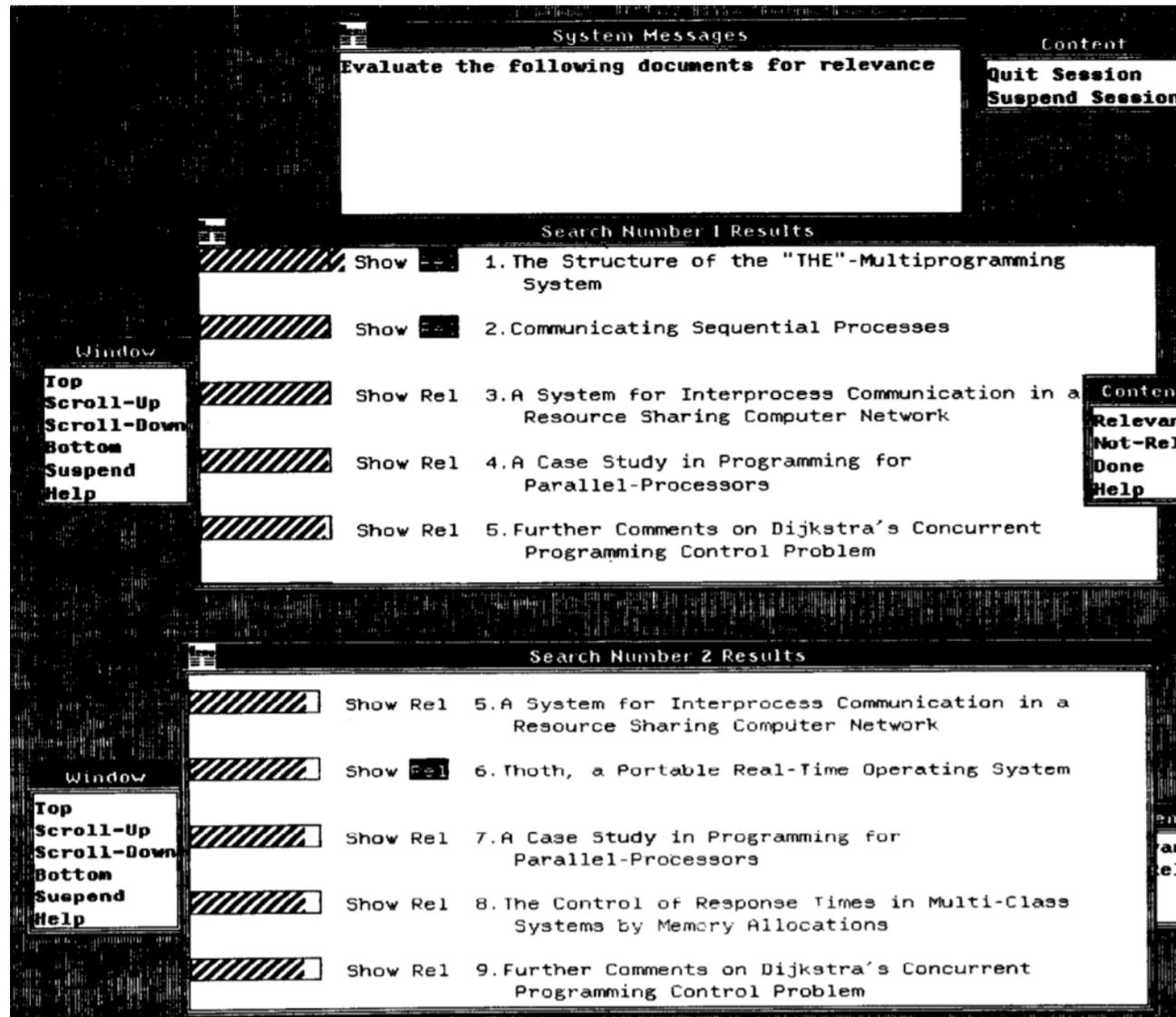
west\_germany.

What countries is Denmark adjacent to?

I don't understand!

# The I<sup>3</sup>R System





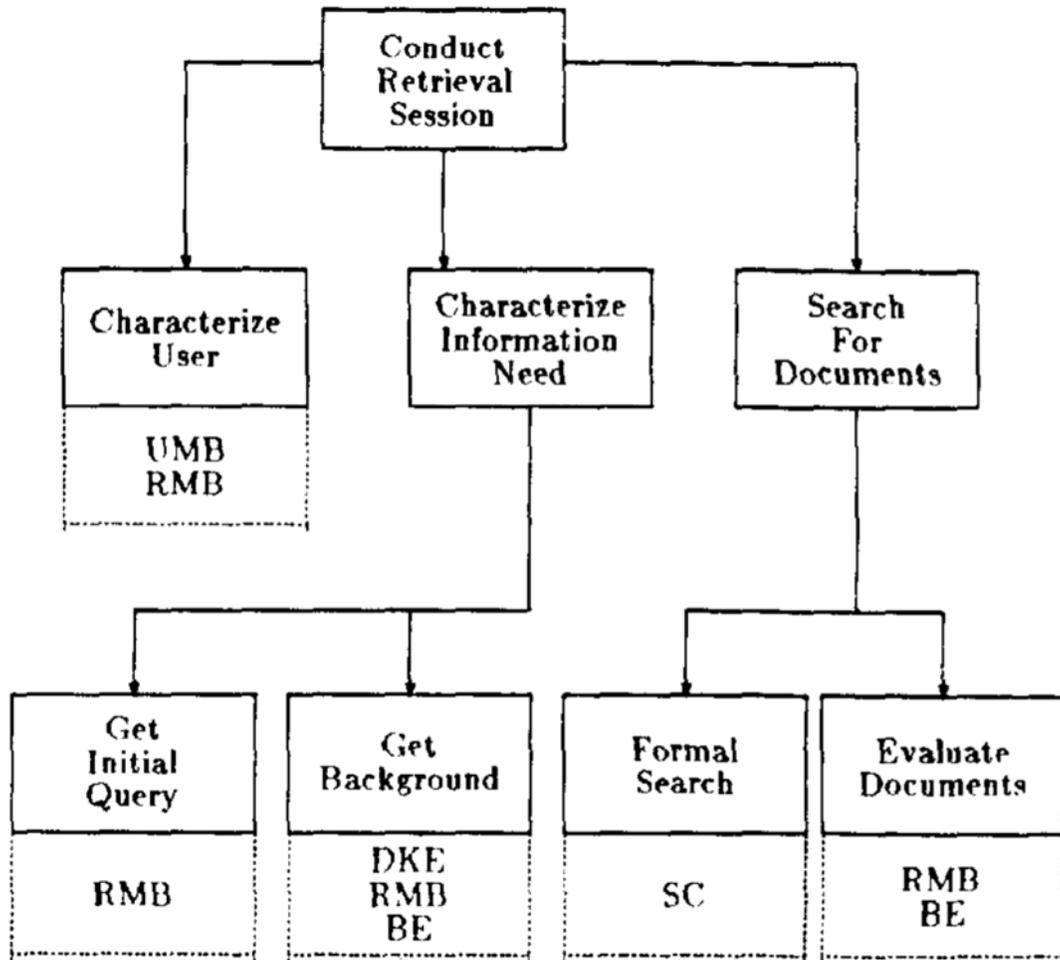


FIG. 3. Default scheduler plan.

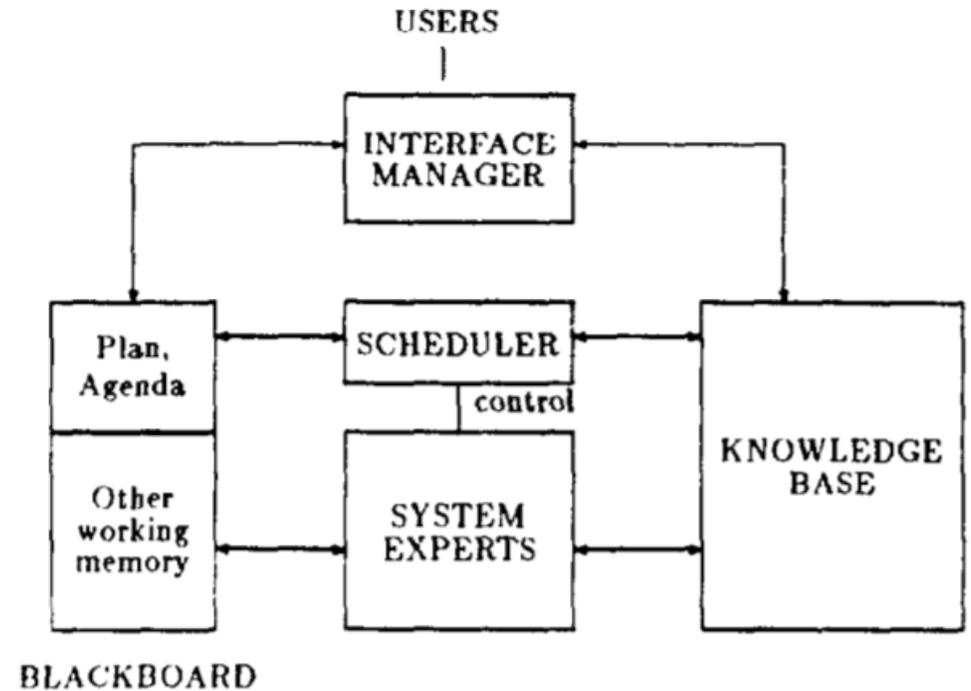
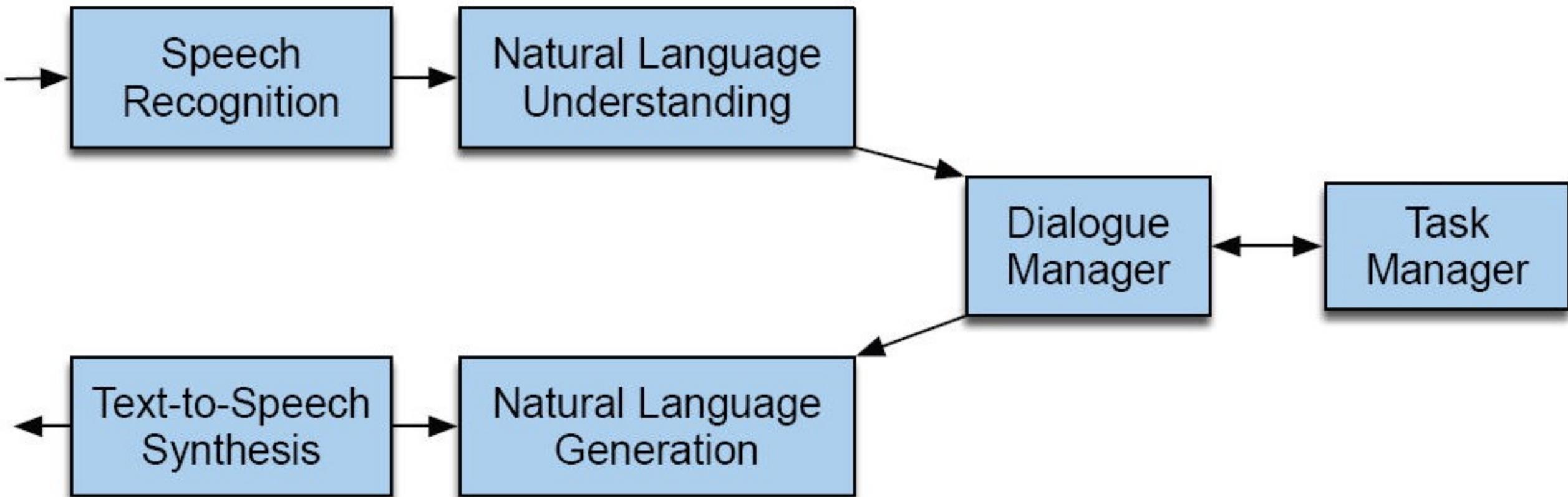


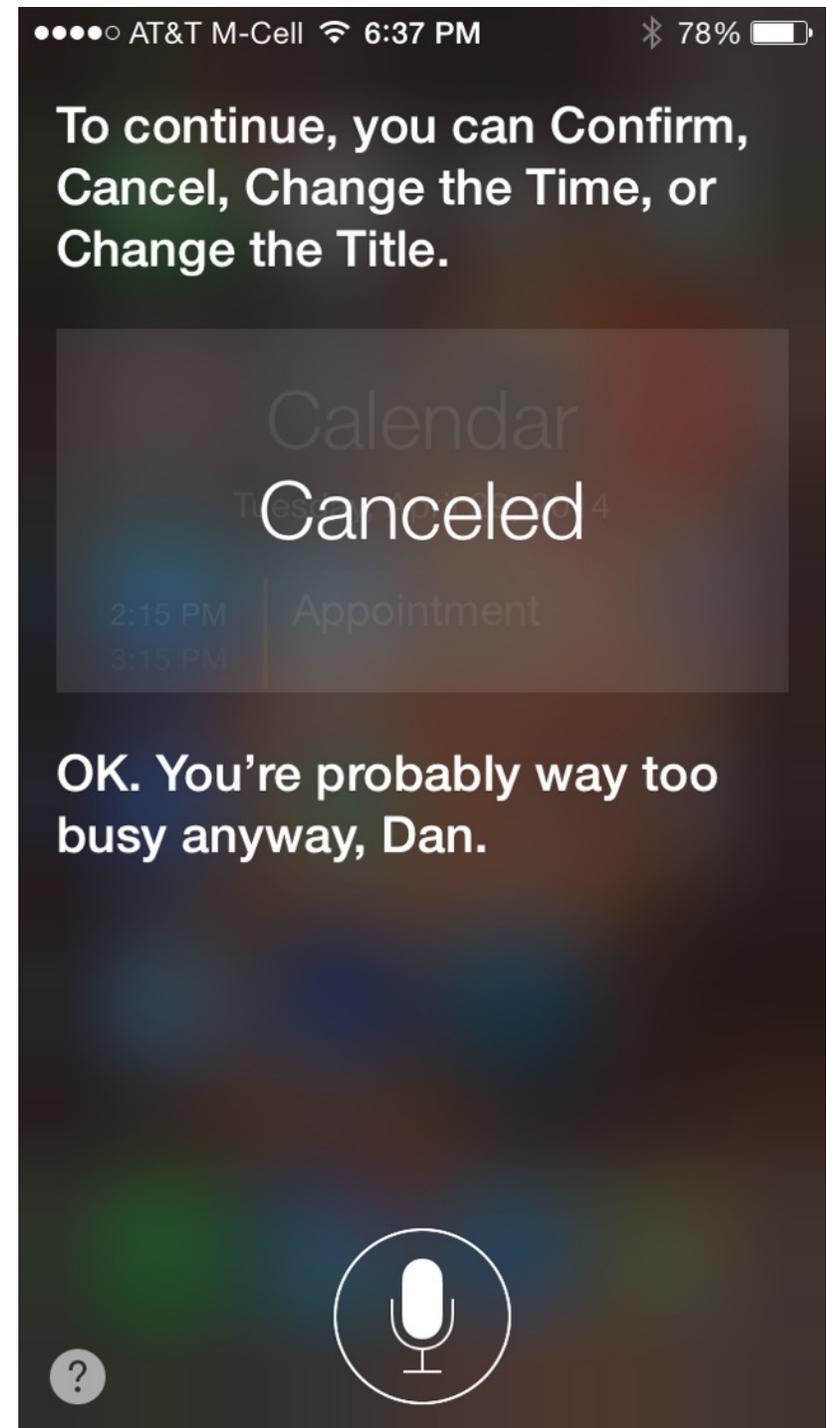
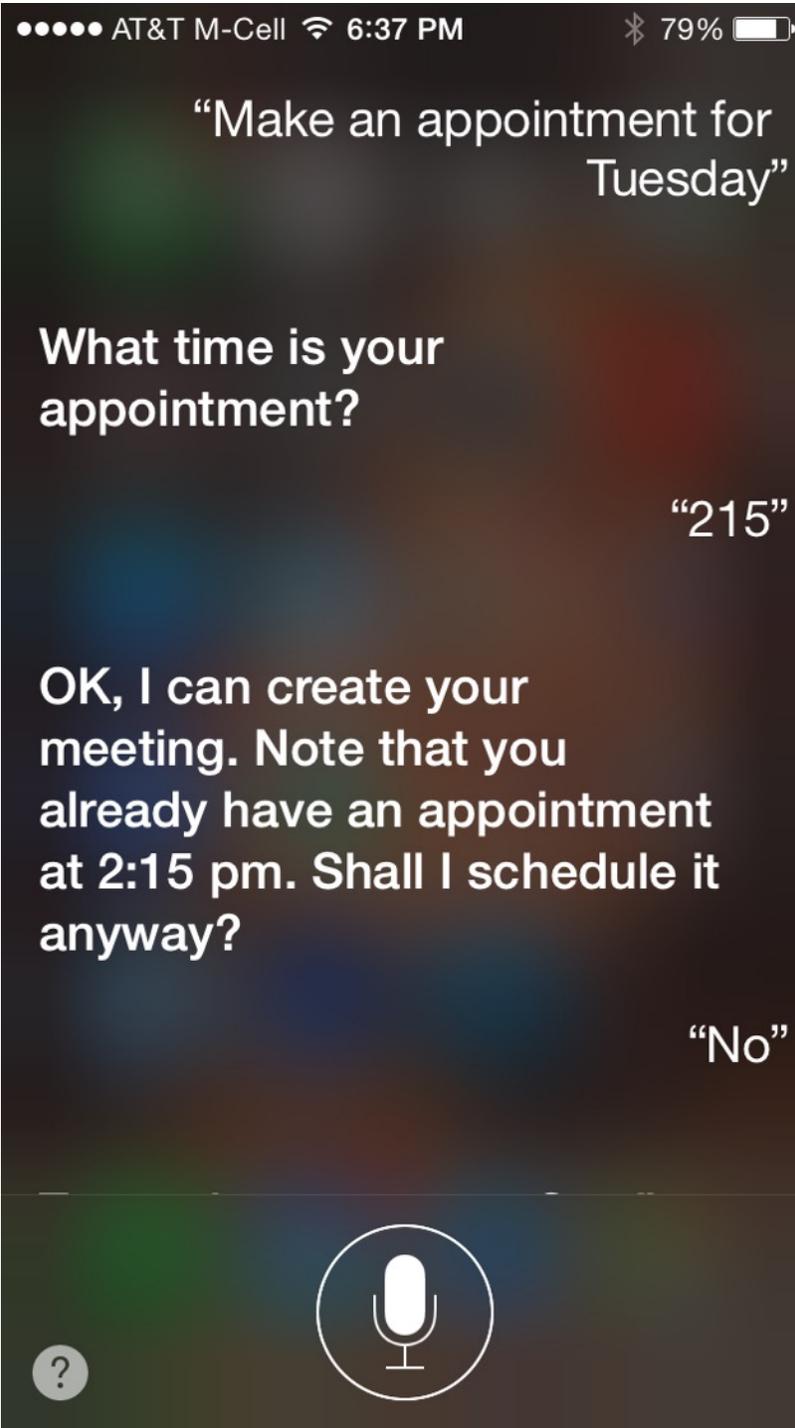
FIG. 2. Major system components.



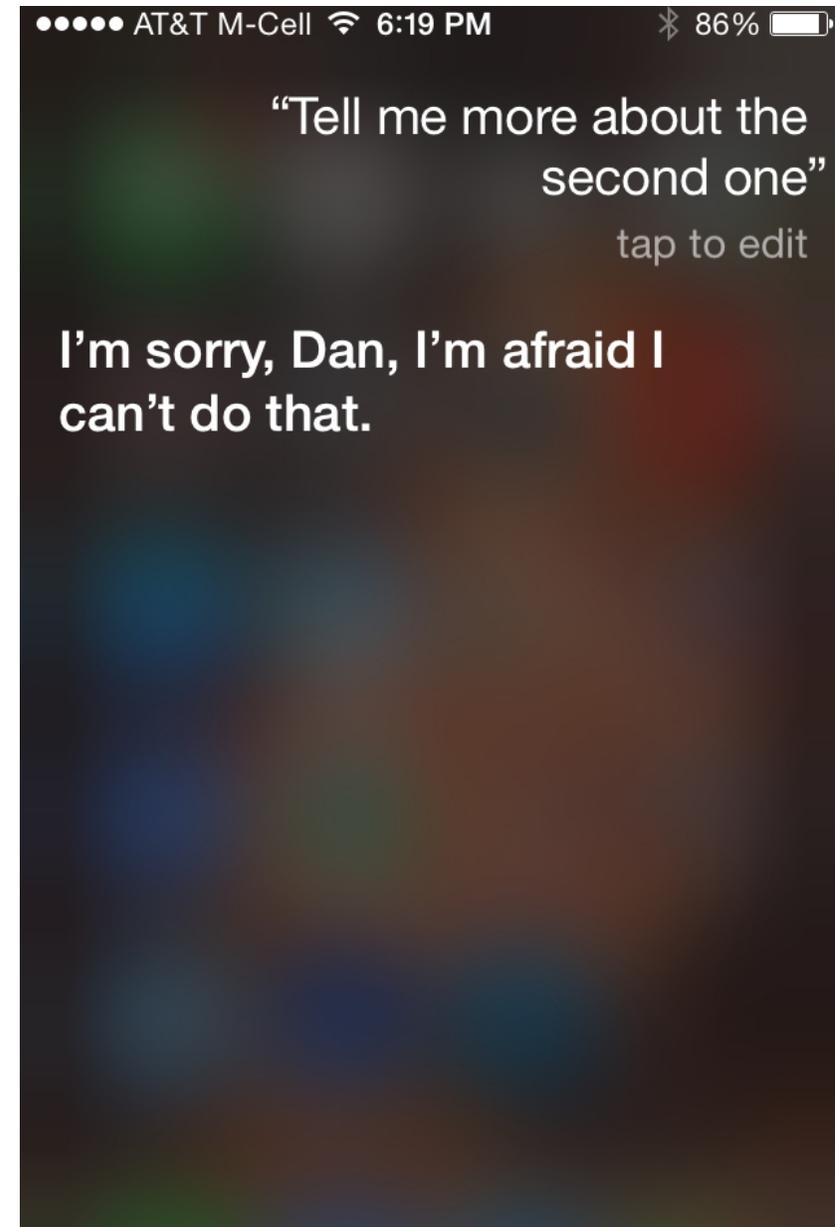
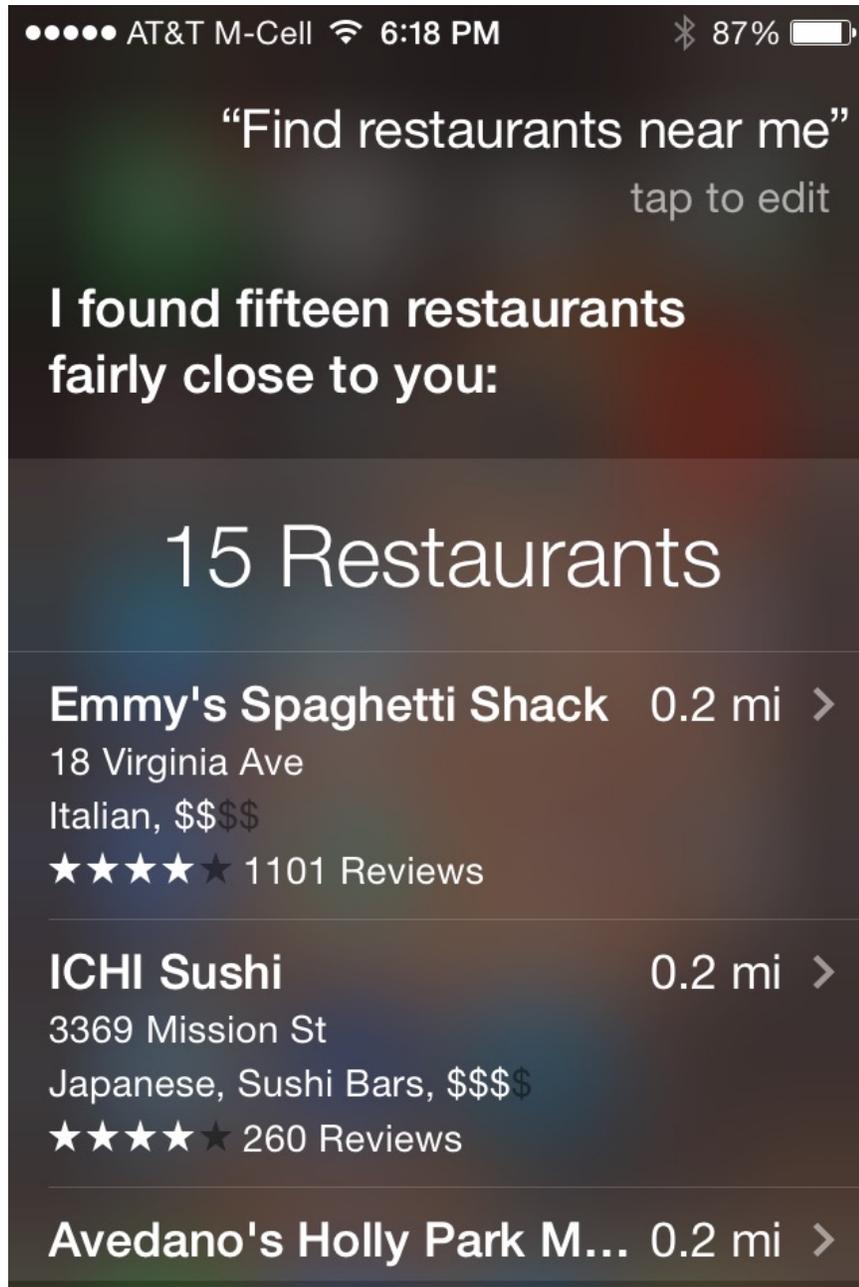
# Classical Voice Assistant Architecture



# SIRI in 2014



Slide from Manning et al.



Slide from Manning et al.

“Find restaurants near me”

tap to edit

I found fifteen restaurants fairly close to you:

# 15 Restaurants

**Emmy's Spaghetti Shack** 0.2 mi >

18 Virginia Ave

Italian, \$\$\$\$

★★★★★ 1101 Reviews

**ICHI Sushi** 0.2 mi >

3369 Mission St

Japanese, Sushi Bars, \$\$\$\$

★★★★★ 260 Reviews

**Avedano's Holly Park M...** 0.2 mi >

“Are any of them Italian”

tap to edit

My web search turned this up:

# Web Search

Are any of them Italian

**any - Dizionario inglese-italiano WordReference**

[www.wordreference.com](http://www.wordreference.com)

English-Italian Dictionary | any ... of any sort  
adj (of an unspecified variety) di qualsiasi

**Italian language - Wikipedia, the free encyclopedia**

[en.wikipedia.org](http://en.wikipedia.org)

Italian or lingua italiana) is a Romance

SIRI in February 2017



"Find restaurants near me"  
tap to edit

Here's what I found:

15 Results

Nearby

Ichi Sushi & Ni Bar

Japanese · 800 feet · Closed now

★★★★☆ (311) on Yelp · \$\$\$



Coco's Ramen

Vietnamese · 800 feet · Closed now

★★★★☆ (280) on Yelp · \$\$



Blue Plate

Restaurant · 0.2 miles · Closed now

★★★★☆ (1251) on OpenTable · \$\$\$



Jasmine Tea House

Chinese · 900 feet · Closed now



"Tell me more about the first one"  
tap to edit

Ichi Sushi & Ni Bar is a Japanese restaurant on Mission St 800 ft to your west. It averages 3½ stars and is somewhat pricey.



Slide from Manning et al.

"Find restaurants near me"  
tap to edit

Here's what I found:

15 Results

Nearby

Ichi Sushi & Ni Bar

Japanese · 800 feet · Closed now

★★★★★ (311) on Yelp · \$\$\$



Coco's Ramen

Vietnamese · 800 feet · Closed now

★★★★★ (280) on Yelp · \$\$



Blue Plate

Restaurant · 0.2 miles · Closed now

★★★★★ (1251) on OpenTable · \$\$\$



Jasmine Tea House

Chinese · 900 feet · Closed now



"Are any of them Italian"  
tap to edit

OK, here's what I found:

15 Results

Nearby

Emmy's Spaghetti Shack

Italian · 0.2 miles · Closed now

★★★★★ (233) on Yelp · \$\$



Vega

Pizza · 0.2 miles · Closed now

★★★★★ (423) on Yelp · \$\$



Pizza Hut

Pizza · 800 feet

★★★★★ (69) on Yelp · \$



La Ciccia

Sardinian · 0.4 miles

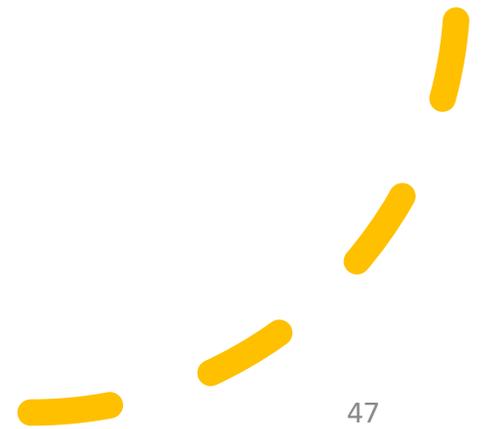


Slide from Manning et al.



## High-level Conversational Search Tasks

- **Conversational Language Understanding**
- **Conversational Ranking & Generation**

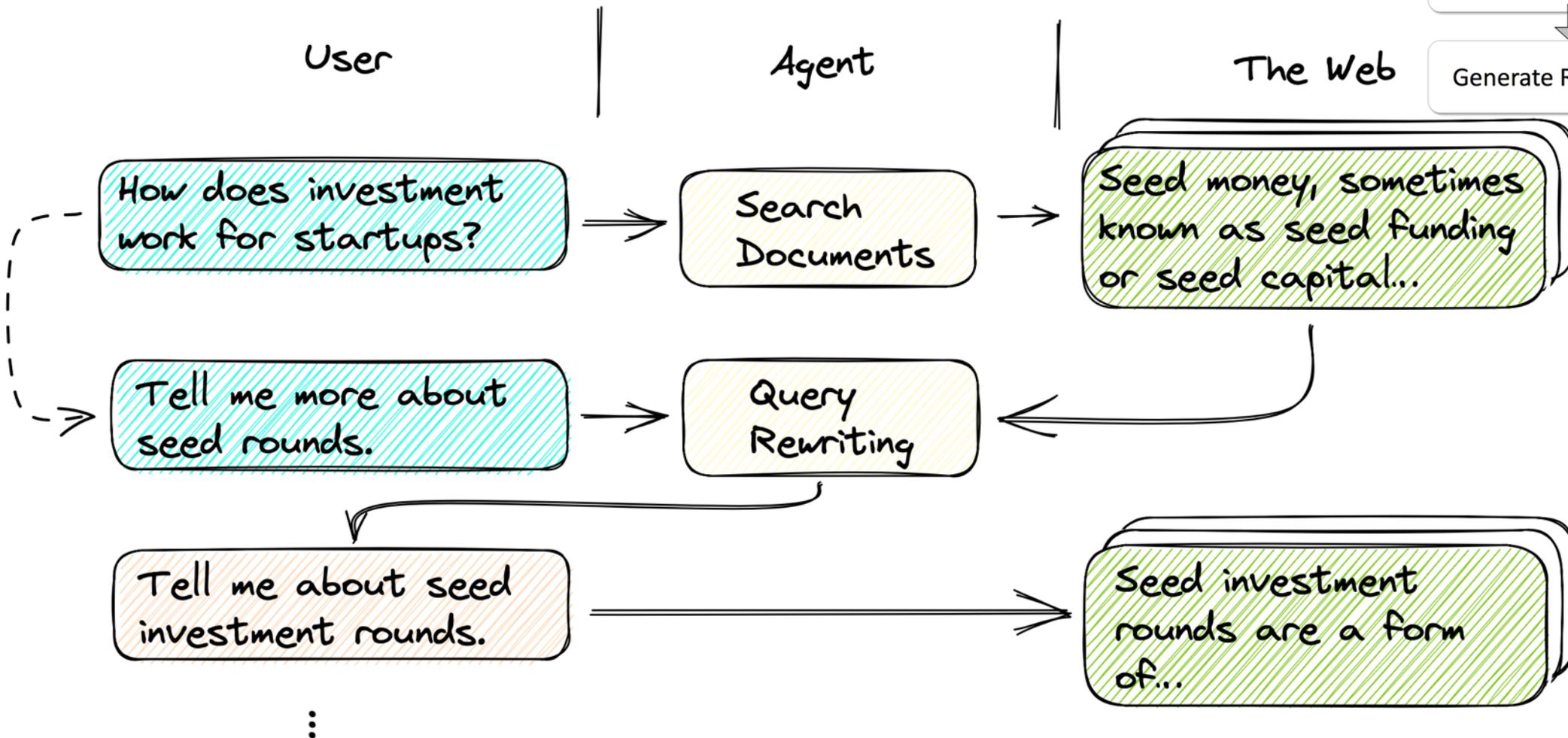
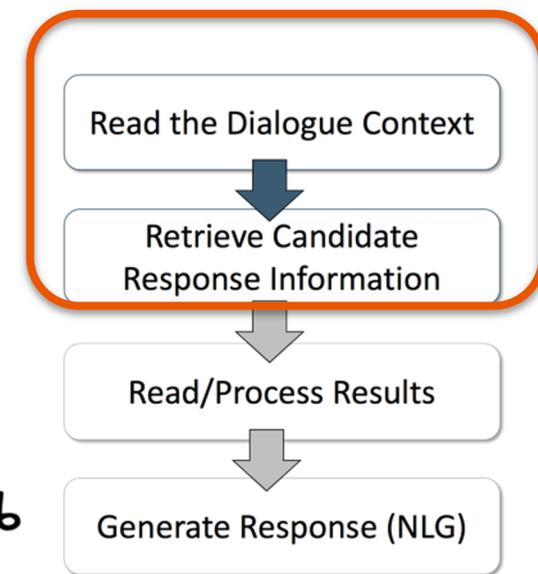


# Conversational QA & Search Benchmarks

- Conversational Question Answering (**ConvQA**)  
SequentialQA, **QuAC**, **CoQA**, QReCC, TopioCQA  
(new!)
- Knowledge graph Conversational Question  
Answering (**KG-ConvQA**)  
Complex Sequential QA, ConvQuestions
- Conversational Passage Retrieval (**ConvPR**)  
**TREC CAsT**, QReCC
- Conversational Document Retrieval (**ConvDR**)  
**TREC CAsT**

# Web-scale Exploratory Conv. Retrieval

TREC Conversational Assistance Track (2019-Present)



# CAsT Information Tasks



- 20-25 conversational topics per year

## Topics

- Informational
- No temporal or outside context
- Not personal / subjective
- Avoid sensitive content
- “Right level” of specificity
- “Meaningful” trajectory
- Approximately 10 turns

## Turns

- Majority should be non-factoid
- Majority should be context dependent
- Coherent (topic flow)
- Use well-formed natural language with coreference

## Example anonymized session

cost of Lamborghini  
food truck price  
license and permits of food truck  
good locations  
best hours of the day  
how to run it successfully  
small business training

### People also ask

What is a typical day for a food truck?

What does a food truck owner do?

How much does it cost to run a food truck daily?

How much does a typical food truck make?



Interactive  
CAsT (iCAsT)

<https://bit.ly/i-cast>



# Conversational Language Understanding

Chapter 4

# Conversational modeling



# Example CAsT Y3 topic: genetic engineering



- How do genes work?
- What other diseases are caused by a single change?
- What are the other types of diseases?
- You missed the second type of disease. What was that?
- That's not what I wanted. How about recent developments in gene therapy to treat those defects?

A large genetic mistake typically occurs in the woman's egg, which may partially explain why older women are more likely to have babies with Down syndrome... Down syndrome is the most common and well-known chromosome defect, but there are many. Types of chromosome diseases: There are several common types of chromosome errors that cause disease. The effects of errors in the sex chromosomes (X and Y) differ greatly from errors in the autosomes (chromosomes 1..22).  
**(MARCO\_D76761)**

# Y3 topic: genetic engineering

- How do genes work?
- What other diseases are caused by a single change?
- What are the other types of diseases?
- You missed the second type of disease. What was that?
- That's not what I wanted. How about recent developments in gene therapy to treat those defects?
- What are they worried about?
- No, I meant in humans.
- It sounds like it could be used in many places. What other types of organisms has it been tried on?
- I've heard a lot about RNA recently. Can it be used to edit that too?
- What's the difference between the types you mentioned?
- That's too basic, I'd like a more scientific explanation.
- The developments sound exciting. What are the commercial issues using it?
- What are the alternatives to avoid licensing issues?

# Y3 topic: genetic engineering

How do genes work?  
What other diseases are caused by a single change?  
What are the other types of diseases?  
You missed the second type of disease. What was that?  
That's not what I wanted. How about recent developments in gene therapy to treat those defects?  
What are they worried about?  
No, I meant in humans.  
It sounds like it could be used in many places. What other types of organisms has it been tried on?  
I've heard a lot about RNA recently. Can it be used to edit that too?  
What's the difference between the types you mentioned?  
That's too basic, I'd like a more scientific explanation.  
The developments sound exciting. What are the commercial issues using it?  
What are the alternatives to avoid licensing issues?

**Dependence on previous results**

**Sequence reference**

**Feedback**

**Topic shift**

Widely varying  
discourse  
structure

## QuAC

- Continuation
  - (follow up, maybe follow up, or don't follow up)
- Affirmation
  - (yes, no, or neither)
- Answerability
  - (answerable or no answer)

## TREC CAsT (Y3)

- Questions (~85%),
- Feedback (10%),
- Revelation (5%),
- Elaboration (5%)

Most existing CIS datasets and models have limited discourse types with **users asking questions**, with the **system responding with answers** or a clarifying question.

# CAsT Y3 discourse examples

**Feedback** ~ 10% of turns

**Revelment** ~5% of turns

**Elaboration** ~5% of turns

[Dalton et al. 2021]

- Does the article have more about it?
- Could you expand on some of these methods?
- Give me some examples.
- Tell me more about them.

four weeks.

# Conceptualizing Agent-Human Interactions

- A taxonomy of User and Agent CIS behaviors.

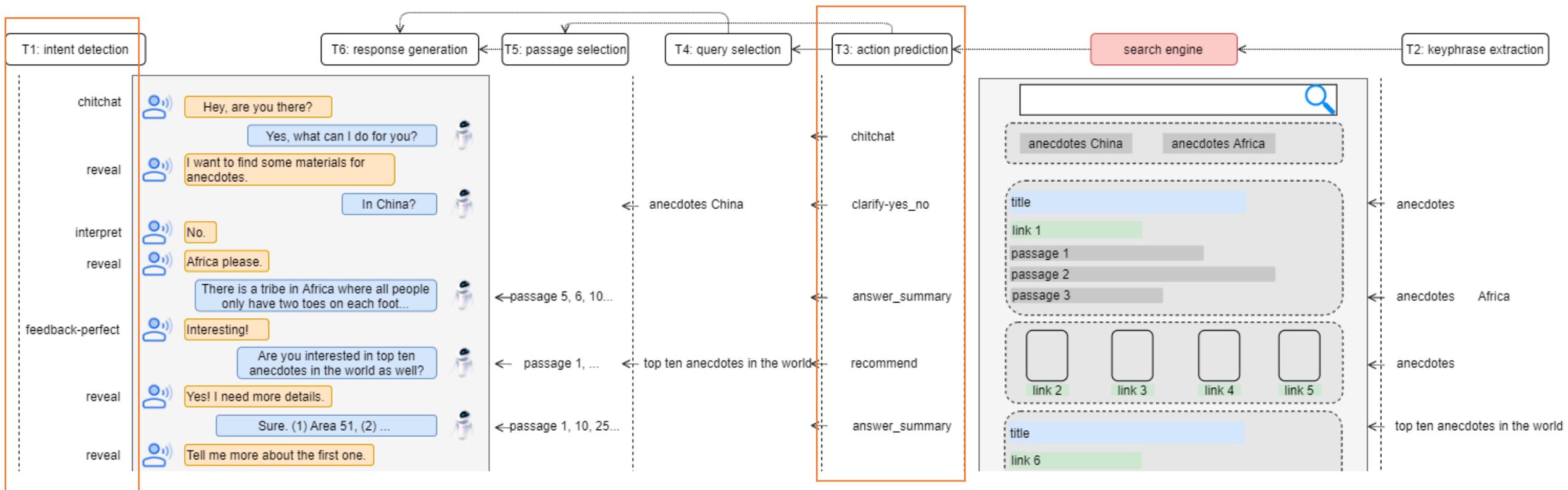
		User	Agent		
	Query Formulation	<b>Reveal</b>	<b>Inquire</b>	User Revealment	Memory
		Disclose, Non-Disclose	Extract, Elicit, Clarify		
		Revise, Expand	Elaborate		
Set Retrieval	Result Exploration	<b>Inquire</b>	<b>Reveal</b>	System Revealment	
		List, Summarize, Compare	List, Summarize, Compare		
		Subset, Similar	Subset, Similar		
		<b>Navigate</b>	<b>Traverse</b>		
		Repeat, Back, More,..., Note	Repeat, Back, More,..., Record		
Mixed Initiative		<b>Interrupt</b>	<b>Suggest</b>		
		Question, Stop, Change	Recommend, Hypothesize		
			<i>Voice Opinion</i>		
		<b>Interrogate</b>	<b>Explain</b>		
		Understand, Explain	Report, Reason		

User: I would to arrange a holiday to Italy [Disclose - Volunteer]  
 Agent: When would to go on holidays?  
 User: The 4th of May [Disclose - Inquire].  
 ...  
 Agent: Do you where in Italy you like to go on holidays?  
 User: I'm not sure [Disclose - Unsure].  
 ...  
 Agent: What is your budget?  
 User: I'd prefer not to say [Disclose - Not].

User: Tell me about all the different things you can do in Tuscany? [Inquire List]  
 ...  
 User: Can you give me an overview of the things to do there? [Inquire Summarize]  
 ...  
 User: What is the best thing to do in Tuscany? [Inquire Subset]  
 ...  
 User: Which things are suitable for children? [Inquire Subset]  
 ...  
 User: What are the main differences between Tuscany and Galicia? [Inquire Compare]  
 ...  
 User: What other regions in Europe are like that? [Inquire Similar ]

# Wizard of Search-Engine

- User intents: reveal, revise, interpret, request-rephrase, chitchat
- Intermediary: clarify, answer-type, answer-form, no-answer, request-rephrase, chitchat





Building blocks

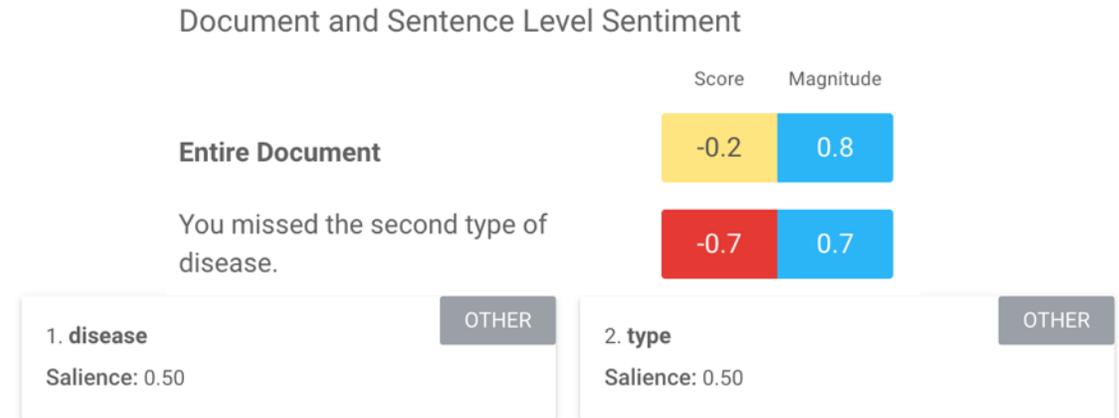
What is CIS turn state?

# Representing a single utterance

- Words
- Standard NLP annotations
  - Sentiment
  - Entities
  - ...
- Discourse/intent classification(s)
- Embedding at token or sequence level
  - BERT, S-BERT, ANCE, etc.

Others...

[You, missed, the, second, type, of, disease,. What, was, that,?]

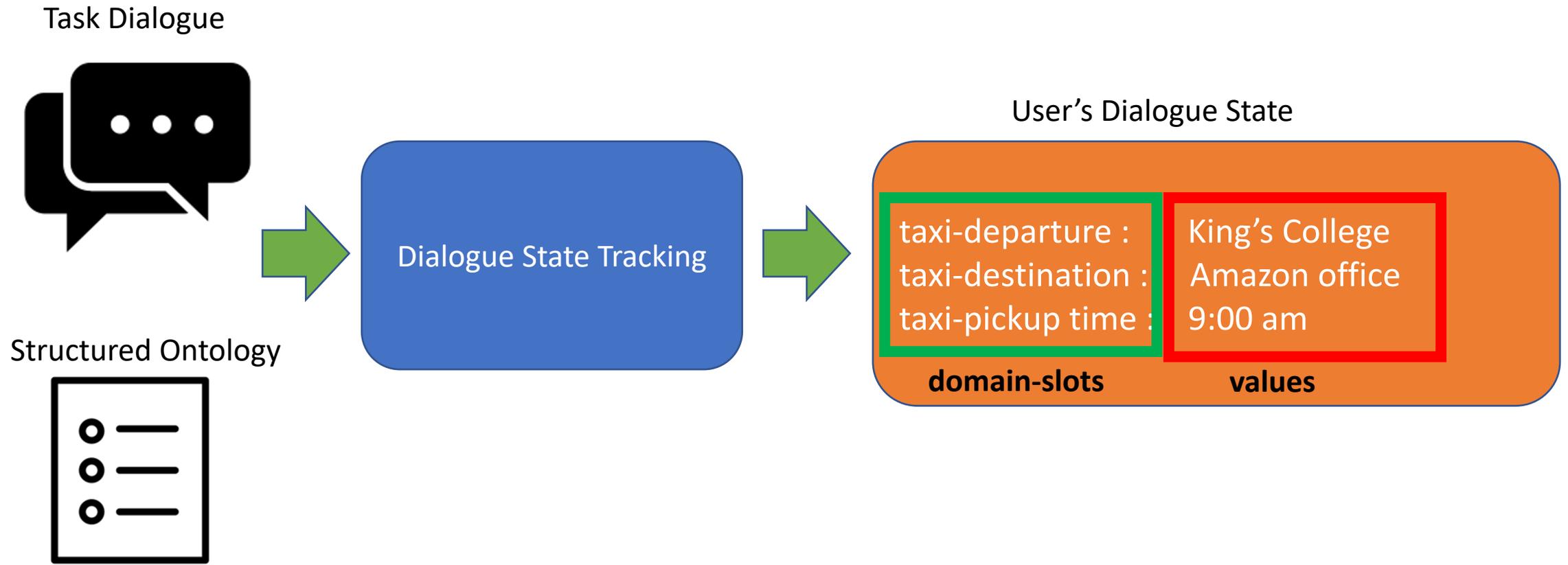


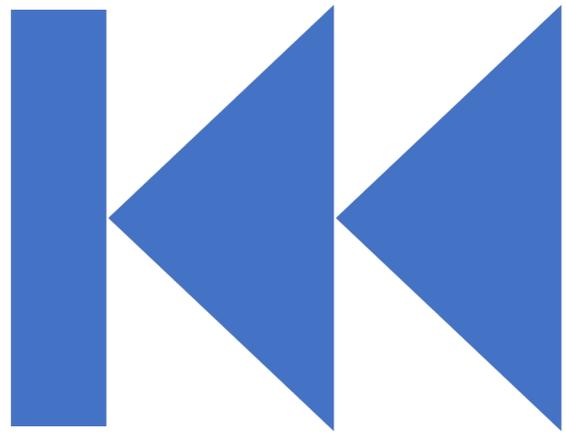
[You missed the second type of disease. What was that?] – **Intent: Clarification / Feedback**

[0.4, -0.11, 0.55, 0.3 . . . 0.1, 0.02]

With increased adoption of pre-trained language models, “just the text” is the most common.

# A Comparison versus Dialogue State Tracking





# Tracking Multi-Turn...

---

The most important feature in modeling history is the **positional relationship between turns** to capture common patterns of conversational discourse.

---

# Last-K context - a simple, but effective heuristic

Append the previous **K-context** (ctx) turns

- Previous user utterances (queries)
- Previous system utterances (responses)

Best  $K$  is typically 2-3

Effective across diverse models and tasks

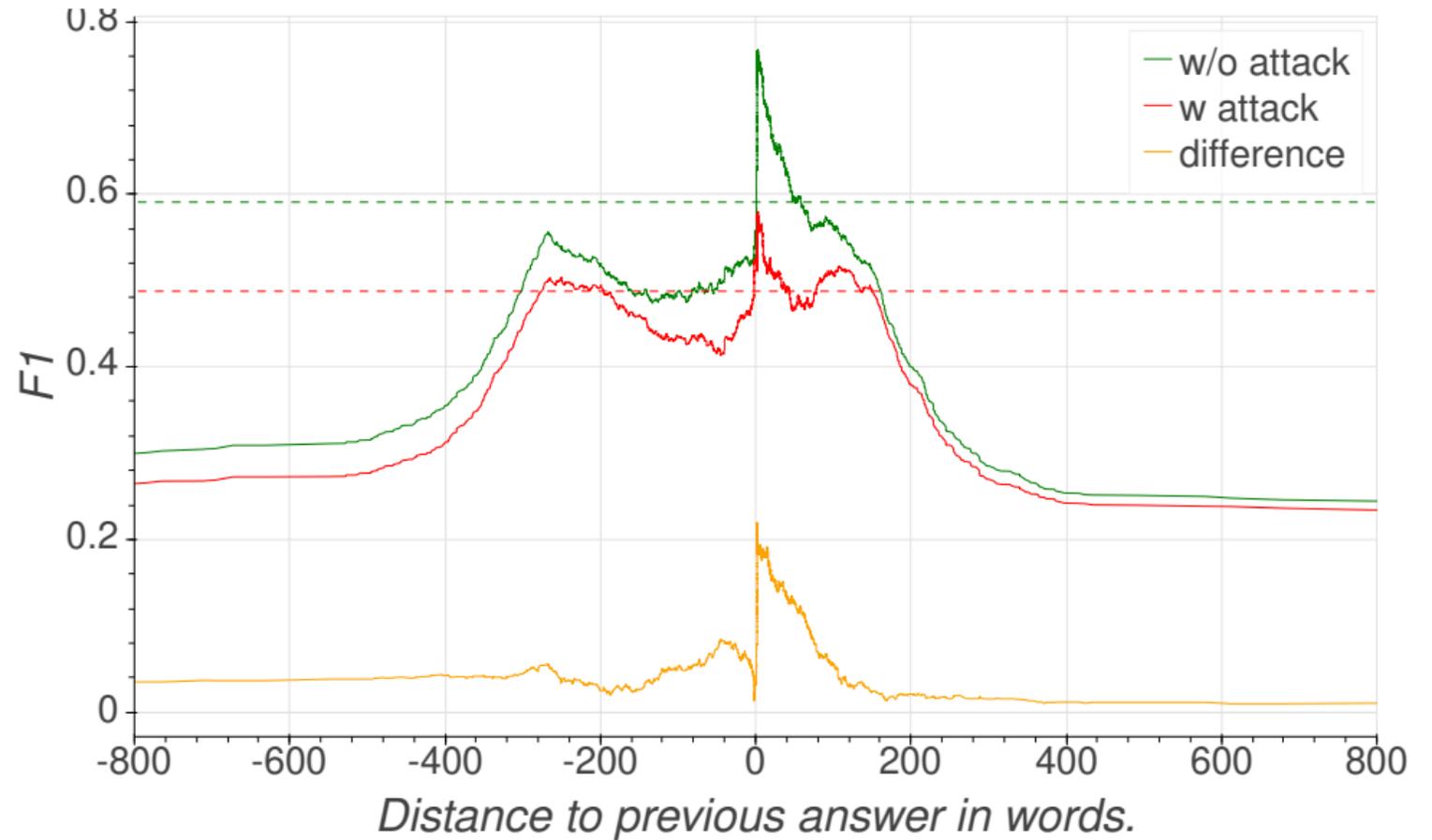
- ConvQA models
- Dialog State Tracking  
(BERT w/3-ctx) – [Mehri et al., 2020]

	# contexts	CoQA	QuAC
BERT w/ 0-ctx	0	72.8	55.0
BERT w/ 1-ctx	1	79.2	63.4
BERT w/ 2-ctx	2	79.6	<b>65.4</b>
BERT w/ 3-ctx	3	79.6	65.3
BERT w/ 4-ctx	4	79.4	64.8
BERT w/ 5-ctx	5	<b>79.7</b>	64.5
BERT w/ 6-ctx	6	79.5	64.9
BERT w/ 7-ctx	7	<b>79.7</b>	64.4

Simple but Effective Method to Incorporate Multi-turn Context with BERT for Conversational Machine Comprehension [Ohsugi et al., 2019]

# Position Robustness Attack

Simply repeat the answer to make the distance between answers longer.



# Context Modeling Summary

- **Most context modeling approaches so far do not beat simple rules.**

## **Possible reasons**

- Most datasets have 2-15 turns
- Single session on one topic
- Most turns have only 'local' dependence on context

## **Beyond *k*-ctx → advanced models of history**

- Open research area
- May require new conversational collections

# Conversational Language Understanding Tasks



**Conversational Query Expansion (CQE)**



**Conversational Query Rewriting (CQR)**



**Conversational Entity Detection and Linking (CEDL)**



**Context Salience (word, turn)**



**Intent & Sub-topic Classification**



**Context Ranking & Summarization**

# Query Expansion

## Task

$$Q \rightarrow Q_{exp}$$

Select and weight important words from the conversation history and/or PRF.

### • Rules

- First turn, previous turn [Clarke, 2019]
- Historical Query Expansion (HExp) – [Yang et al., 2019]

### • Supervised

- Conversational Term Selection (CVT) [Kumar et al., 2020]
- QuReTeC [Voskarides et al., 2020]

---

#### Algorithm 1: Historical Query Expansion

---

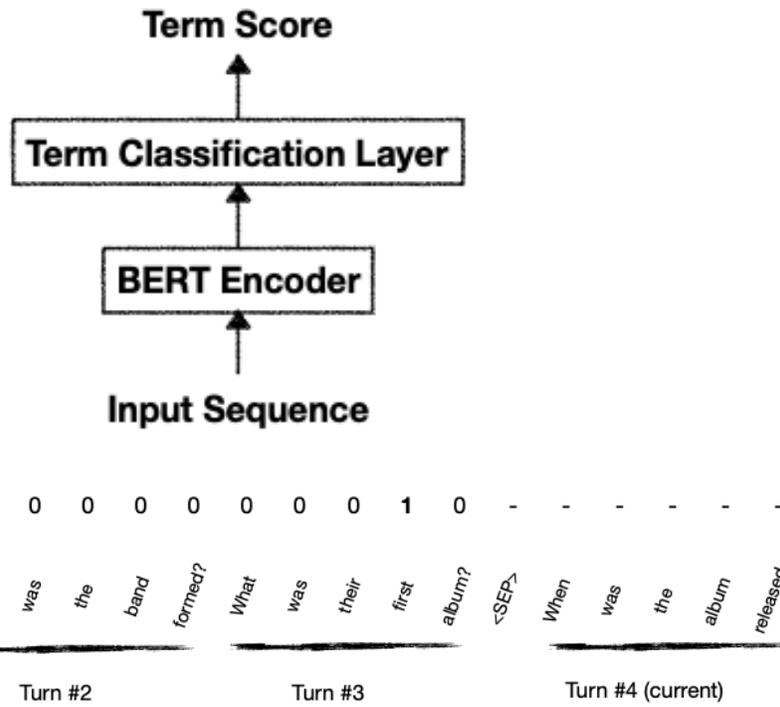
```
Input:  $u_i, u_{<i}, C$   
Output:  $\bar{u}_i$   
1  $\bar{u}_i \leftarrow ()$ ;  $W_{\text{topic}} \leftarrow \{\}$ ;  $W_{\text{sub}} \leftarrow \{\}$   
2 for  $j = 1$  to  $i$  do  
3   for  $k = 1$  to  $n(u_j)$  do  
4      $\mathcal{R}_j^k = \text{KE}(t_j^k, C)$   
5     if  $\mathcal{R}_j^k > \mathcal{R}_{\text{topic}}$  then  
6        $W_{\text{topic}}.\text{insert}(t_j^k)$   
7     if  $(\mathcal{R}_j^k > \mathcal{R}_{\text{sub}})$  and  $(j \geq i - M)$  then  
8        $W_{\text{sub}}.\text{insert}(t_j^k)$   
9 if  $i > 1$  then  
10    $\mathcal{A}_i = \text{QPP}(u_i, C)$   
11    $\bar{u}_i.\text{insert}(t)$  for all  $t \in W_{\text{topic}}$   
12   if  $\mathcal{A}_i < \eta$  then  
13      $\bar{u}_i.\text{insert}(t)$  for all  $t \in W_{\text{sub}}$   
14  $\bar{u}_i.\text{append}(u_i)$   
15 return  $\bar{u}_i$ 
```

---

Figure from [Yang et al., 2019]

# Supervised Query Expansion

- Query Resolution by Term Classification (QuReTeC)
- BERT binary term classification from the history
- Labels from distant supervision from rel. passages



**Table 10: Qualitative analysis for initial retrieval (extrinsic) when using QuReTeC or RM3 (cur+first) for query resolution. The example is sampled from the TREC CAsT dataset.**

- Q1: What is a real-time database?
- Q2: How does it differ from traditional ones?
- Q3: What are the advantages of real-time processing?
- Q4: What are examples of important ones?
- Q5: What are important applications?
- Q6: What are important cloud options?
- Q7: Tell me about the Firebase DB?
- Q8 (current): How is it used in mobile apps?

**Predicted terms – QuReTeC:** {"database", "firebase", "db"}

**Top-ranked passage – QuReTeC**

Firestore is a mobile and web application platform ... Firestore's initial product was a realtime database, ... Over time, it has expanded its product line to become a full suite for app development

**Predicted terms – RM3 (cur+first):** {"real", "time", "database"}

**Top-ranked passage – RM3 (cur+first)**

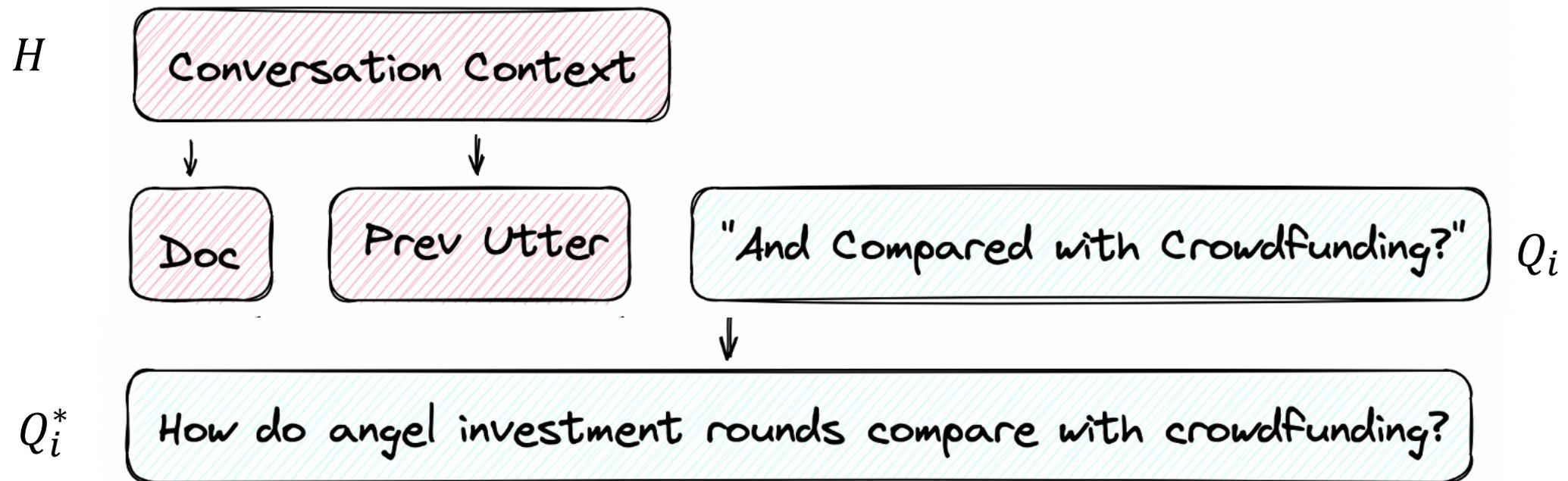
There are two options in Jedox to access the central OLAP database and software functionality on mobile devices: Users can access reports through the touch-optimized Jedox Web Server ... on their smart phones and tablets.

(b) Example input sequence and gold standard term labels (1: relevant, 0: non-relevant) for QuReTeC.

# Conversational Query Rewriting (CQR)

- Generate contextualized query from conversation context
- Sequence-to-sequence approach =  $P(\mathbf{y}|\mathbf{x}; \theta) = \prod_{j=1}^J P(y_j|y_{<j}, \mathbf{x}; \theta)$

$$Q_i^* = CQR(Q_i, H; \theta)$$



# CAsT Y1: NLP-based rewriting baseline

- Run AllenNLP or other NLP toolkit to identify entities and mentions.
- **Rewriting:** Replace 'coreferent' mentions with 'canonical name'

How much does 0 a used Lamborghini cost? How does 0 it compare to a Ferrari? Interesting. What about for a pimped – out food truck? What licenses and permits are needed? What is a typical day like? How can I run 1 it successfully? What are some good examples to learn from? Besides inventive flavors, what made 1 it successful?

# CAsT Y1 Coreference Phenomena

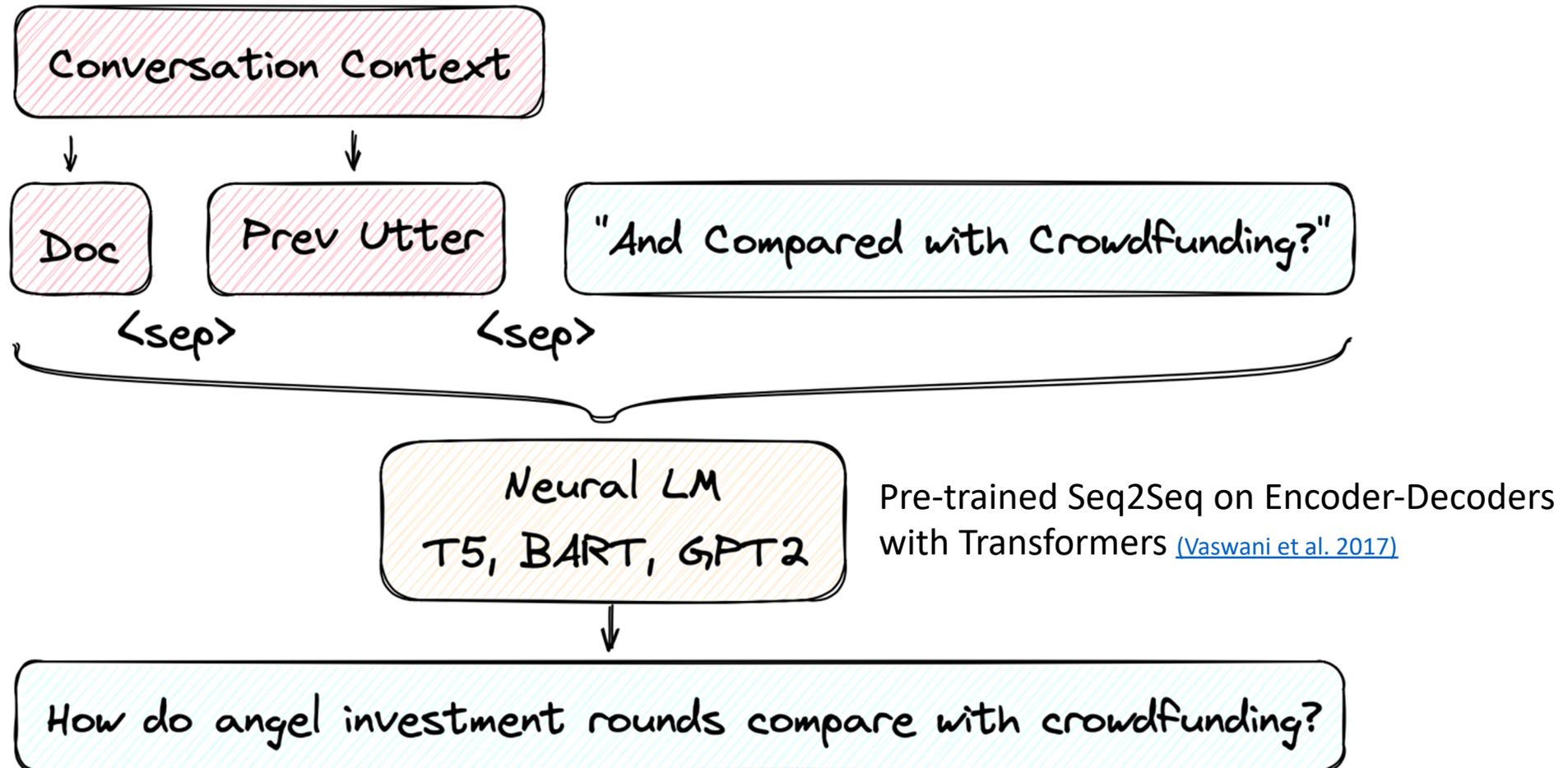
## Examples

Type	Utterance	Mention
<b>Pronominal</b>	How do <b>they</b> celebrate Three Kings Day?	they -> Spanish people
<b>Zero</b>	What cakes are traditional?	Null -> Spanish, Three Kings Day
<b>Groups</b>	Which <b>team</b> came first?	which team -> Avengers, Justice League
<b>Abbreviations</b>	What are the main types of <b>VMs</b> ?	VMs -> Virtual Machines

## Statistics

Dataset	Pronominal	Zero	Groups	Abbreviations
TRAIN	102	82	6	29
EVALUATION	128	111	4	15

# Neural Query Rewriting



# CQR Datasets

- **CAsT (2019-present)**

- Few-Shot Generative Conversational Query Rewriting [Yu et al., 2020]
- A few hundred turns

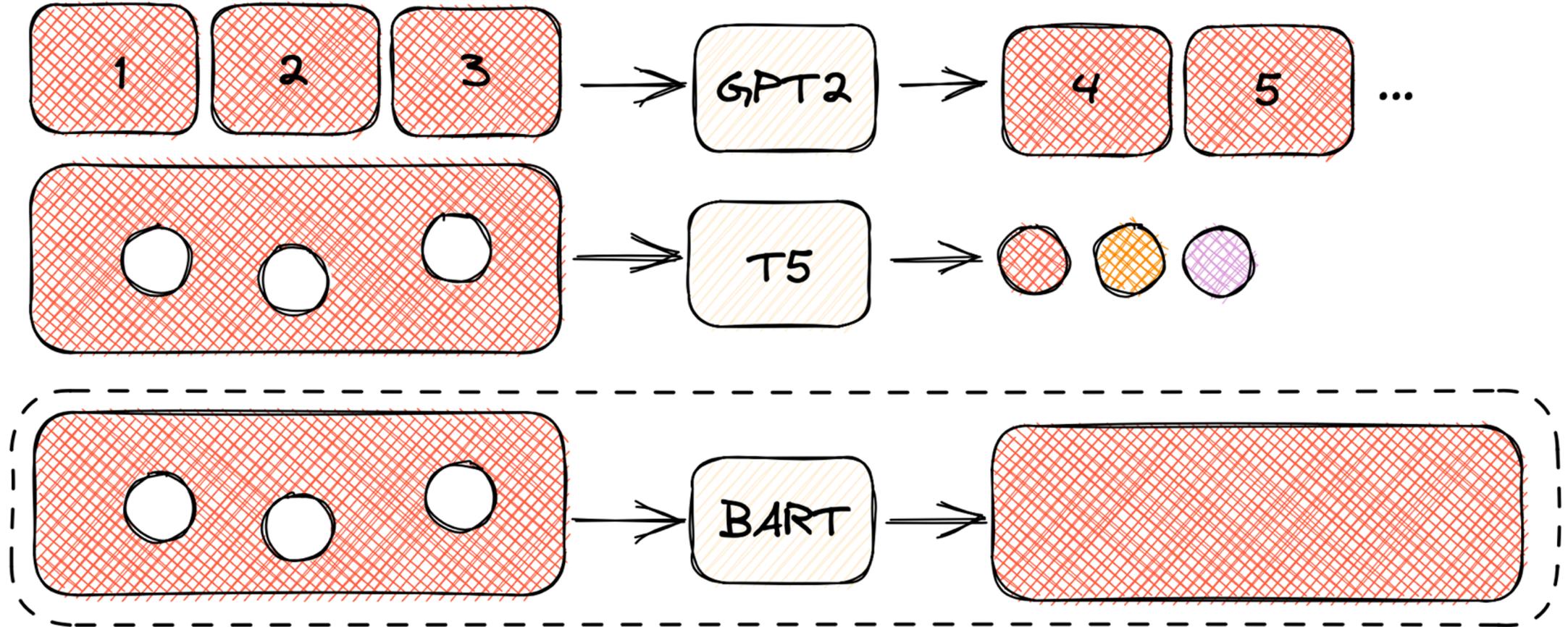
- **CANARD – (2019)**

- Based on QuAC - 40,527 turns
- Conversational Question Reformulation via Sequence-to-Sequence Architectures and Pretrained Language Models [Lin et al., 2020]

- **QReCC (2021)**

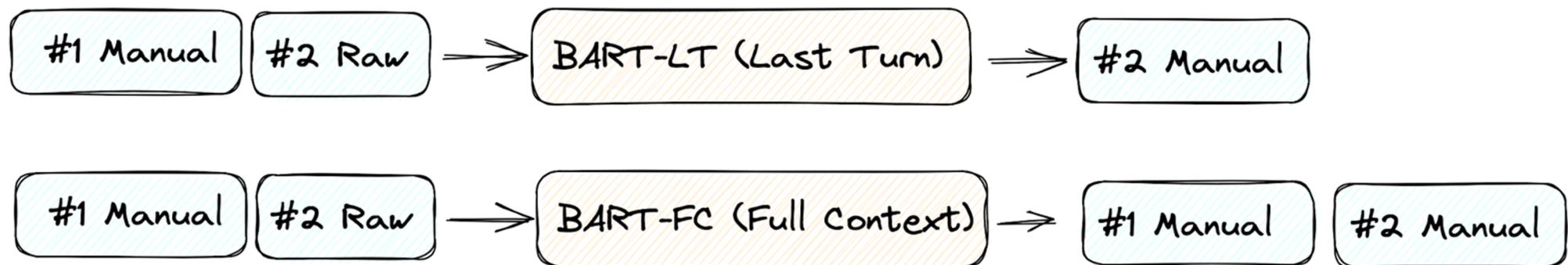
- 14K conversations with 81K question-answer pairs

# Neural LM Pre-Training



# BART FC: Aligning Pre-Training with Fine-Tuning

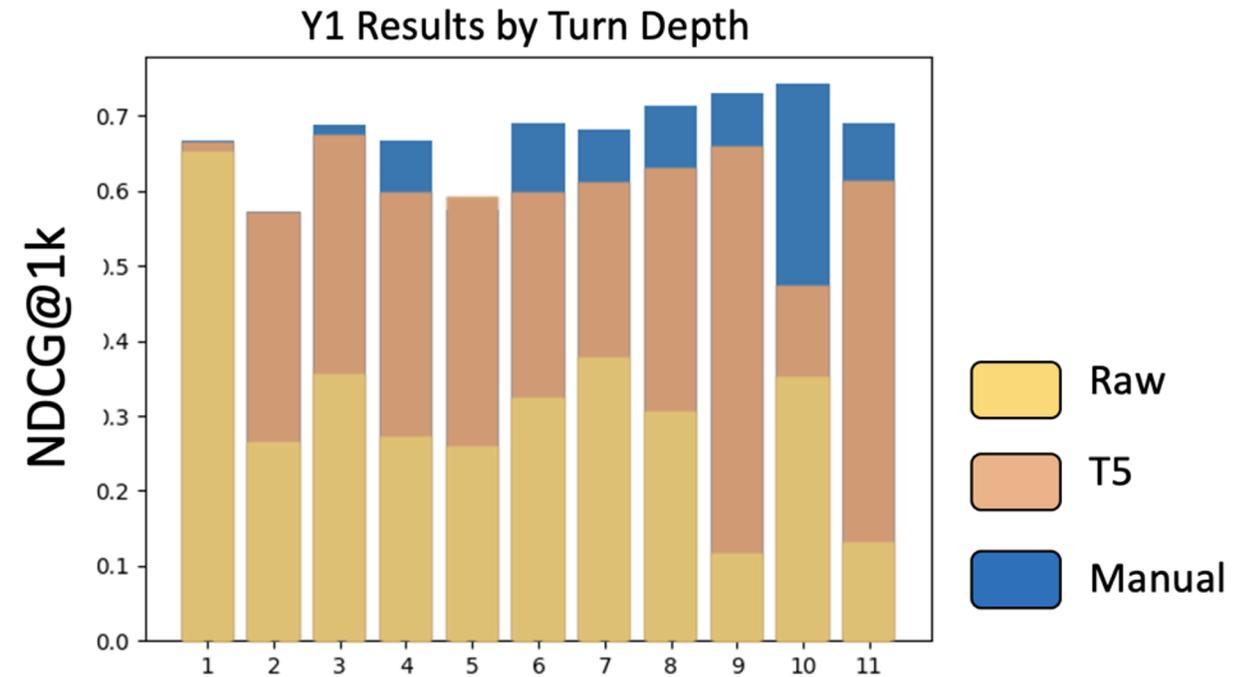
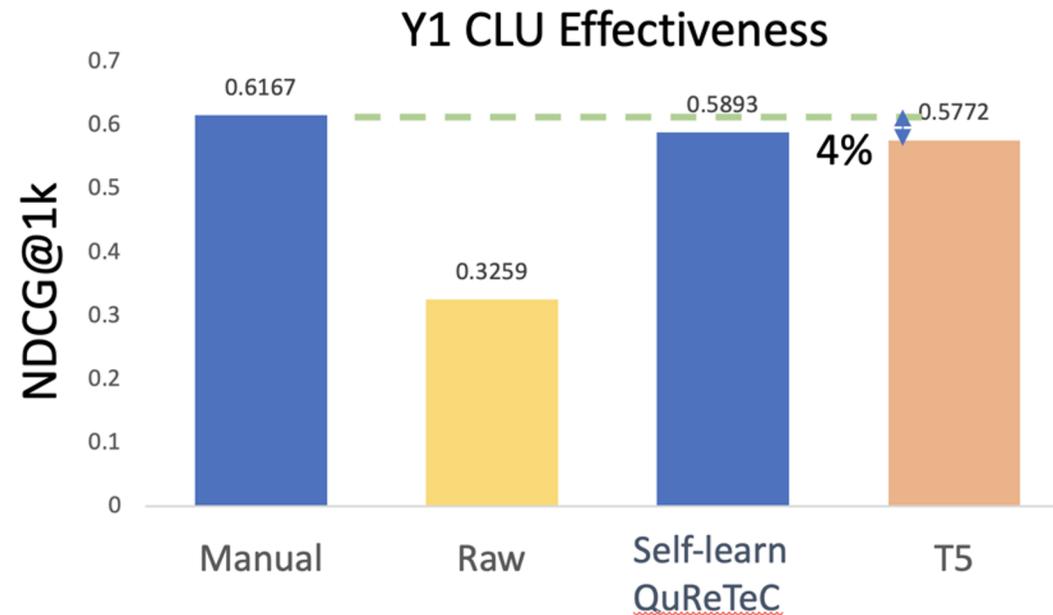
Training



Inference

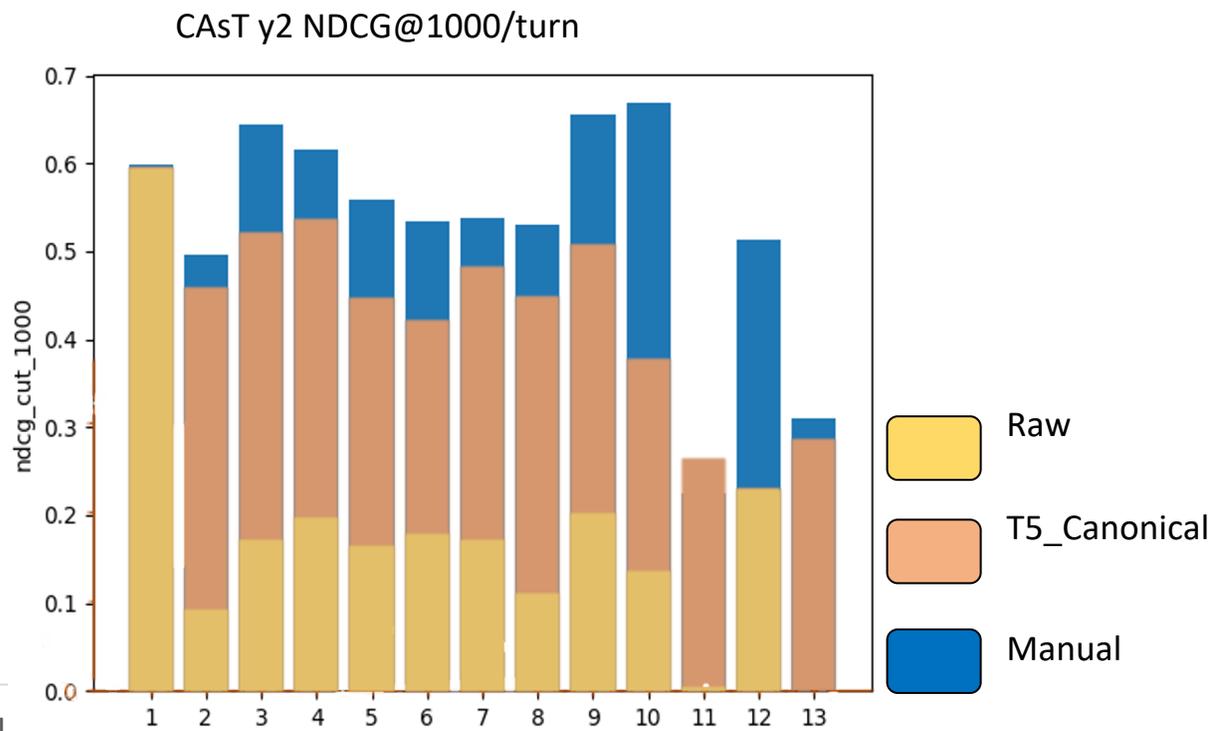
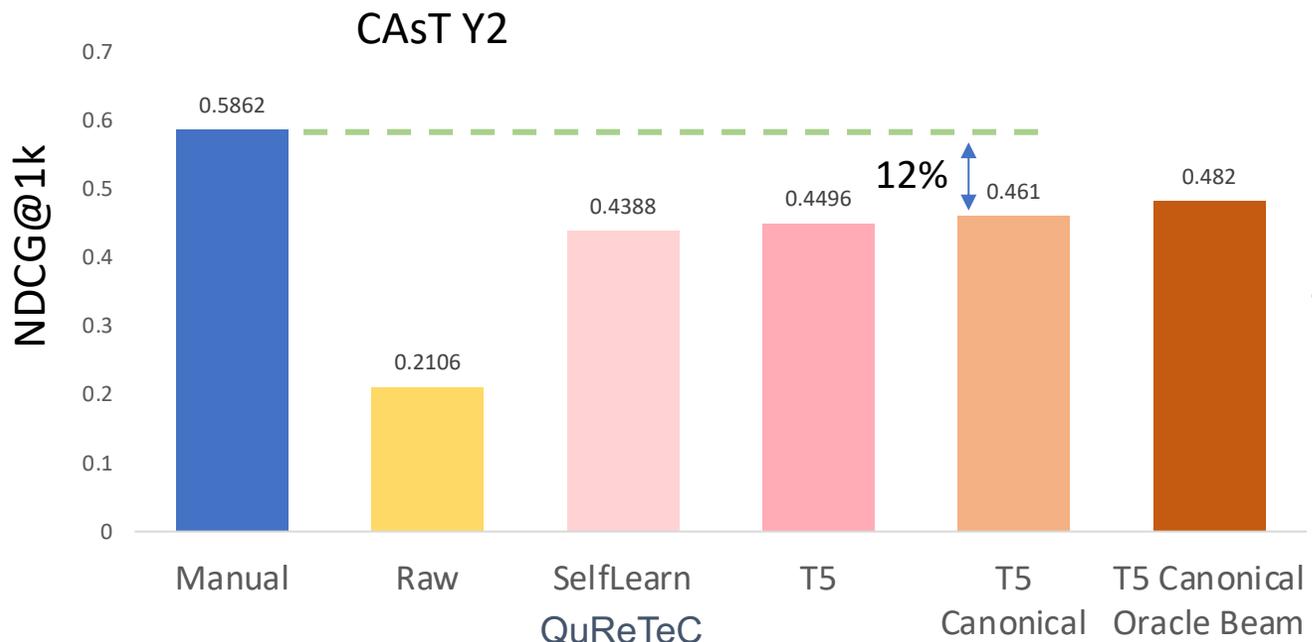


# CAsT Y1 Results: Rewriters Do Well



No dependence on results, only previous user queries.

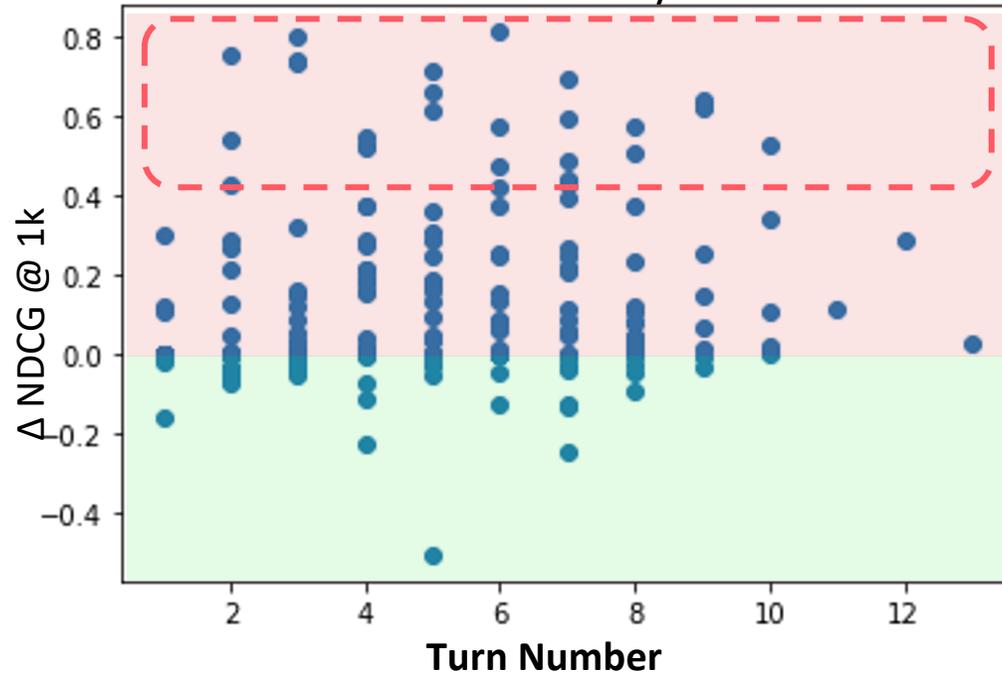
# CAsT Y2 Results: Rewriters Struggle



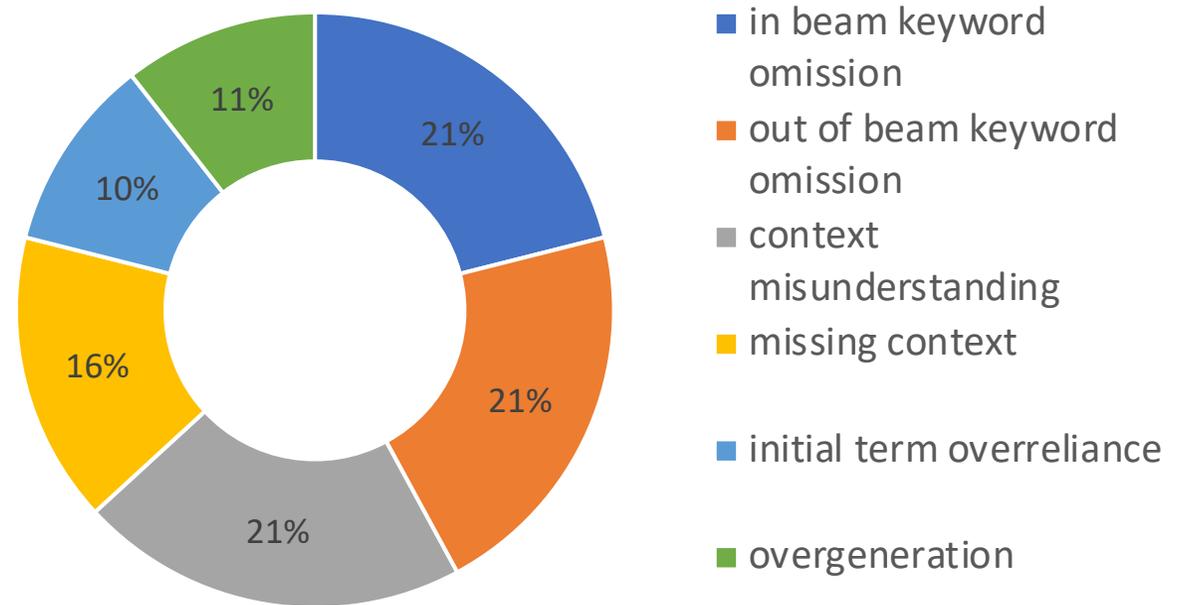
Adds long response result dependence

# CAsT Analysis: Year 2 Rewriters Break Down

Absolute  $\Delta$  between Manual and T5 Canonical by turn

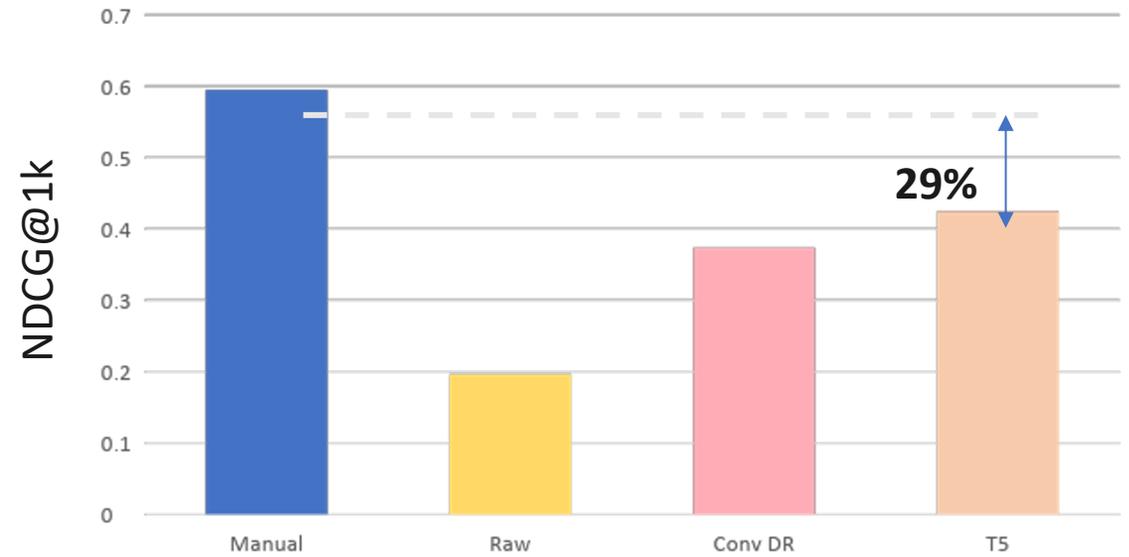
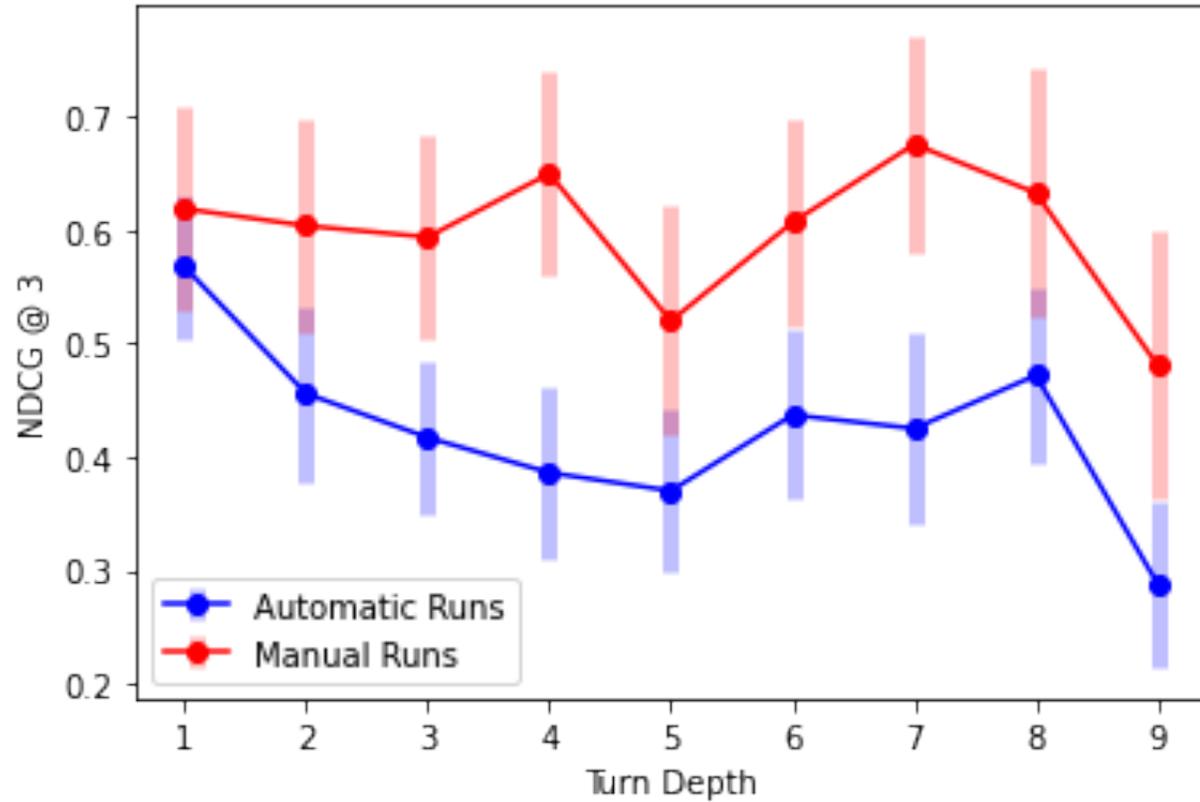


Types of errors for T5 queries with largest gaps

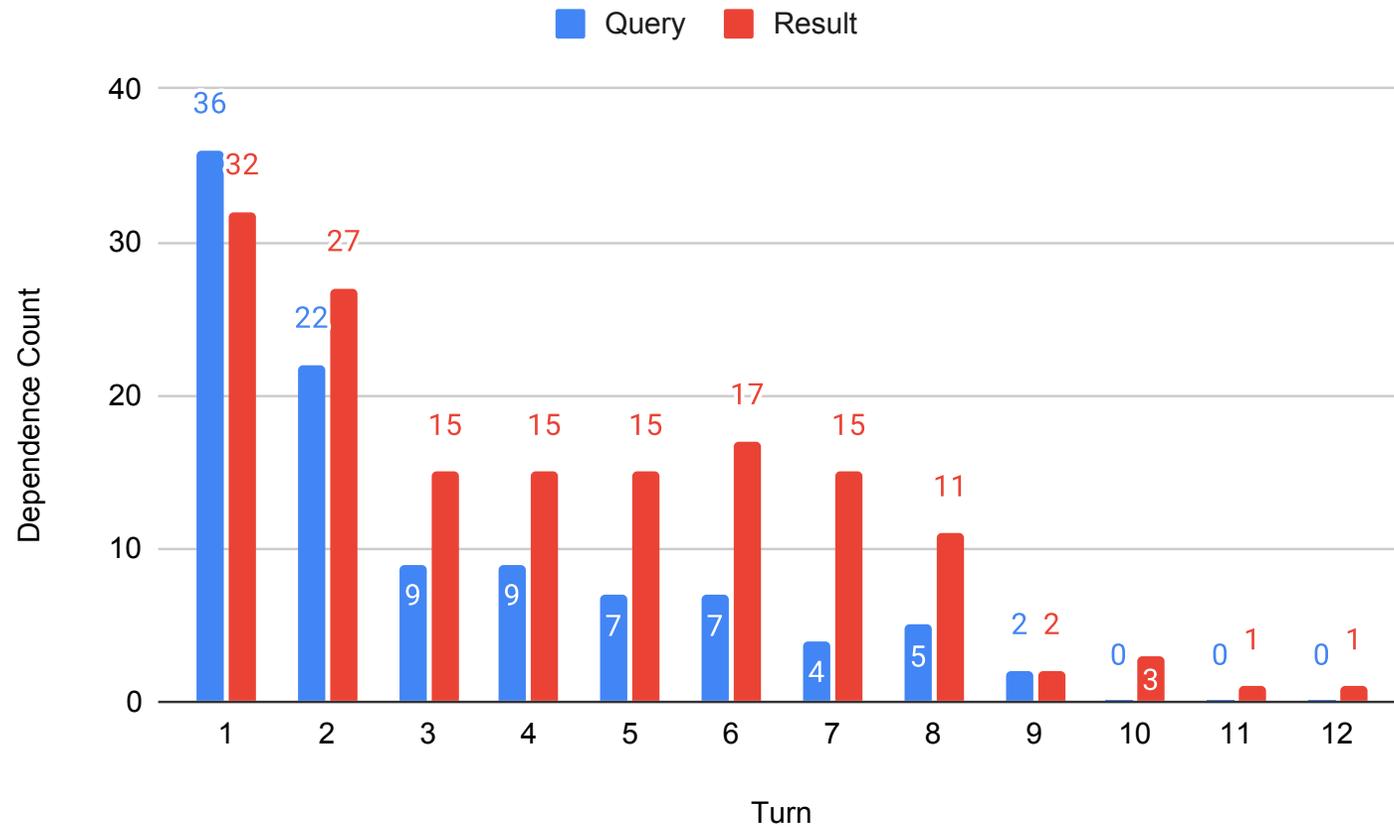


Rewrite failure can lead to catastrophic first phase retrieval failure.

# Results by conversation depth



# Context Dependence in CAsT Y3

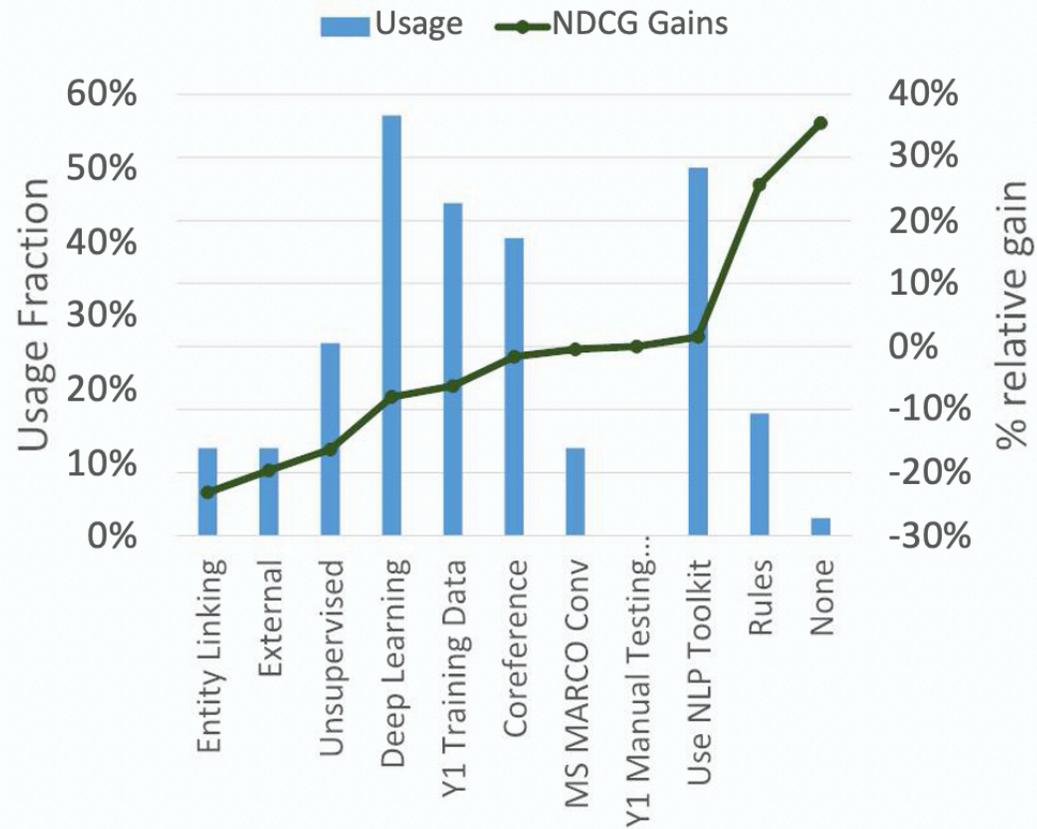


**‘Non-trivial’ dependence**

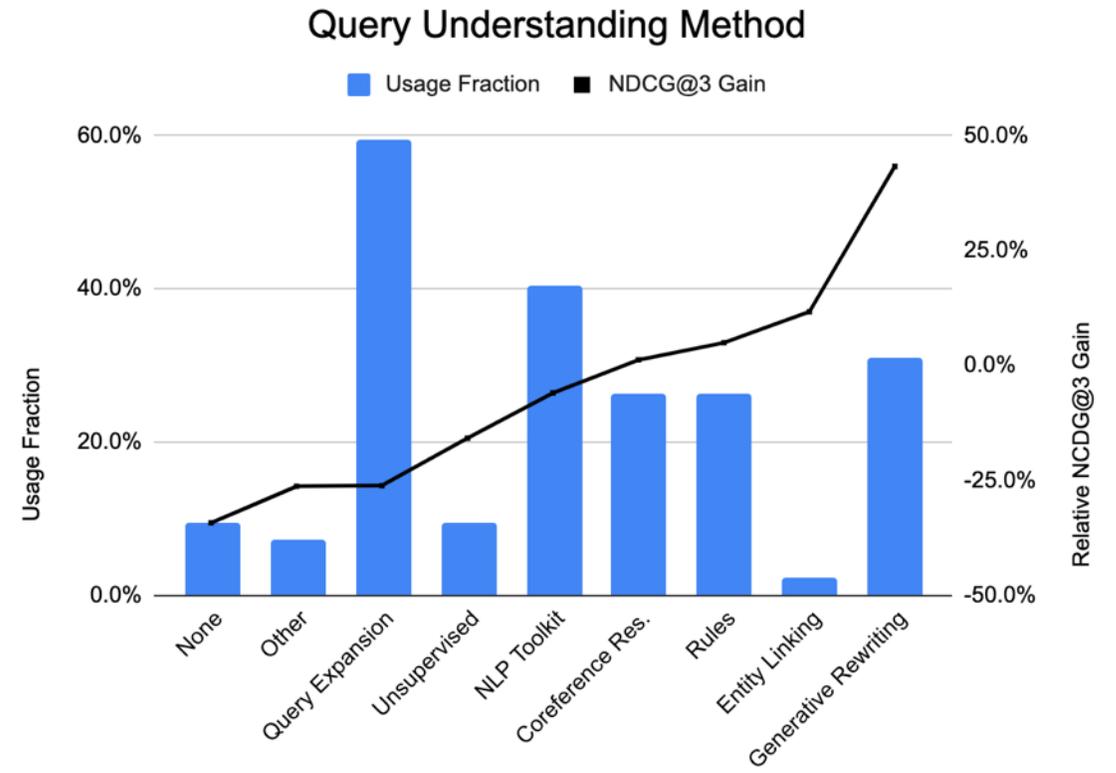
→ Dependence with  $> 1$  turn distance

# Putting the pieces together...

- Evolution of Conversational Language Understanding methods in CAsT



2019



2020

# Putting the pieces together



Effectiveness requires  
combining conversational tasks

- Conversational Query Rewriting
- Conversational Query Expansion ...



CLU may be performed in multiple phases in combination with  
with multi-stage pipelines

[Lin et al., 2021]



Need new methods for generating effective conversational  
representations beyond current methods

# Response ranking and generation

Chapter 5



# Ranking and Generation Tasks

---



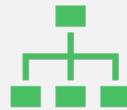
**Short answers**

ConvQA  
KG-ConvQA  
OR-ConvQA



**Long answers**

Conv PR  
Conv DR



**Semi-structured data**

Conv Task



**Recommendation**

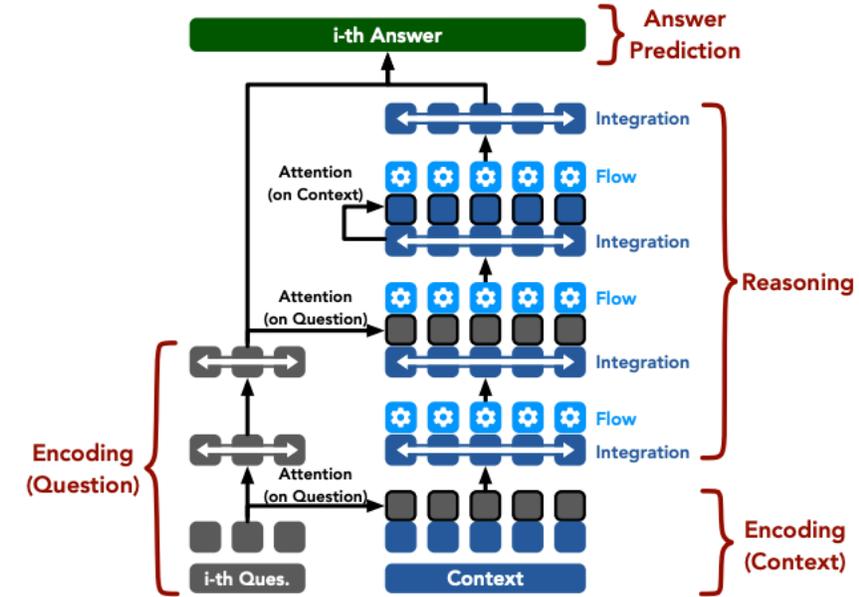
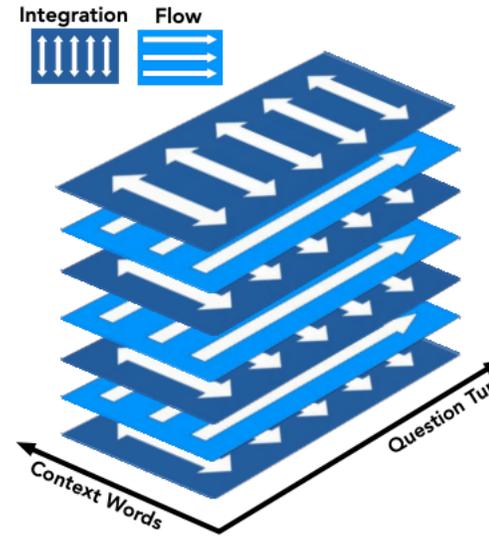
ConvRec

# ConvQA Evolution

- TREC 2004 – Contextual QA w/coref in a pipeline
- ConvQA – Machine Comprehension QA models w/context
  - DrQA, BiDAF++
- OR-ConvQA – Retrieve and Read Models
  - BERT-based – BERTserini (retriever) + BERT QA (reader)
  - Fusion-in-Decoder – Dense Passage Retriever + T5 reader [Izacard and Grave, 2021]

# Flow-based models for ConvQA

- Add layers for reasoning about turn-level evidence
- **Information Flow** layers
  - **Integration** - For each turn, create an integrated contextualized vector for every word
  - **FLOW** layer – Integrate token-level contexts across turns
- FlowDelta - Explicitly model information gain in conversation
- GraphFlow – Model FLOW using RGNNS



[Huang et al., 2019]

Q1: Who went to the farm? -> Q2: Why?

Billy went to the farm to buy some beef for his brother 's birthday . When he arrived there he saw that all six of the cows were sad and had brown spots . The cows were all eating their breakfast in a big grassy meadow . He thought that the spots looked very strange so he went closer to the cows to get a better look ...

Q2: Why? -> Q3: For what?

Billy went to the farm to buy some beef for his brother 's birthday . When he arrived there ... After Billy got a good look at the cows he went to the farmer to buy some beef . The farmer gave him four pounds of beef for ten dollars . Billy thought that ...

[Chen et al., 2020]

# ConvQA with Transformers - BERT

## AllHistory

- Append turns with separators

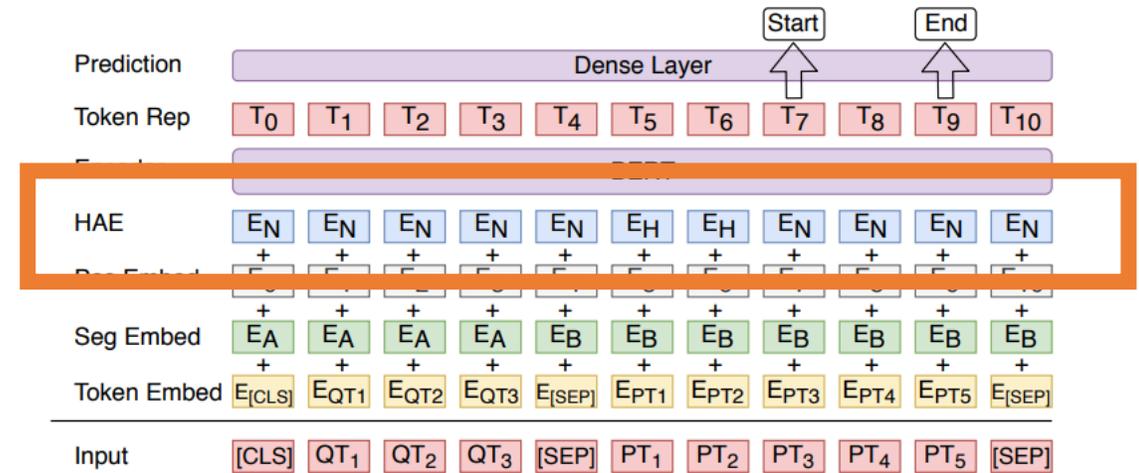
## History Answer Embedding

- Distinguishes between user/response in context using a binary embedding.
- 63.9% F1 on QuAC

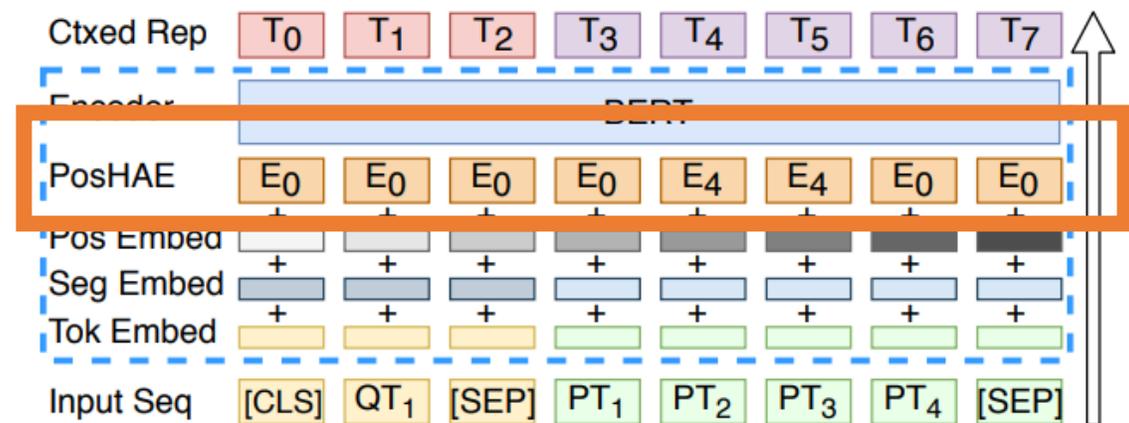
## Positional History Answer Embedding

- A shared **relative** position indicator.
- 64.7% on QuAC

[q1 [SEP] a1 [SEP] q2 [SEP] a2 [SEP] ...  
 [SEP] qn-1 [SEP] an-1 [SEP] qn.]  
 [Adlakha et al. 2021]



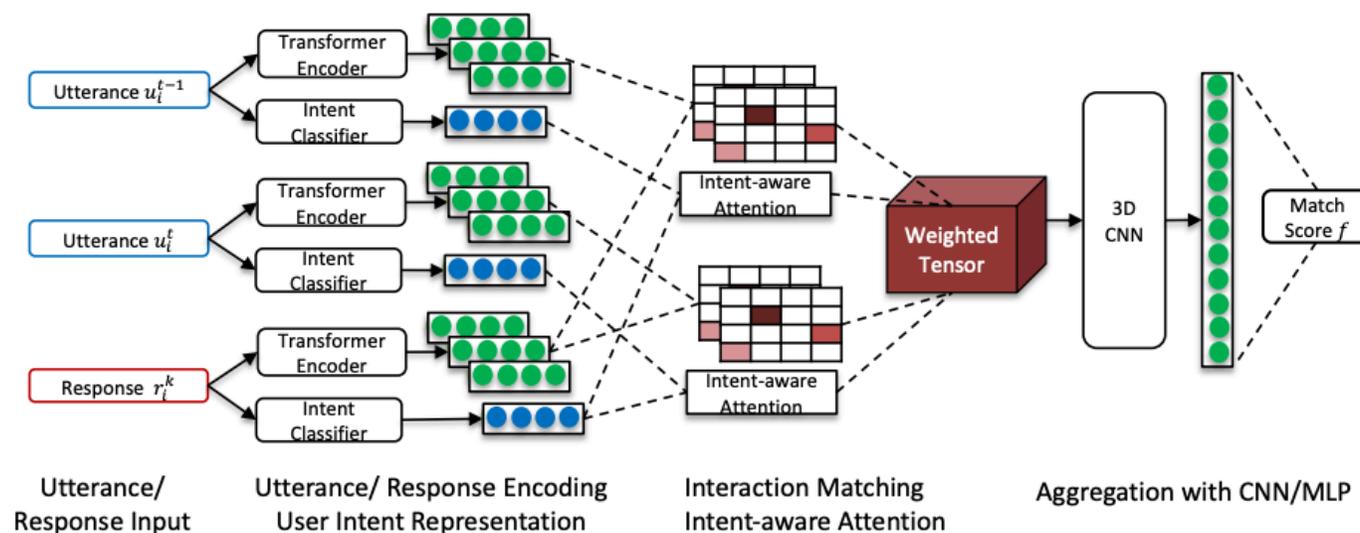
[Qu et al., 2019a]



[Qu et al., 2019b]

# Beyond Answers: Discourse-Aware Ranking

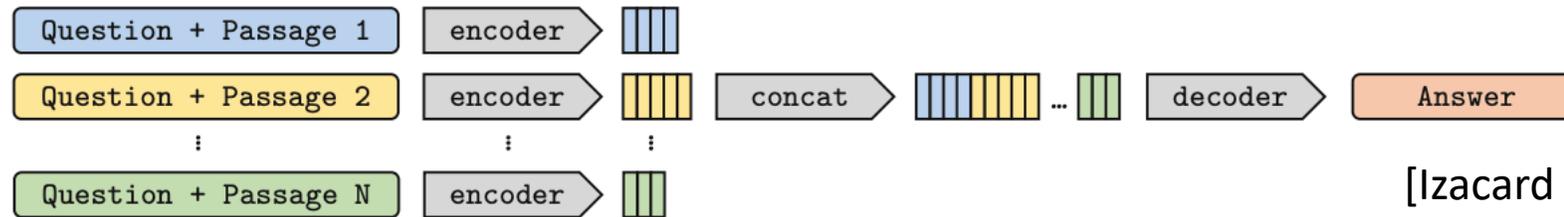
Code	Label	Description	Example	%
OQ	Original Question	The first question by a user that initiates the QA dialog.	If a computer is purchased with win 10 can it be downgraded to win 7?	13
RQ	Repeat Question	Posters other than the user repeat a previous question.	I am experiencing the same problem ...	3
CQ	Clarifying Question	Users or agents ask for clarification to get more details.	Your advice is not detailed enough. I'm not sure what you mean by ...	4
FD	Further Details	Users or agents provide more details.	Hi. Sorry for taking so long to reply. The information you need is ...	14
FQ	Follow Up Question	Users ask follow up questions about relevant issues.	Thanks. I really have one simple question -- if I ...	5
IR	Information Request	Agents ask for information of users.	What is the make and model of the computer? Have you tried installing ...	6
PA	Potential Answer	A potential answer or solution provided by agents.	Hi. To change your PIN in Windows 10, you may follow the steps below: ...	22
PF	Positive Feedback	Users provide positive feedback for working solutions.	Hi. That was exactly the right fix. All set now. Tx!	6
NF	Negative Feedback	Users provide negative feedback for useless solutions.	Thank you for your help, but the steps below did not resolve the problem ...	4
GG	Greetings/Gratitude	Users or agents greet each others or express gratitude.	Thank you all for your responses to my question ...	22
JK	Junk	There is no useful information in the post.	Emojis. Sigh .... Thread closed by moderator ...	1
O	Others	Posts that cannot be categorized using other classes.	N/A	1



# Generative OR-ConvQA with Transformers

## Fusion-in-Decoder

Retrieve (DPR) + Generative QA model (T5)



[Izacard and Grave, 2021]

Model	Question Rep	Dev		Test	
		EM	F1	EM	F1
Human		<b>40.2</b>	<b>70.1</b>	<b>40.3</b>	<b>70.0</b>
GPT-3		12.4	33.4	10.4	31.8
BM25 + DPR Reader	ORIGINAL	7.1	12.8	7.2	13.0
	ALLHISTORY	13.6	25.0	13.8	25.2
	REWRITES	15.4	32.5	15.7	31.7
BM25 + FiD	ORIGINAL	10.1	21.8	10.5	22.6
	ALLHISTORY	24.1	37.2	23.4	36.1
	REWRITES	24.0	41.6	24.9	41.4
DPR Retriever + DPR Reader	ORIGINAL	4.9	14.9	4.3	14.9
	ALLHISTORY	21.0	43.4	19.4	41.1
	REWRITES	17.2	36.4	16.5	35.2
DPR Retriever + FiD	ORIGINAL	7.9	21.6	7.8	21.4
	ALLHISTORY	<b>33.0</b>	<b>55.3</b>	<b>33.4</b>	<b>55.8</b>
	REWRITES	23.5	44.2	24.0	44.7

Results from [Adlakha et al., 2021]

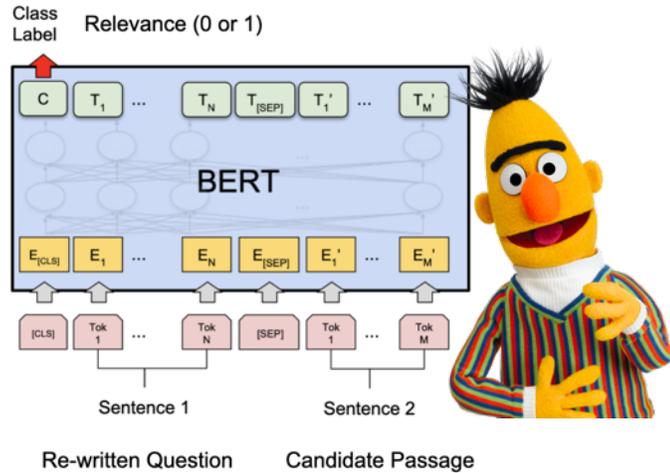
Table 5: Overall performance of all model variants on TOPIOCQA development and test set

- Rewriting is particularly important in initial retrieval; less important for Reader model

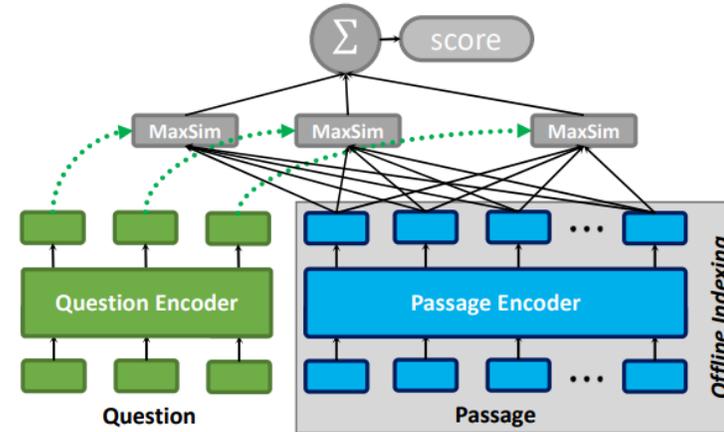
# ConvPR and ConvDR

# ConvPR with Rewritten queries

## BERT for neural ranking

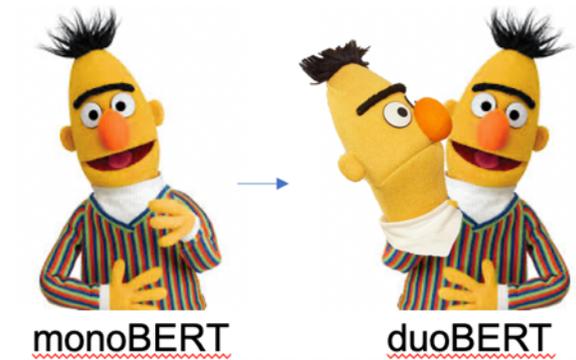


## CoBERT



[Khattab et al. 2020]

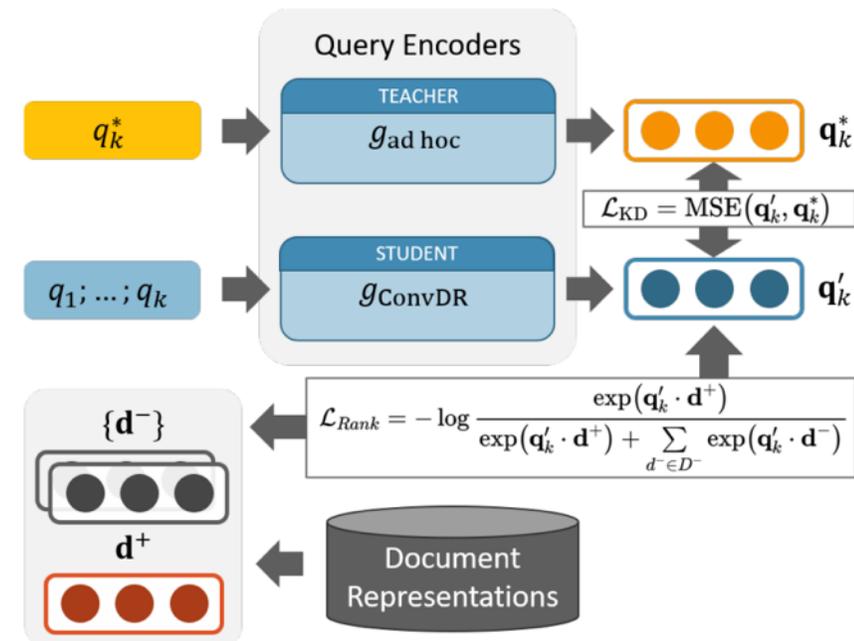
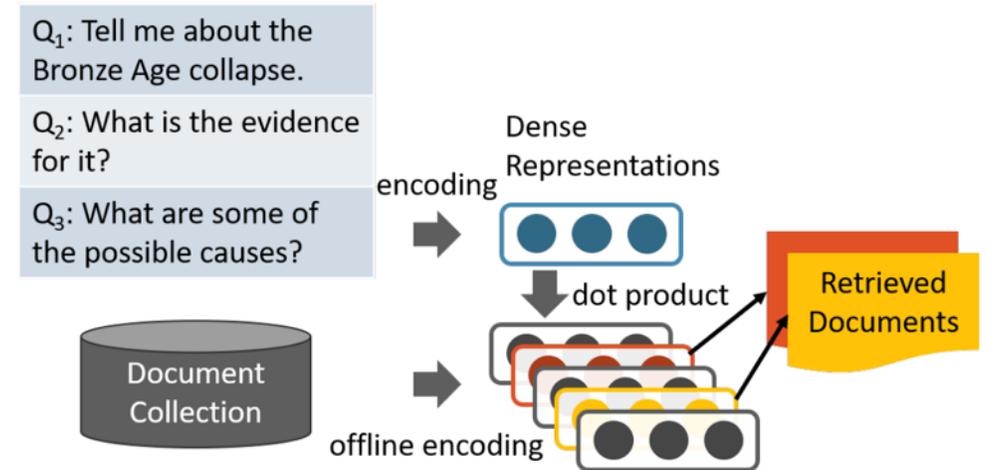
## Multi-stage fusion with neural ranking



T5

# Conversational Dense Retrieval

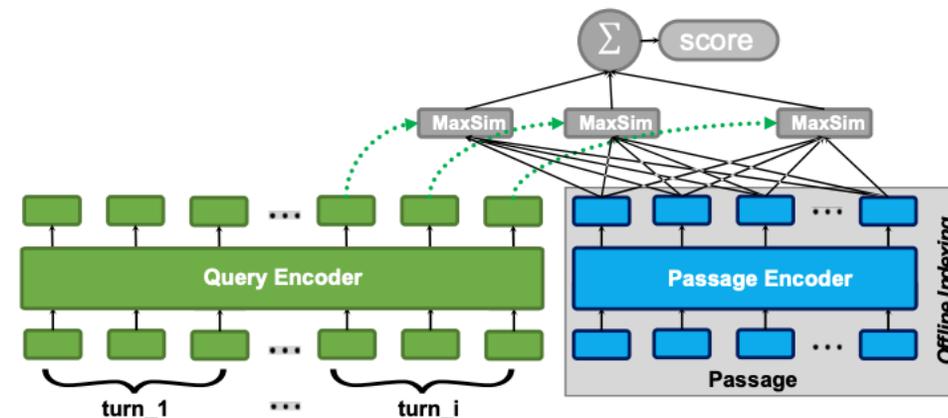
- End-to-end retrieval with dense vector representation to retrieve content
- Can be more effective than rewriting + dense retrieval
- Learned from oracle manual query representations



[Yu et al., 2021]

# Zero-shot Conversational Contextualization

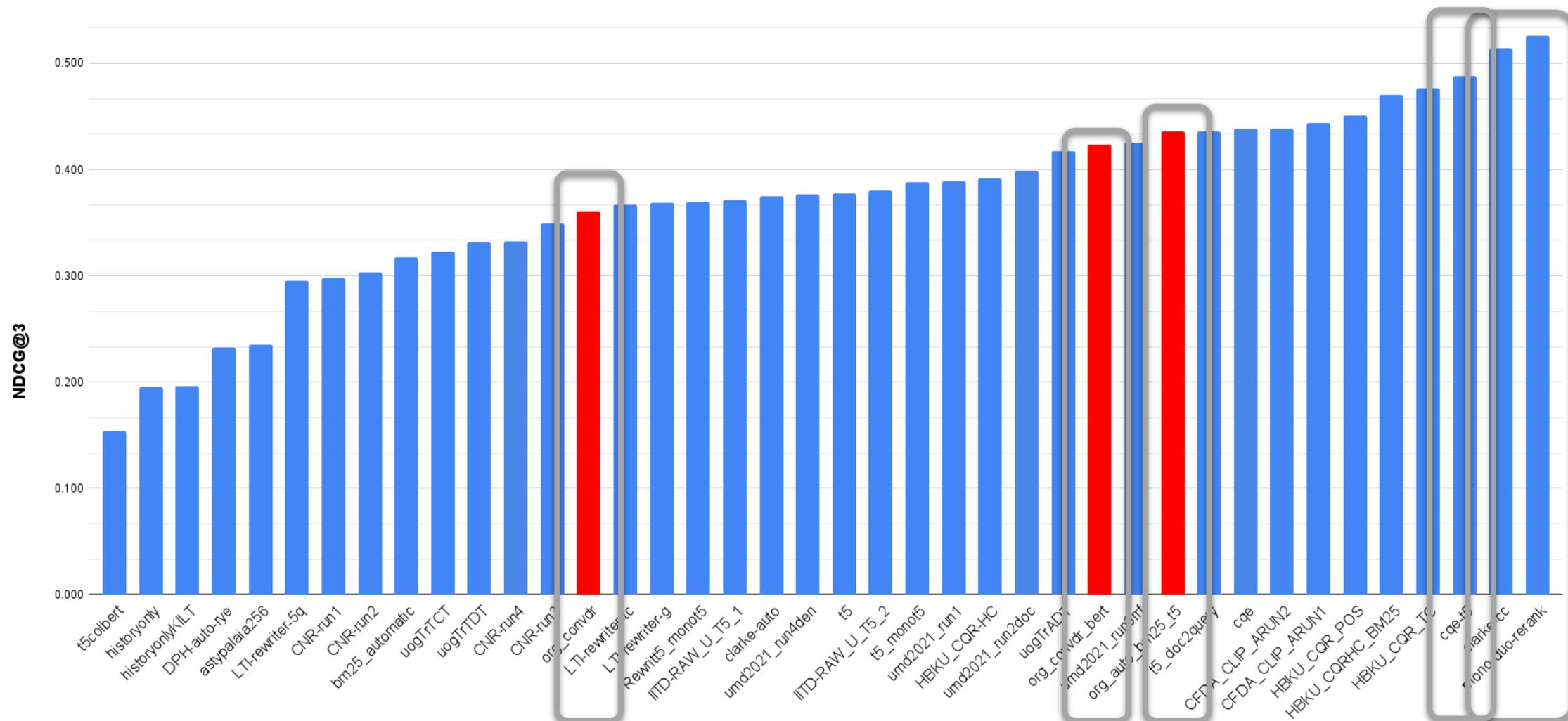
- *ZeCo<sup>2</sup>* -A variant of ColBERT for ConvPR
  - Contextualizes all embeddings within the conversation
  - Matches only the contextualized terms of the user's last question



base-retriever	variant	zero-shot	CAsT'19		CAsT'20		CAsT'21	
			NDCG@3	R@100	NDCG@3	R@100	NDCG@3	R@100
ColBERT	last-turn <sup>a</sup>	✓	0.214	0.157	0.155	0.124	0.140	0.154
	all-history <sup>b</sup>	✓	0.190	0.165	0.150	0.166	<b>0.237</b>	0.265
	<i>ZeCo<sup>2</sup> (ours)</i>	✓	0.238 <sup>b</sup>	<b>0.216</b> <sup>a,b,c</sup>	<b>0.176</b> <sup>b</sup>	<b>0.200</b> <sup>a,b,c</sup>	0.234 <sup>a</sup>	<b>0.267</b> <sup>a</sup>
	human		0.430	0.363	0.443	0.408	0.431	0.403
ConvDR [31]	zero-shot <sup>c</sup>	✓	<b>0.247</b>	0.183	0.150	0.150	–	–
	few-shot		0.466	0.362	0.340	0.345	0.361	0.376
	human		0.461	0.389	0.422	0.454	0.548	0.451

[Krasakis et al., 2022]

# CAsT Year 3 Results



# Evaluation

- Topic vs turn-level effectiveness
  - All topics are approximately the same length in turns
- Turn-level effectiveness
  - Primary measure: NDCG@K with  $K = 3$
  - Mean reciprocal rank, MAP@1000
    - Binary relevance threshold (2)
- **Ongoing work:** Conversation-centric evaluation
  - Responses should be 'consistently' relevant
  - Model abandonment 'hang-ups'
  - Beyond evaluating responses: questions

# QA & Ranking Conclusion

---

- Models evolved from **closed short answer** ConvQA models towards **open-retrieval models** that **generate short (and long)** answer responses
- Most effective methods have **pipelines** involving multiple components: **query rewriting**, query expansion, **dense retrieval**, **multi-pass re-ranking**, and result fusion.
- A common pattern is **stacking models** to add richer conversational modeling capability (Flow, 3D-CNNs, RGNNs).
- **Models, datasets, and evaluation** need to evolve to handle **richer forms of interactions** beyond questions and answers.

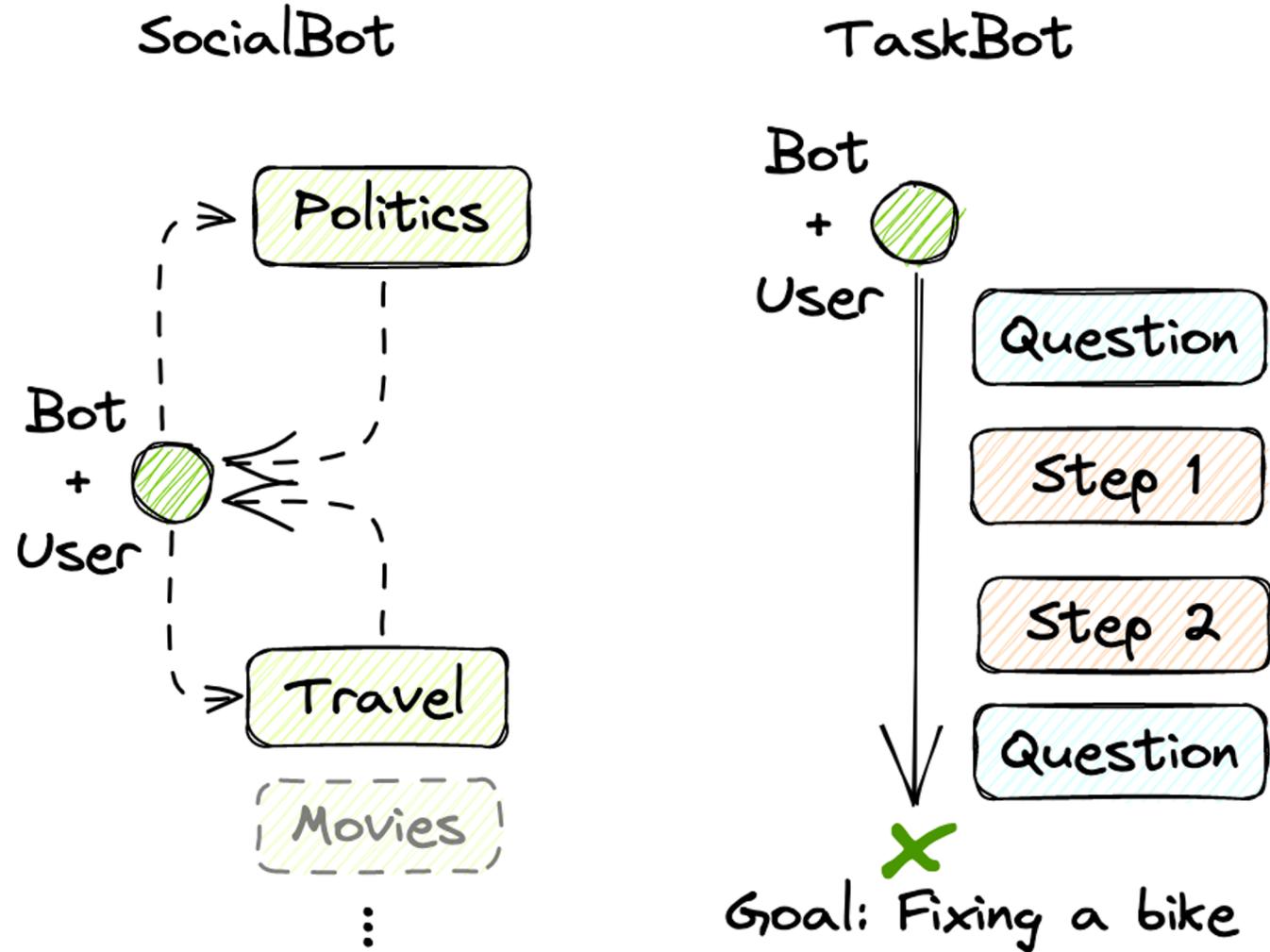
UoG GRILL Bot



Taskbot Challenge



# Real-World Task Assistants



## 2022 AlexaPrize

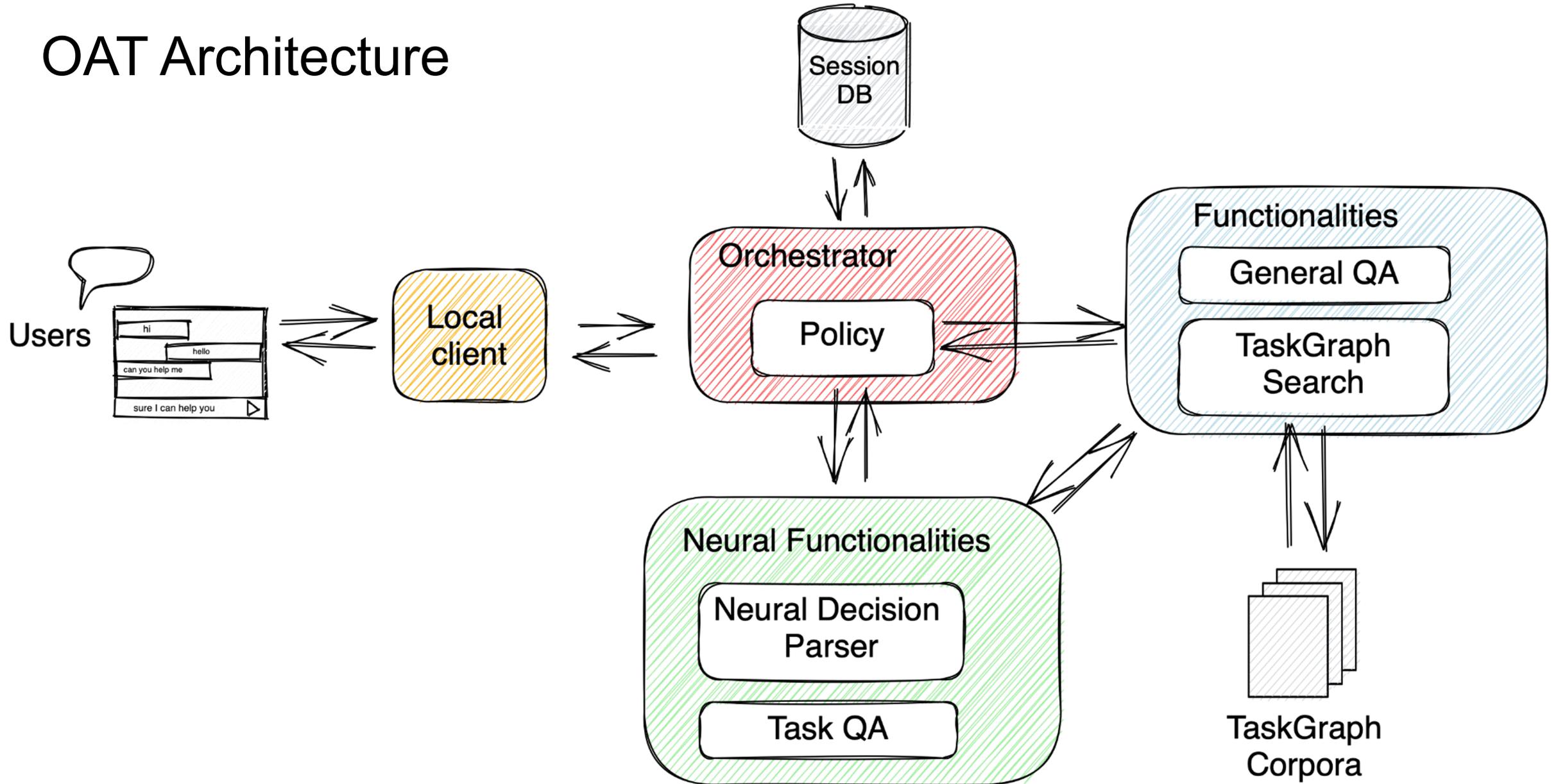
### Cooking



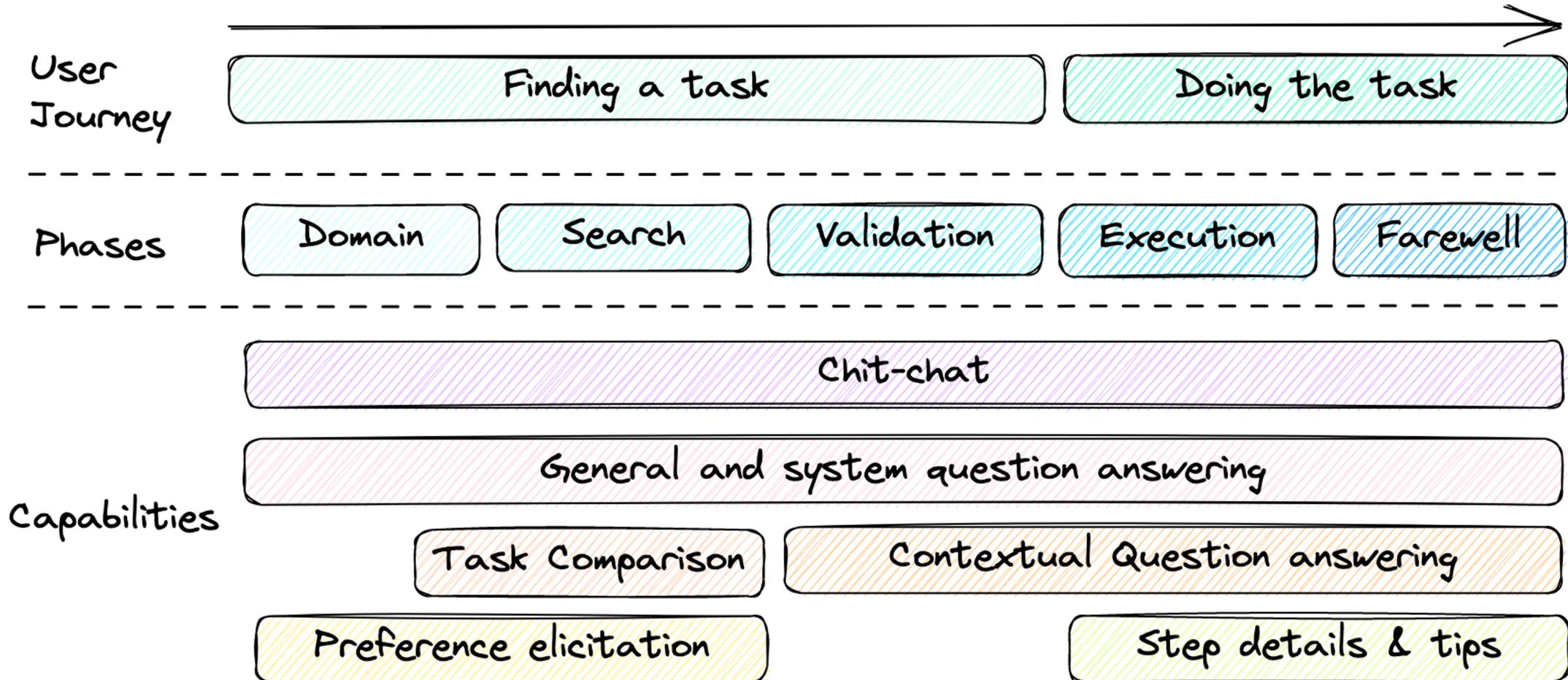
### Crafts & DIY



# OAT Architecture



# Conversation Flow: A Phased Approach



# Multi-modal task conversations

"Soup is my favourite thing to cook!  
What's yours?"

"desserts"

"Great choice! I have 2 recommendations:"

- Apple Pie
- Toffee Pudding

"Hmmm... Pie... convince me!"

"It's by Chef Angela Hartnett's mom,

And Bramley apples are in season now!"

⋮

⋮  
"Sold! But I don't have butter..."

"Sunflower oil, works well with pastries"

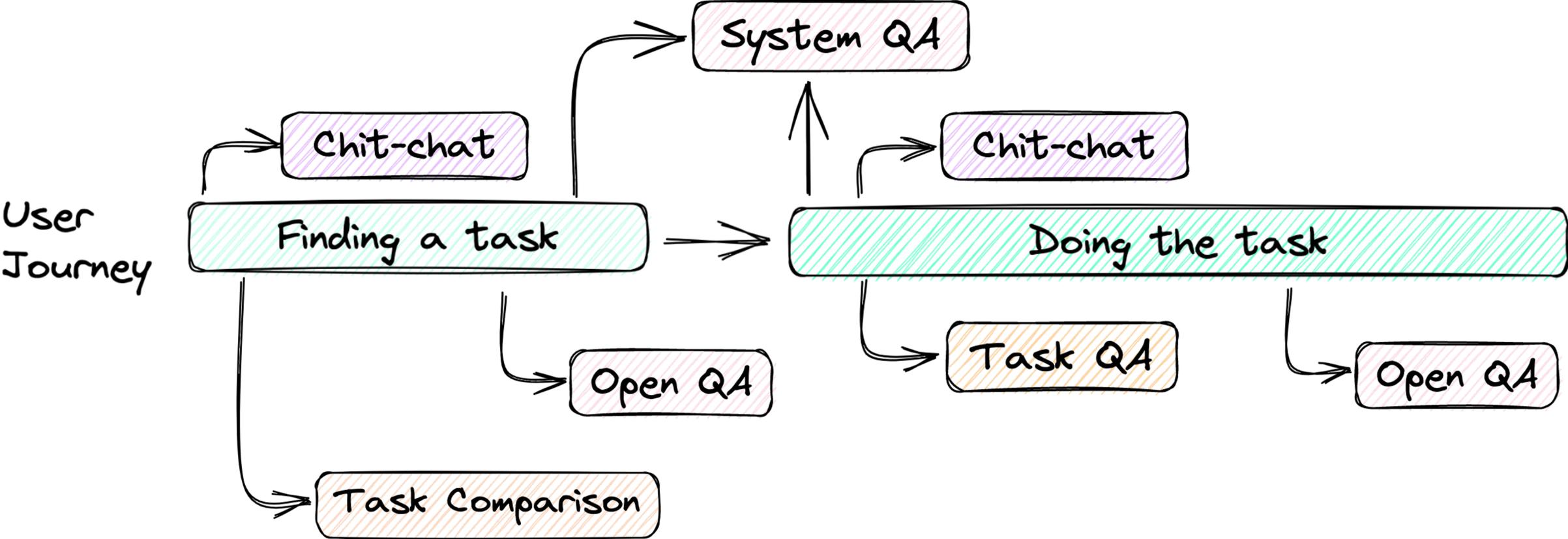
↑  
Stateful Question Answering

Task search

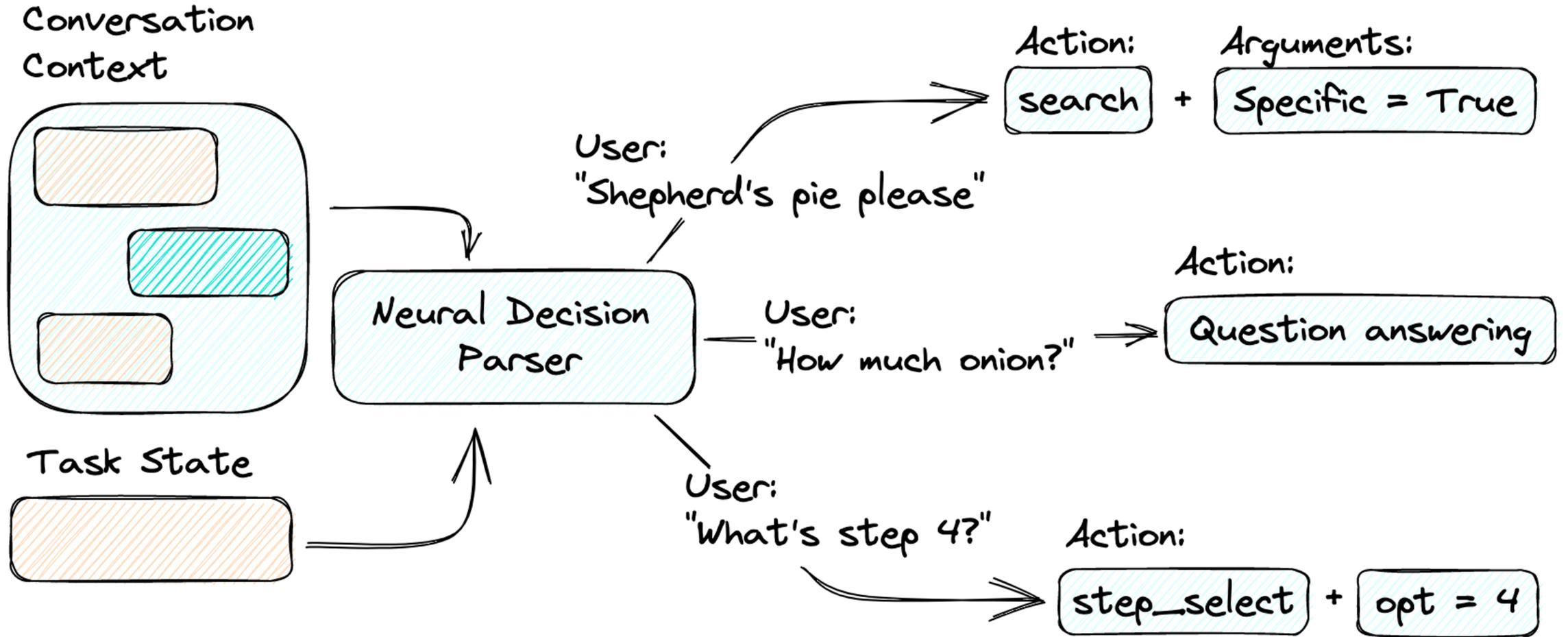
Contextual  
Knowledge  
Addition



# Conversational Question Answering

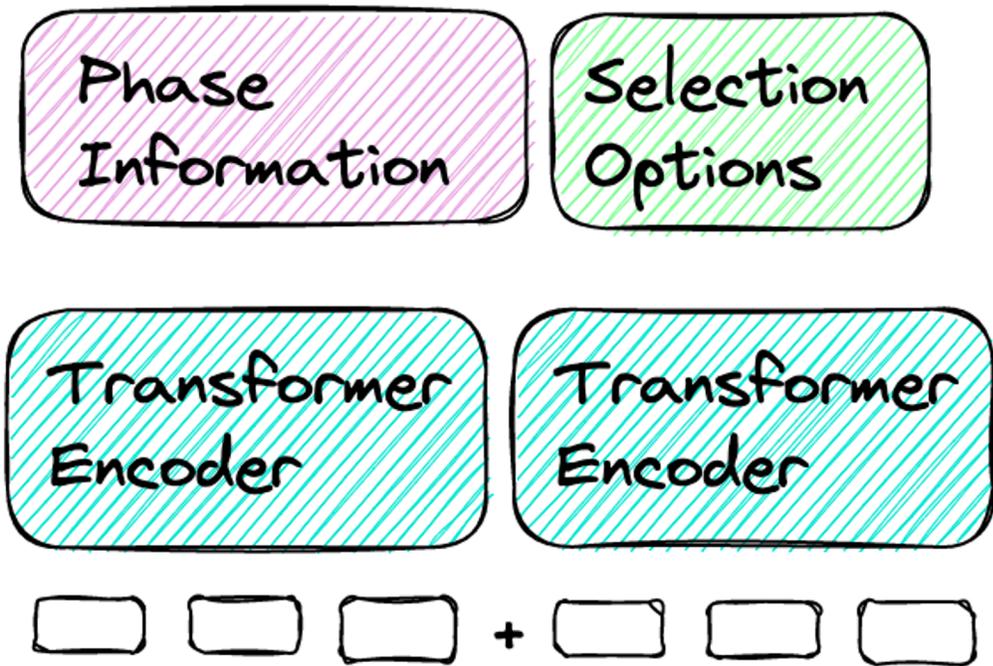


# Code Generation for State Management

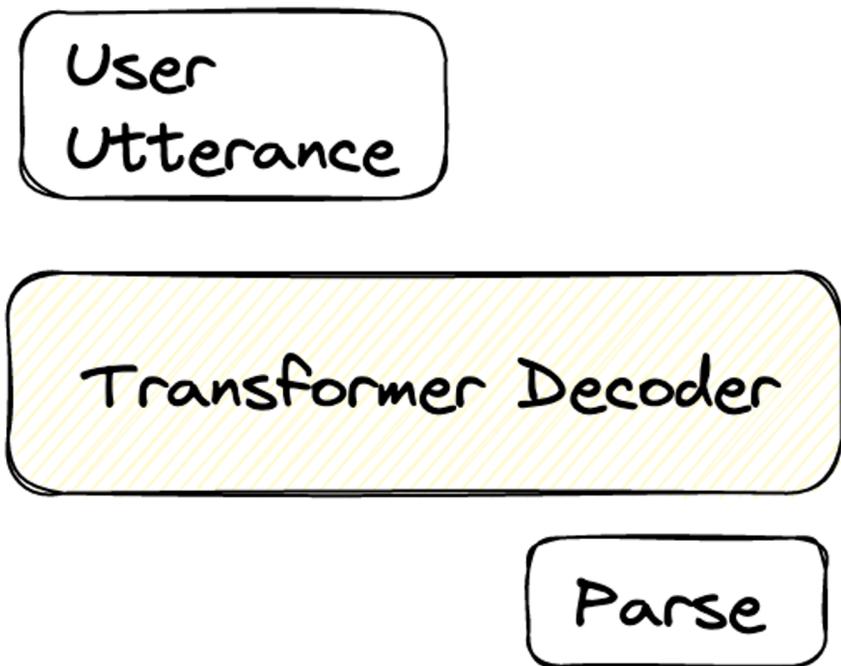


# Neural Decision Parser Architecture

Stateful context



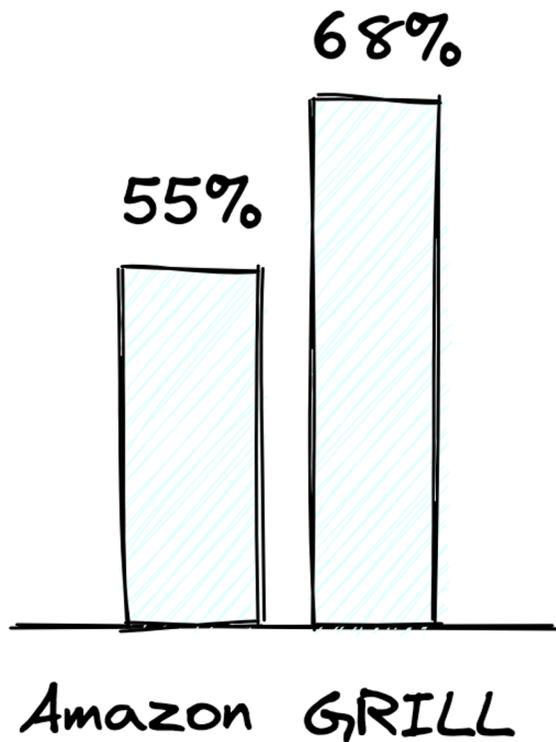
Fixed Autorregressive decoding



# Understanding Model Effectiveness

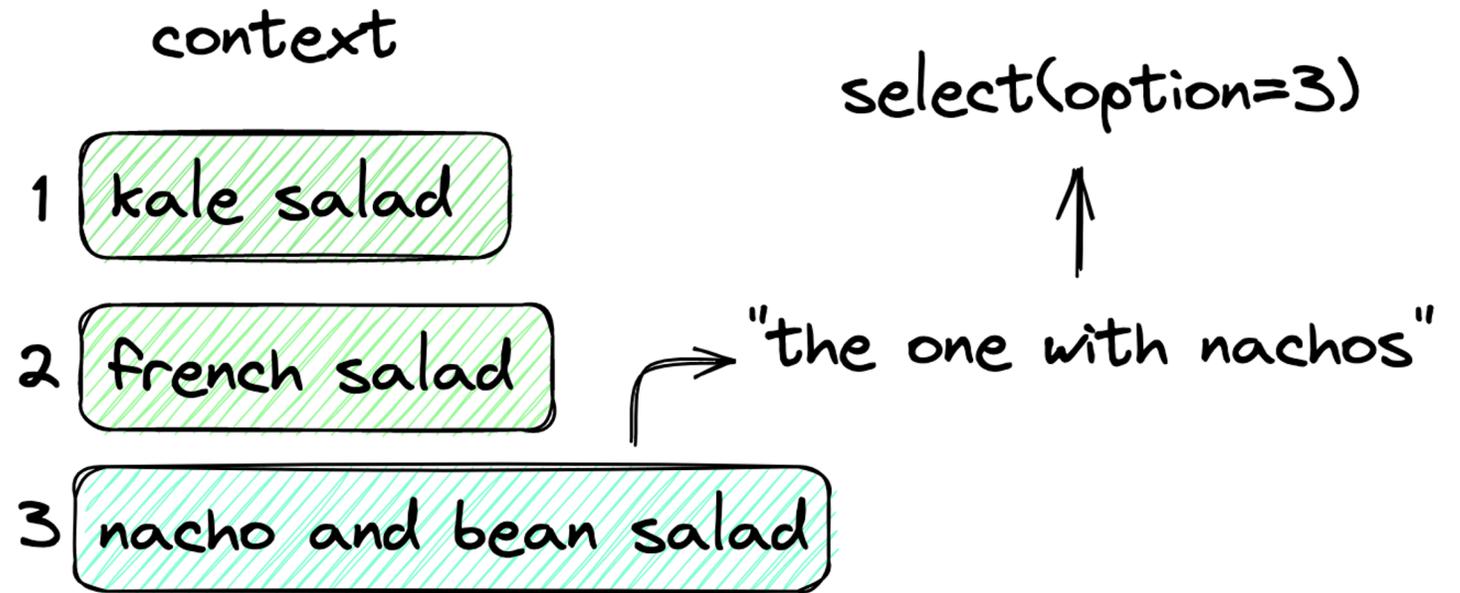
predicted intents

---



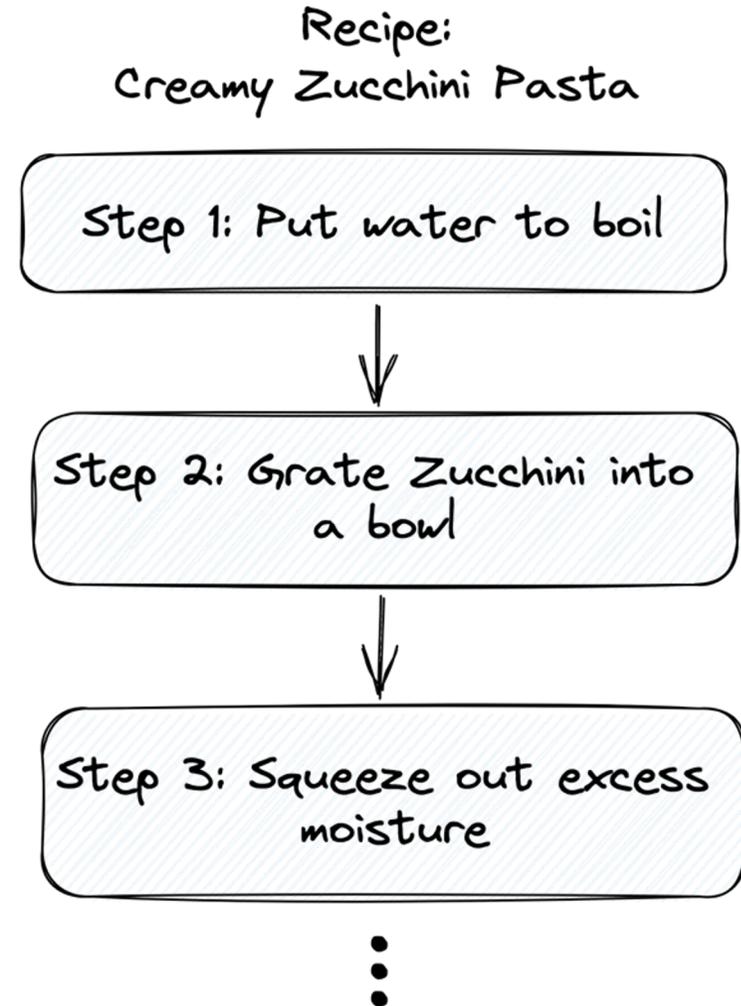
argument prediction

---

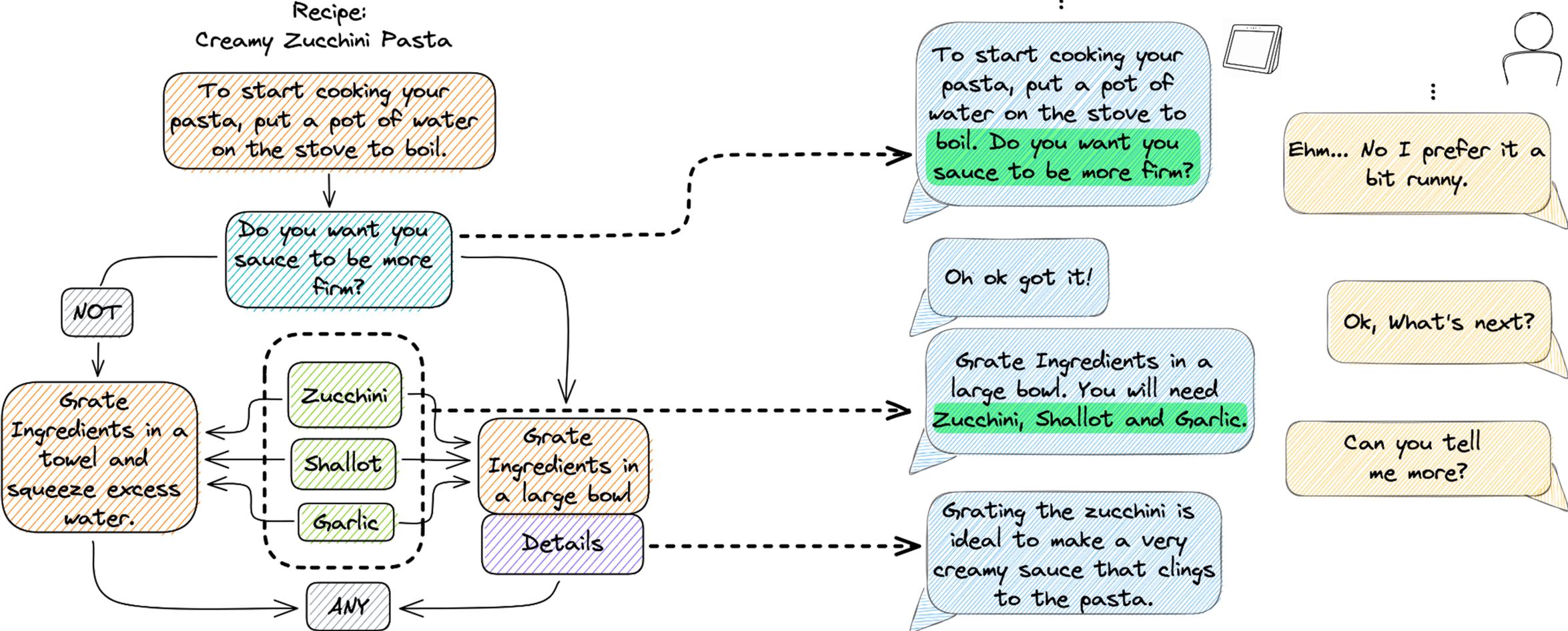


# Standard Task Representation

- No system initiative
- Lacking personalisation
- Dry speech interactions



# Task Graphs: A Live Example



# Task-based Conversational Challenges

...



Extract and ground task materials

Existing linkers have low recall



Adapt task to user preferences and constraints

Material substitution



Support task QA

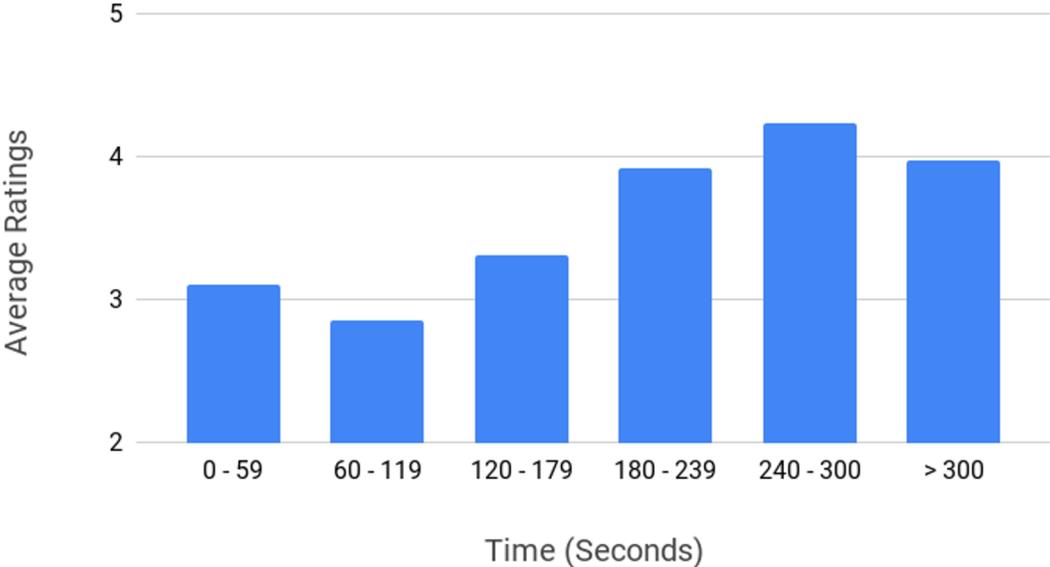
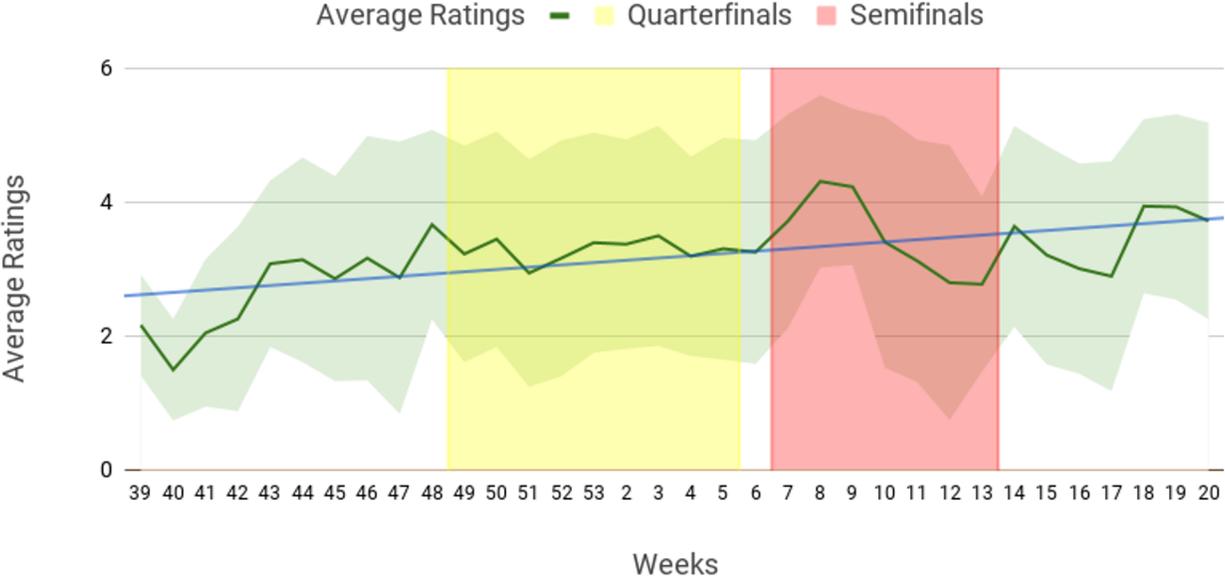
What is kosher salt?



Support step-level explainers

Multimodal images and videos

# Performance over the year

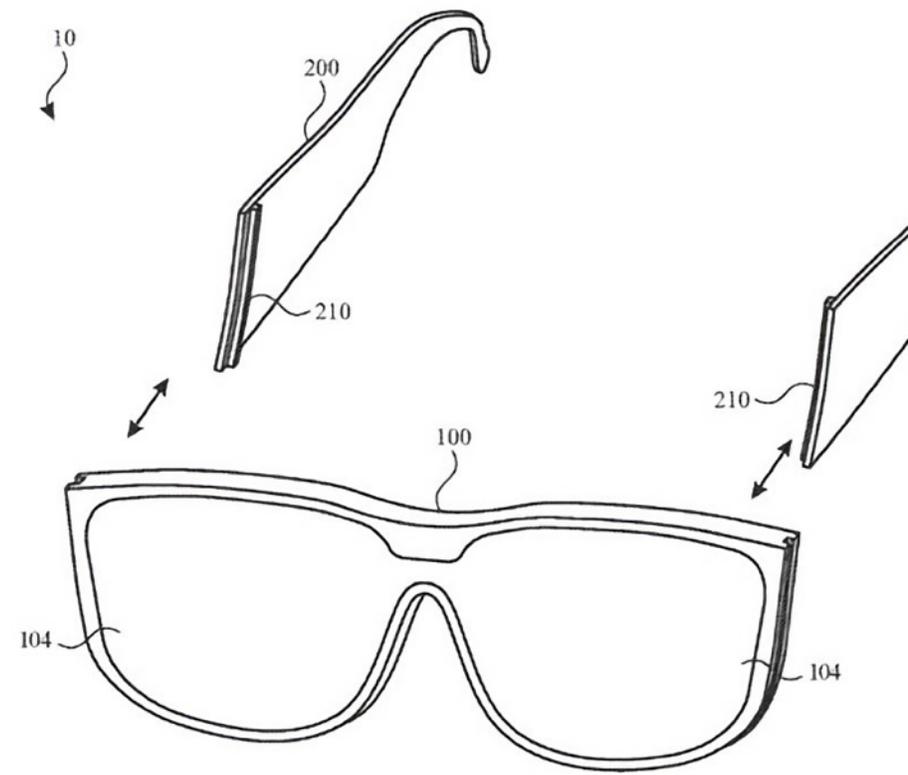




Looking ahead



Slide from Croft, SIGIR 2019



Future: AR & IoT



# Many open challenges...



**Improved use of result context**

Deeper result understanding



**Mixed-initiative and feedback**

Rank a question to answer instead of a result



**Flexible responses**



**Beyond-turn based relevance evaluation**

Move towards “usefulness” of interactions



**Bridging the gap with chat/assistant systems**

# CAsT Y4 2022

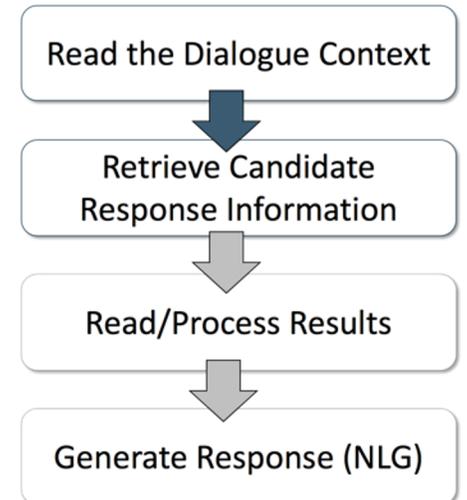
**treccast.ai**

Twitter: @treccast

Google Group:  
<https://groups.google.com/d/forum/trec-cast>

August 30, 2022!

- Candidate 'source' passage/doc retrieval
- **Extract/Generate** responses
  - Grounded to retrieved passages
- Mixed-initiative sub-task
- New organizers!
  - Johanne, Leif, Mohammed, Paul, Svitlana



# Summary

- Conversational Search is an important topic in AI
  - Benchmark for ‘machine reading’ and understanding
  - Distinct from web search and interactive retrieval
- Conversational Query Rewriting is a key building block
  - Sequence-to-sequence models are more effective than traditional NLP pipelines
  - Surprisingly ‘easy’ to fine-tune on small datasets
  - Fail in key areas: hallucination, long-term relationships, scaling to long results
- The goal of meaningful ‘interaction’ for CIS remains challenging
  - Requires new evaluation approaches and measures
- Conversational Search is still a new sub-field
  - Everything we did for ‘search’ is new again!
  - State and interaction changes everything



A new generation...

---