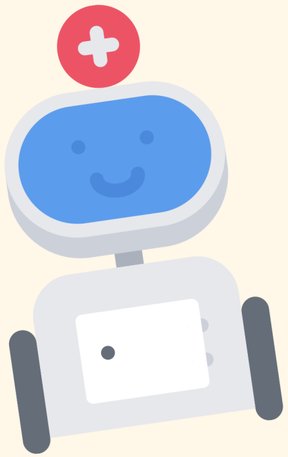
THE AI DOCTOR IS IN:

Towards Intelligent Automated Support for Healthcare Tasks using Natural Language Processing

July 18, 2022
NLG4Health @ INLG
Waterville, Maine, USA



Natalie Parde

Assistant Professor
Department of Computer Science

Co-Director
Natural Language Processing Laboratory

University of Illinois Chicago

AI is increasingly prevalent in healthcare, and offers many opportunities to support patients, clinicians, and caregivers.

- Medical diagnosis
- Health information delivery
- Patient monitoring
- Clinical note processing
- And many more!

Within NLG specifically....

Key healthcare tasks include:

- Healthcare dialogue agents
- Medical summarization
- Clinical decision support
- Clinical note generation
- Health coaching
- Behavior change coaching
- Personalized reporting

15.00 – 15.20 In-Domain Pre-Training Improves Clinical Note Generation from Doctor-Patient Conversations – Colin A. Grambow, Longxiang Zhang, Thomas Schaaf

15.20 – 15.40 Towards Development of an Automated Health Coach – Leighanne Hsu, Rommy Marquez Hernandez, Kathleen McCoy, Keith Decker, Ajith Kumar Vemuri, Greg Dominic

15.40 – 16.15 Coffee/Tea break

16.15 – 16.30 DrivingBeacon: Driving Behaviour Change Support System ConAsidering Mobile Use and Geo-information -Jawwad Baig, Guanyi Chen, Chenghua Lin, Ehud Reiter

16.30 – 16.45 Personalizing Weekly Diet Reports -Elena Monfroglio, Luca Anselma, Alessandro Mazzei

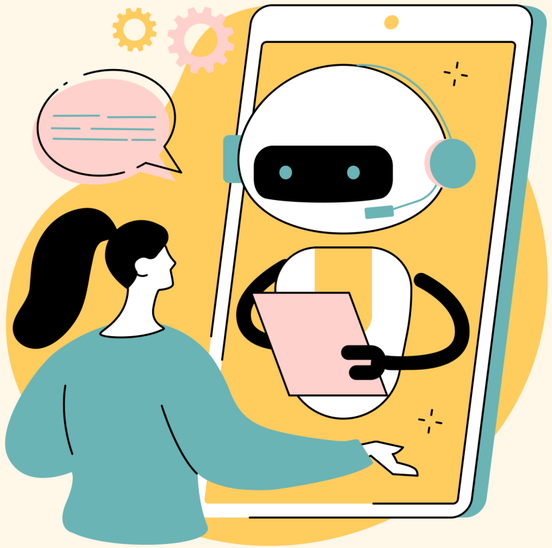
16.45 – 17.00 LCHQA-Summ: Multi-perspective Summarization of Publicly Sourced Consumer Health Answers – Abari Bhattacharya, Rochana Chaturvedi, Shweta Yadav

OUTLINE

Focus: Task-oriented healthcare dialogue systems, following a recent survey conducted on this topic.

- Current state of task-oriented dialogue systems in the healthcare domain
- Recommendations for the future
- Ongoing work in this area

Dialogue systems have a daily presence in our lives.



They may take different forms:

- Chatbots conduct unstructured conversations in open domains
- Task-oriented dialogue systems help users complete tasks in a specific domain

Task-oriented dialogue systems have been adopted by growing numbers of patients, caregivers, and clinicians.

However, there is a gap between cutting-edge, foundational work in dialogue systems and prototypical or deployed dialogue agents in healthcare settings.

This restricts the potential benefits of fundamental research!

We comprehensively analyze task-oriented healthcare dialogue systems with two underlying objectives.

1

Explore how these systems have been employed

2

Map their characteristics, shortcomings, and opportunities for follow-up work

How did we do this?

Our primary contributions include:

1

We search through 4070 papers from well-known venues and identify 70 fitting our inclusion criteria

2

We analyze these systems based on many factors

3

We identify common limitations across systems, and offer practical suggestions for addressing these

Background

Existing surveys of healthcare dialogue systems have been conducted by medical and clinical communities to examine real-world use of deployed systems.

- Vaidyam et al. (2019) surveyed 10 systems in mental health settings
- Laranjo et al. (2018) and Kearns et al. (2019) surveyed 17 and 46 systems in general healthcare settings, respectively

Since these articles were written for a medical audience and focused on healthcare issues and impact, they covered few articles from technical venues.

Background

A couple recent surveys have sought to conduct investigations with greater technical depth:

- Montenegro et al. (2019) and Tudor Car et al. (2020) surveyed 40 and 47 systems, respectively
- These surveys used a wide search breadth, examining conversational agents in the healthcare domain at a general level (e.g., including both chatbots and task-oriented dialogue systems)

We reviewed 70 papers focusing on a specific class of conversational agents (task-oriented dialogue systems), allowing us to apply a more thorough technical taxonomy when analyzing papers.

Search Criteria and Screening

Source Databases

- ACM
- IEEEE
- ACL Anthology
- AAAI Digital Library

Inclusion Criteria

- Main focus is on the technical design or implementation of a task-oriented dialogue system
- System is designed for health-related purposes
- Article is not dedicated to one specific module of the system's architecture

Search Criteria and Screening

Initial Search

Populate initial list of papers using predefined search query

Title Screening

Read the titles for the initial list of papers, keeping those that satisfy inclusion criteria

Abstract Screening

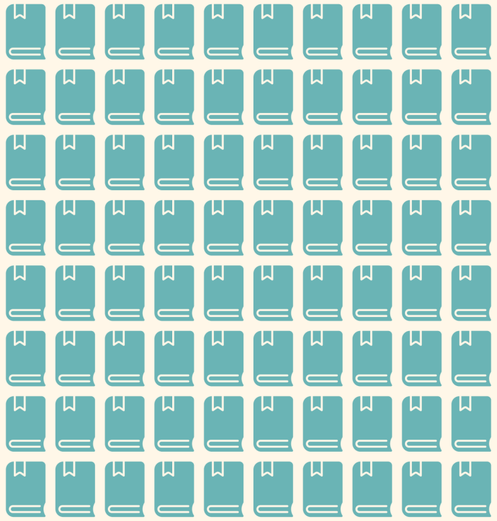
Read the abstracts of the remaining papers, keeping those that satisfy inclusion criteria

Final Screening

Read the full text for the remaining papers, keeping those that satisfy inclusion criteria

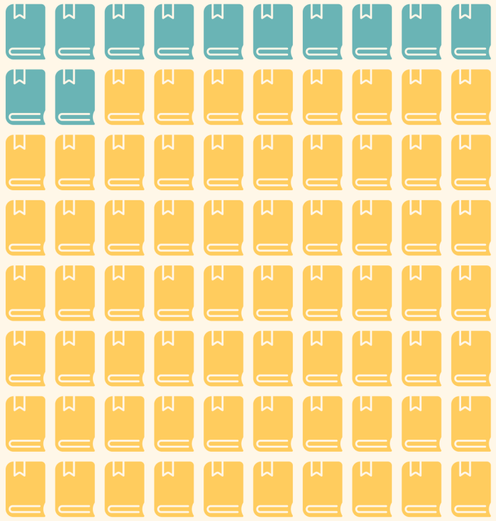
Search Criteria and Screening

Initial Search



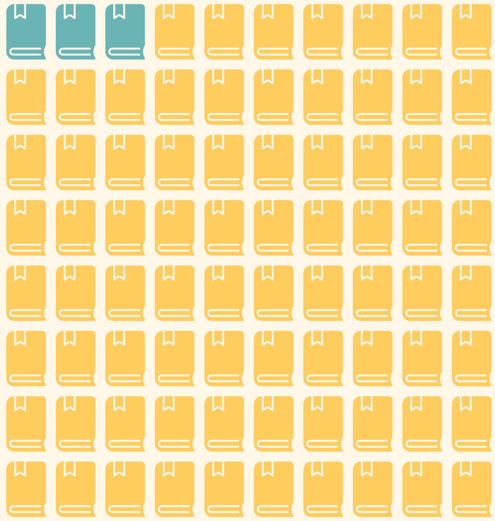
4070 papers

Title Screening



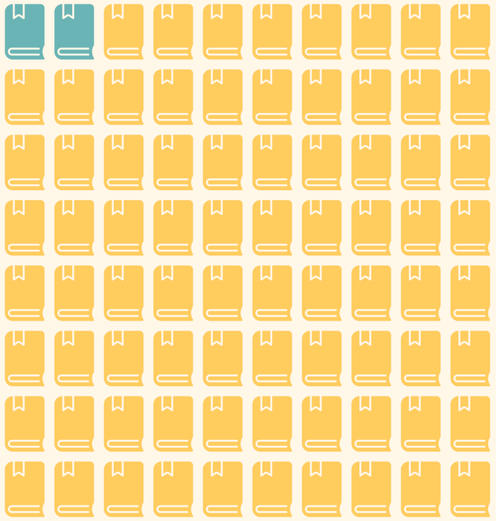
585 papers

Abstract Screening



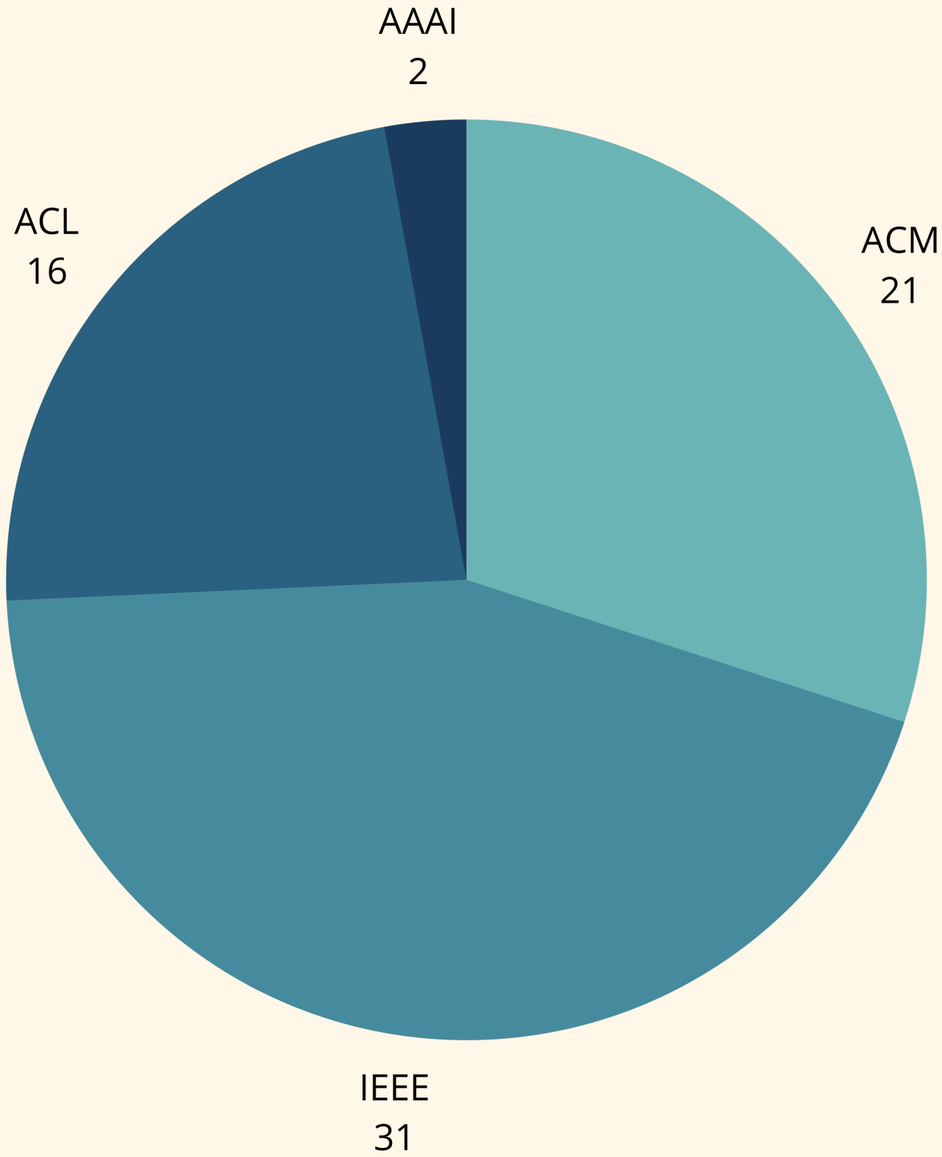
110 papers

Final Screening



70 papers

Paper Distribution



Analysis of Included Papers

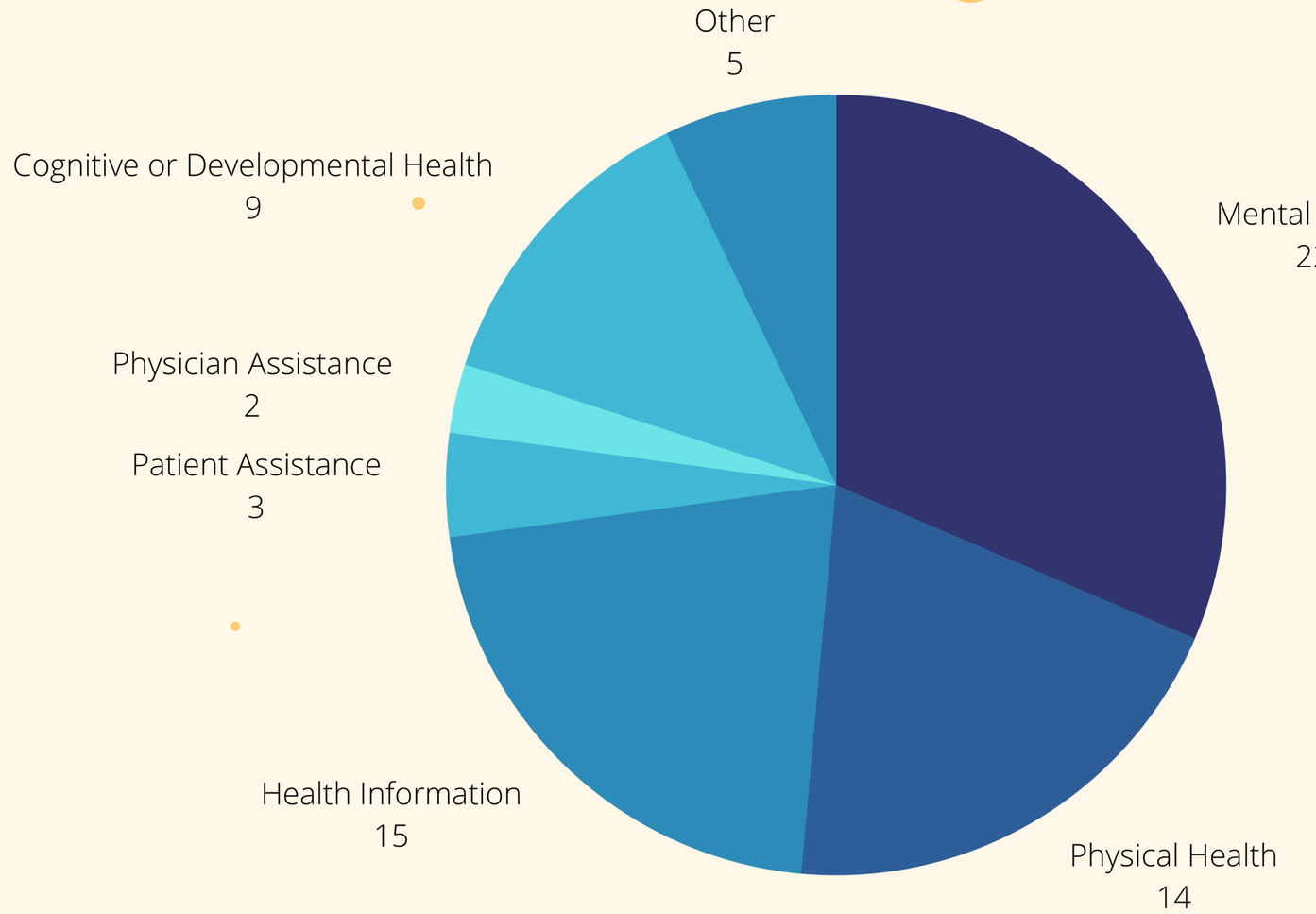
Each included paper was annotated according to its:

- Research domain
- System objective
- Target audience
- Language
- Architecture
- Design
- Dataset
- Evaluation

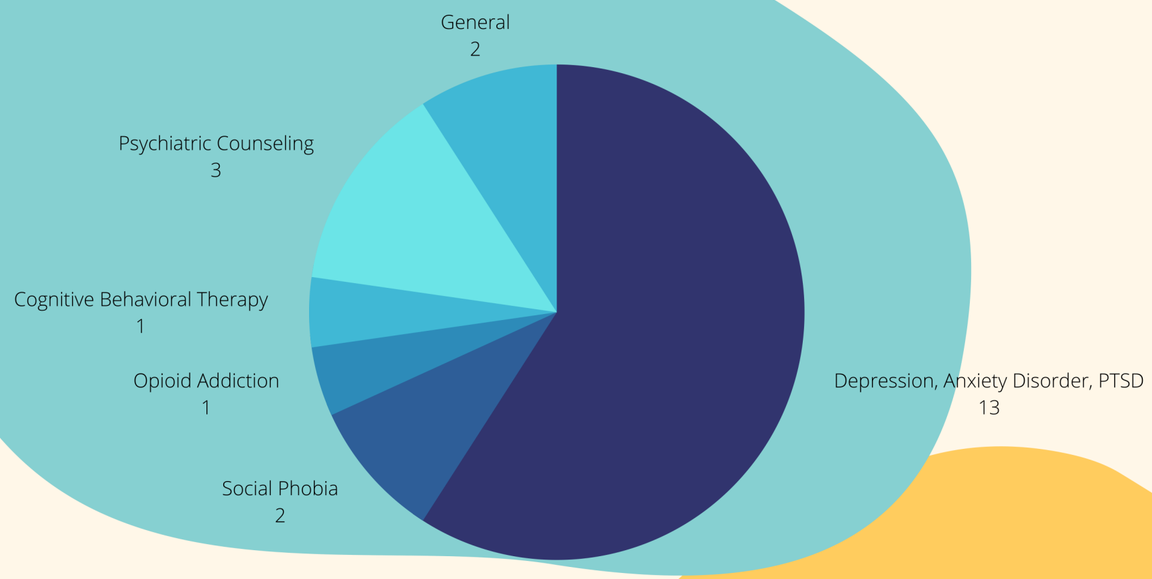
Domain of Research

- Healthcare area in which the system operates
- Papers were annotated according to both broad domains and specific subcategories
- Broad domains included:
  - Mental Health: Systems that support individuals with mental or psychological health conditions
  - Cognitive or Developmental Health: Systems that support individuals with conditions impacting memory, executive, or other cognitive function
  - Physical Health: Systems targeted towards individuals with specific physical health concerns, including infectious, non-infectious, and temporary conditions
  - Health Information: Systems that perform general-purpose actions designed to inform or suggest diagnoses
  - Patient Assistance: Systems that support patient-focused healthcare tasks
  - Physician Assistance: Systems that support physician-focused healthcare tasks

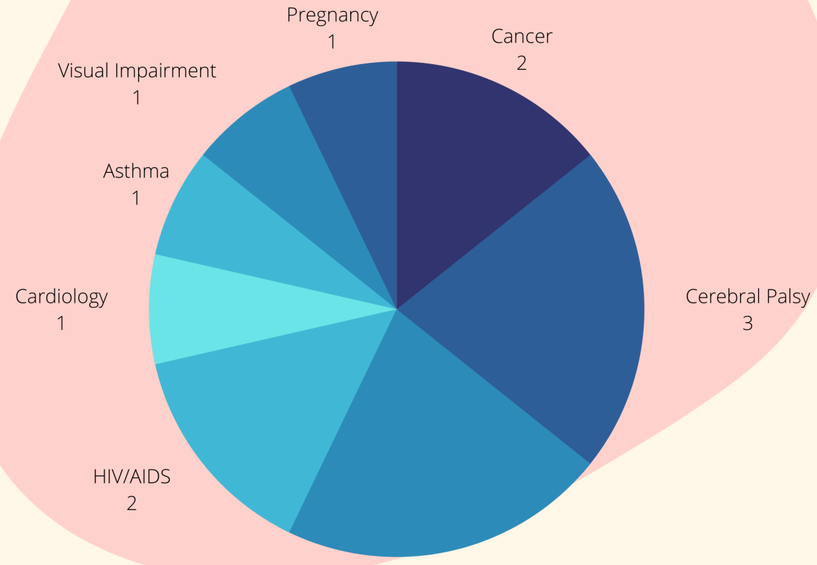
Broad Domain Categories



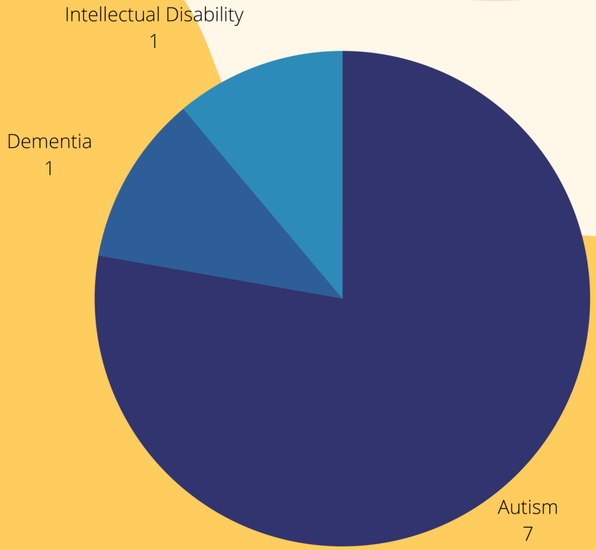
Mental Health



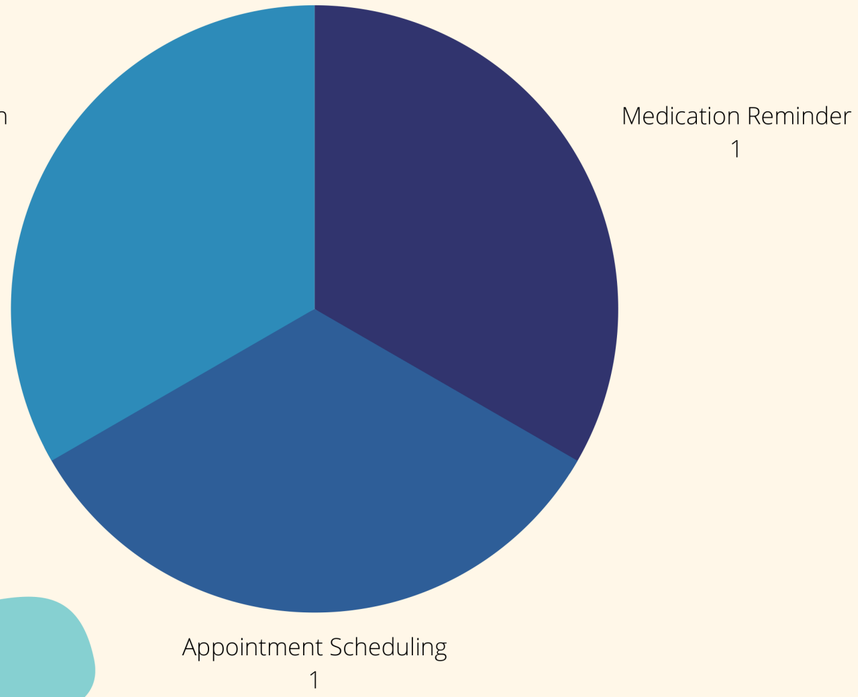
Physical Health



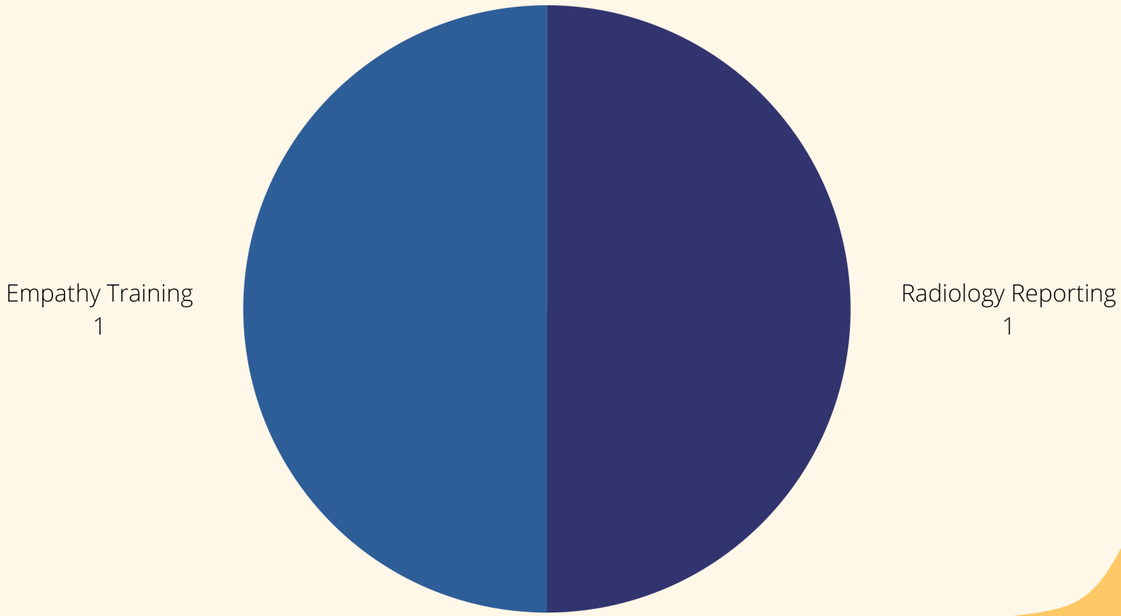
Cognitive or Developmental Health



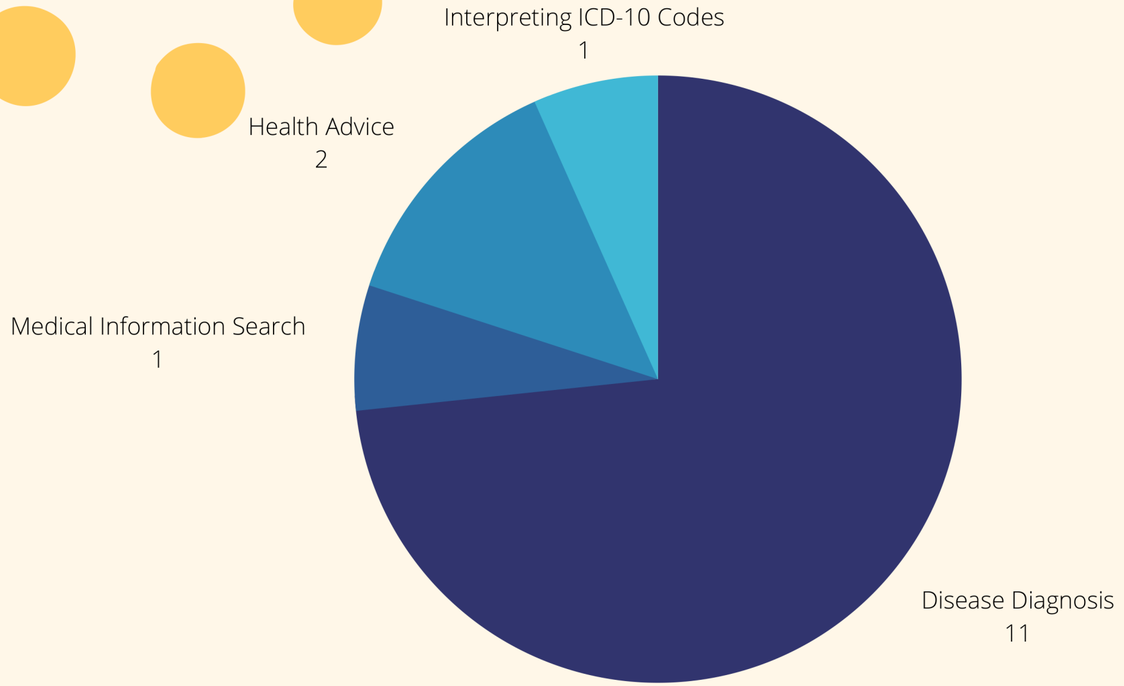
Patient Assistance



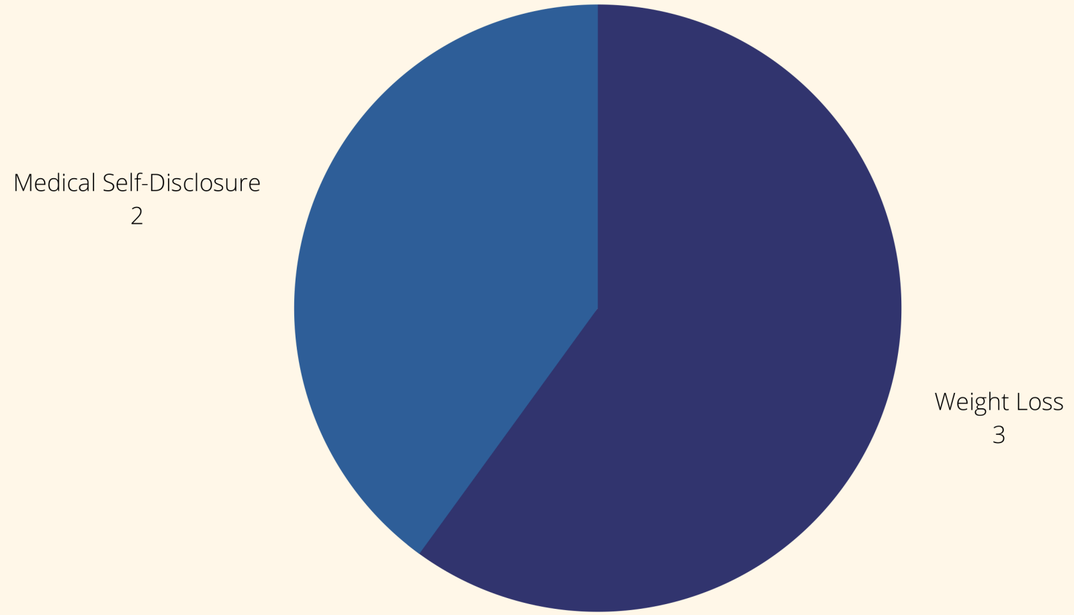
Physician Assistance



Health Information



Other

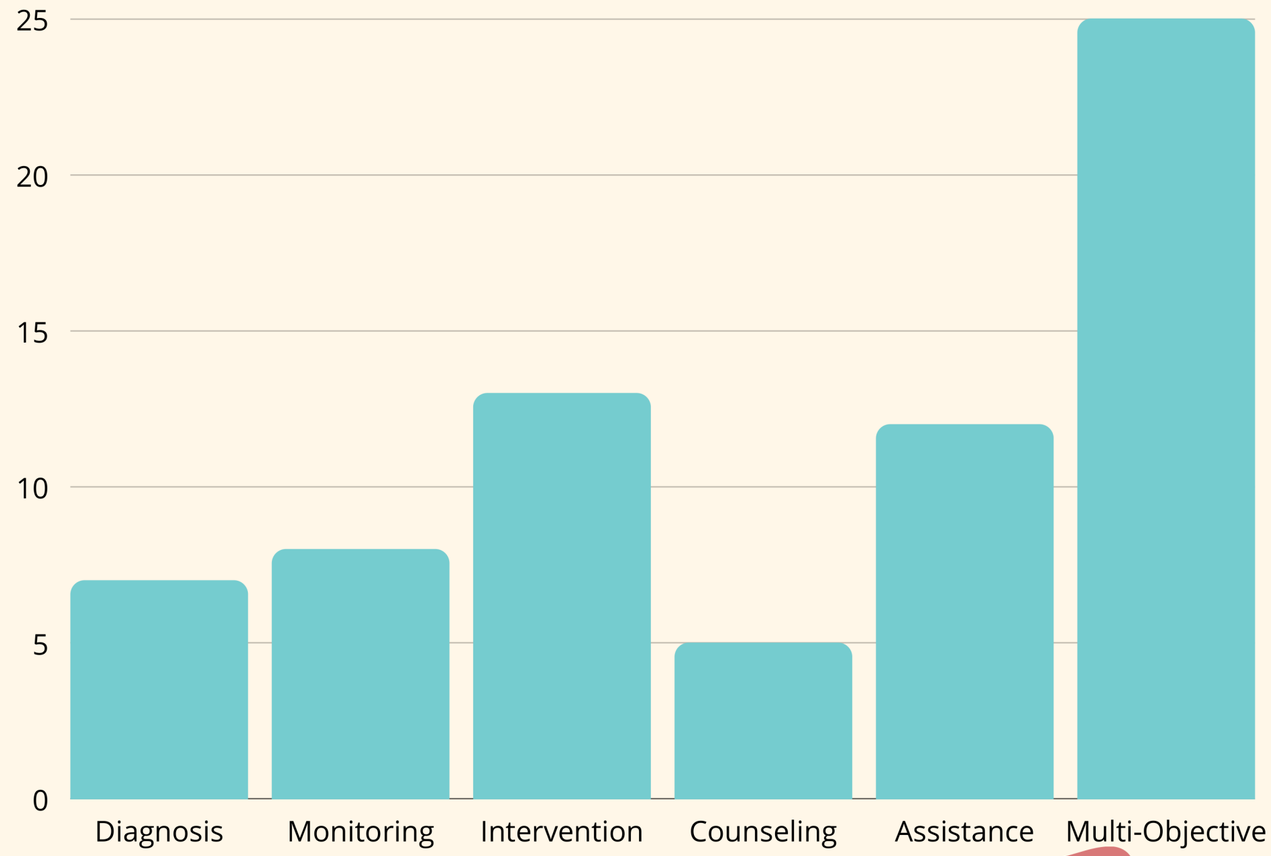


System Objective

The healthcare task for which a system is designed

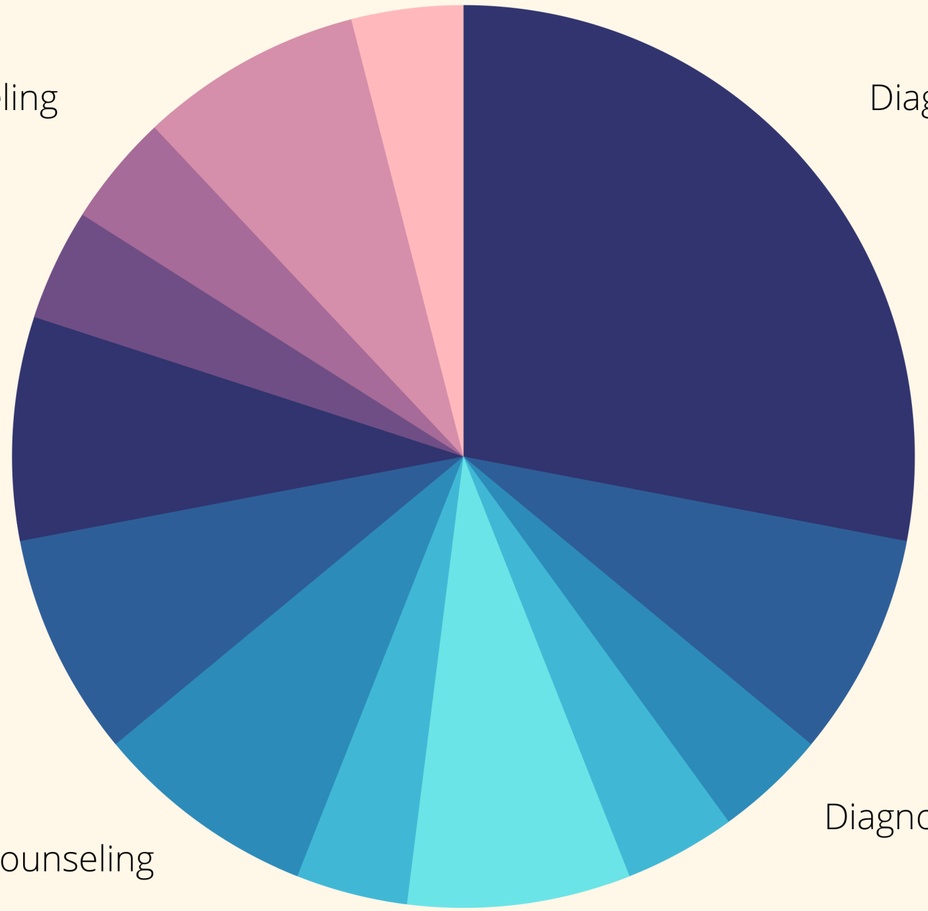
- **Diagnosing** a health condition
- **Monitoring** user states
- **Intervening** by addressing users' health concerns or improving their states
- **Counseling** users without directly intervening
- **Assisting** users by providing information or guidance

System Objective



Multi-Objective: Designed for more than one objective

How do systems combine objectives?



Target Audience

59 systems were designed for patients

3 systems were designed for caregivers

11 systems were designed for clinicians

2 systems were designed for patients and caregivers

Language

Most (56%) of systems were designed for English speakers.

Other languages were also represented, including Chinese (9%), Korean (6%), Japanese (4%), Swedish (3%), French (2%), Telugu (2%), Bengali (2%), Dutch, multiple Indian languages including Indian English, Italian, Arabic, and Setswana (all 1%).

Some papers (10%) did not specify the system's language.

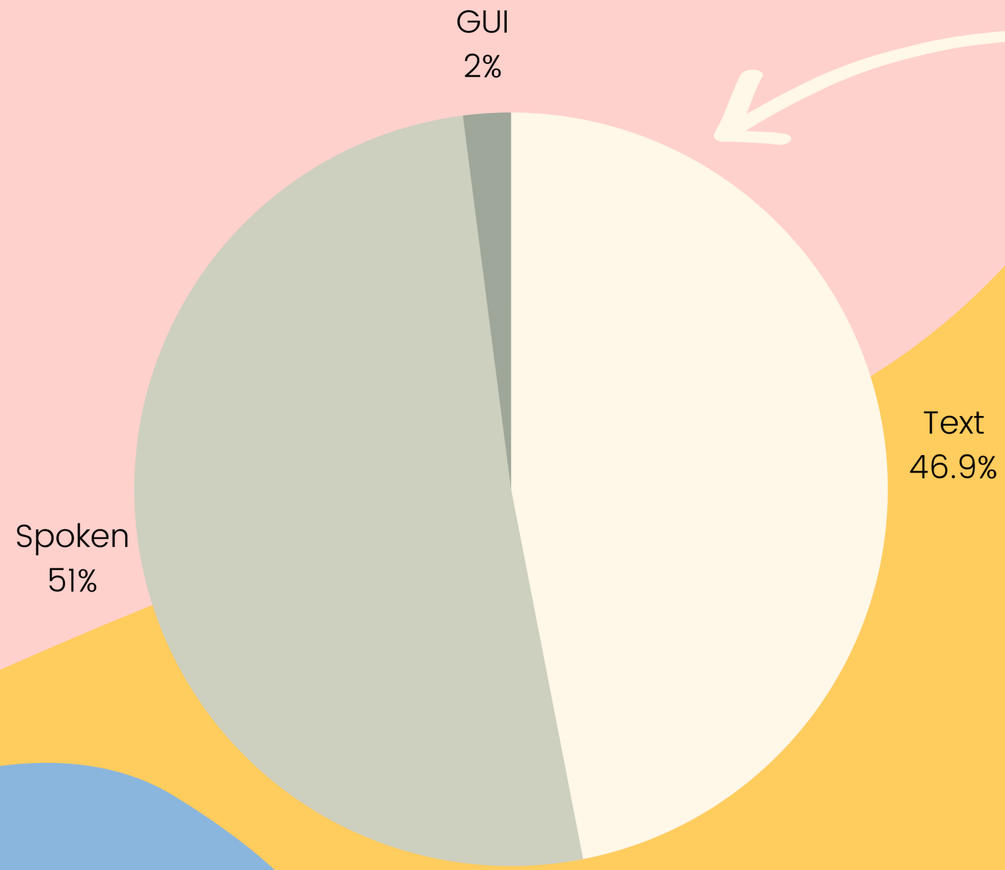
Architecture

General Architecture

- Pipeline architectures including separate components for core tasks (i.e., dialogue policy, natural language generation, etc.)
  - 58 systems
- End-to-end architectures training a single model to produce output for a given input
  - 2 systems
- 10 systems did not specify their architecture

Dialogue Management Architecture

- Among the 58 systems with pipeline architectures, 17 used a rule-based dialogue manager
- 20 systems used intent-based dialogue managers
- 21 systems used hybrid dialogue managers that combined elements of rule-based and intent-based approaches

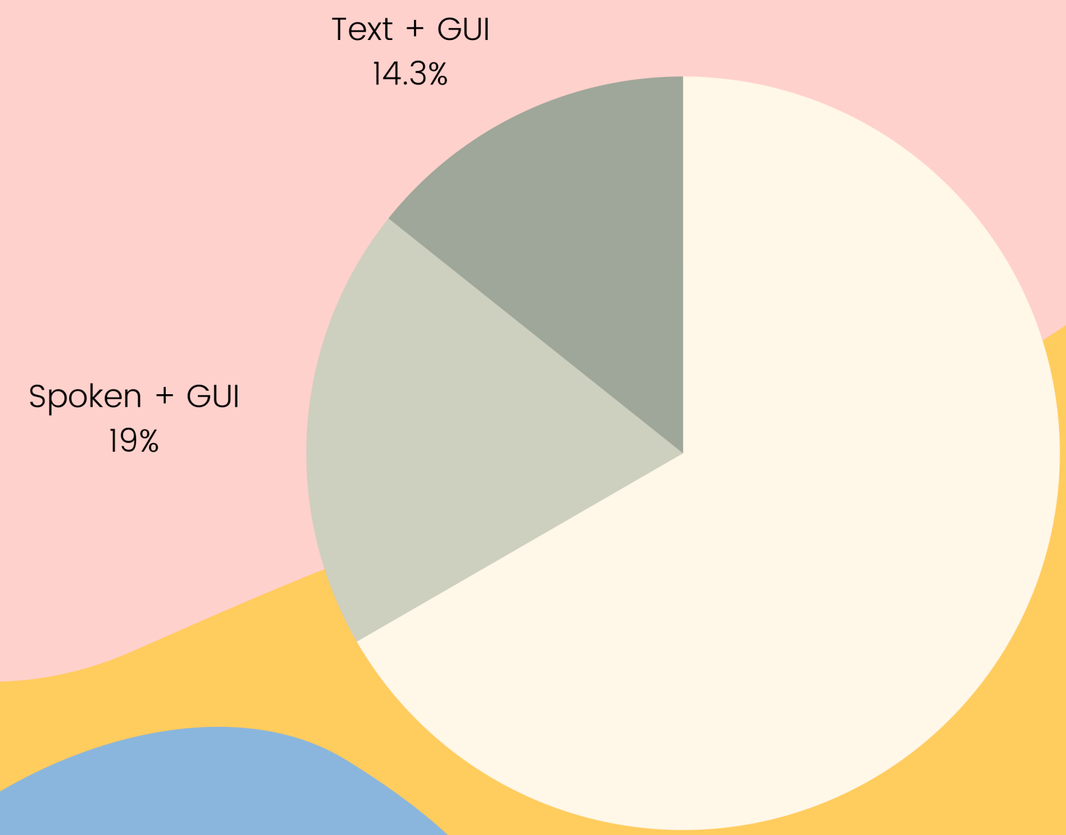


System Design

Modality

The channel through which information is exchanged between the dialogue system and human conversant

- Systems may be unimodal (49 systems) or multimodal (21 systems)
- Included modalities:
  - Text-based: Users interact with the system by typing
  - Spoken: Users interact with the system by speaking
  - Graphical User Interface (GUI): Users interact with the system through the use of visual elements



System Design

Modality

The channel through which information is exchanged between the dialogue system and human conversant

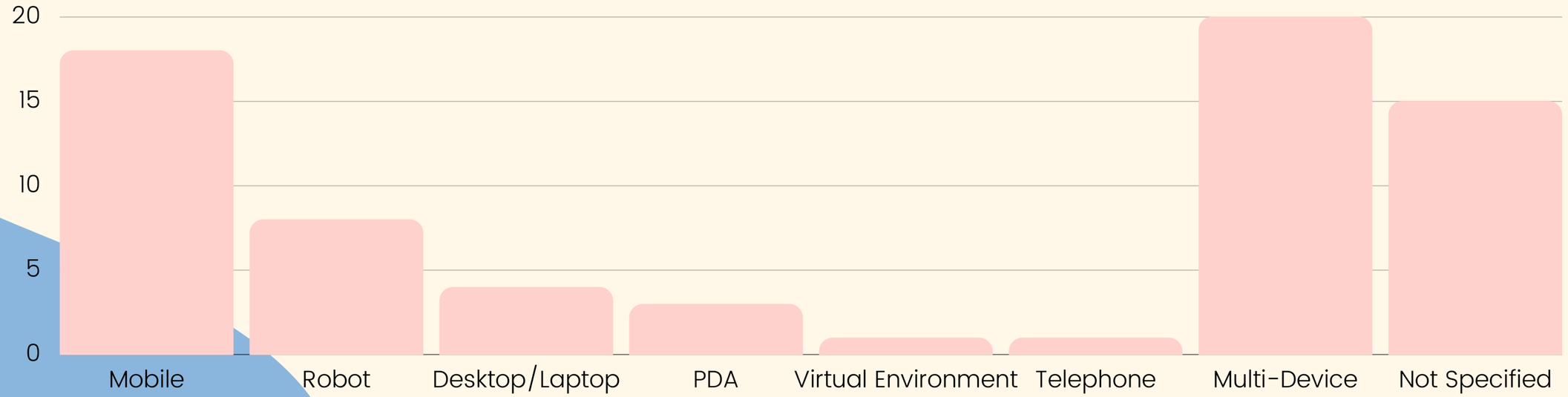
- Systems may be unimodal (49 systems) or multimodal (21 systems)
- Included modalities:
  - Text-based: Users interact with the system by typing
  - Spoken: Users interact with the system by speaking
  - Graphical User Interface (GUI): Users interact with the system through the use of visual elements

System Design

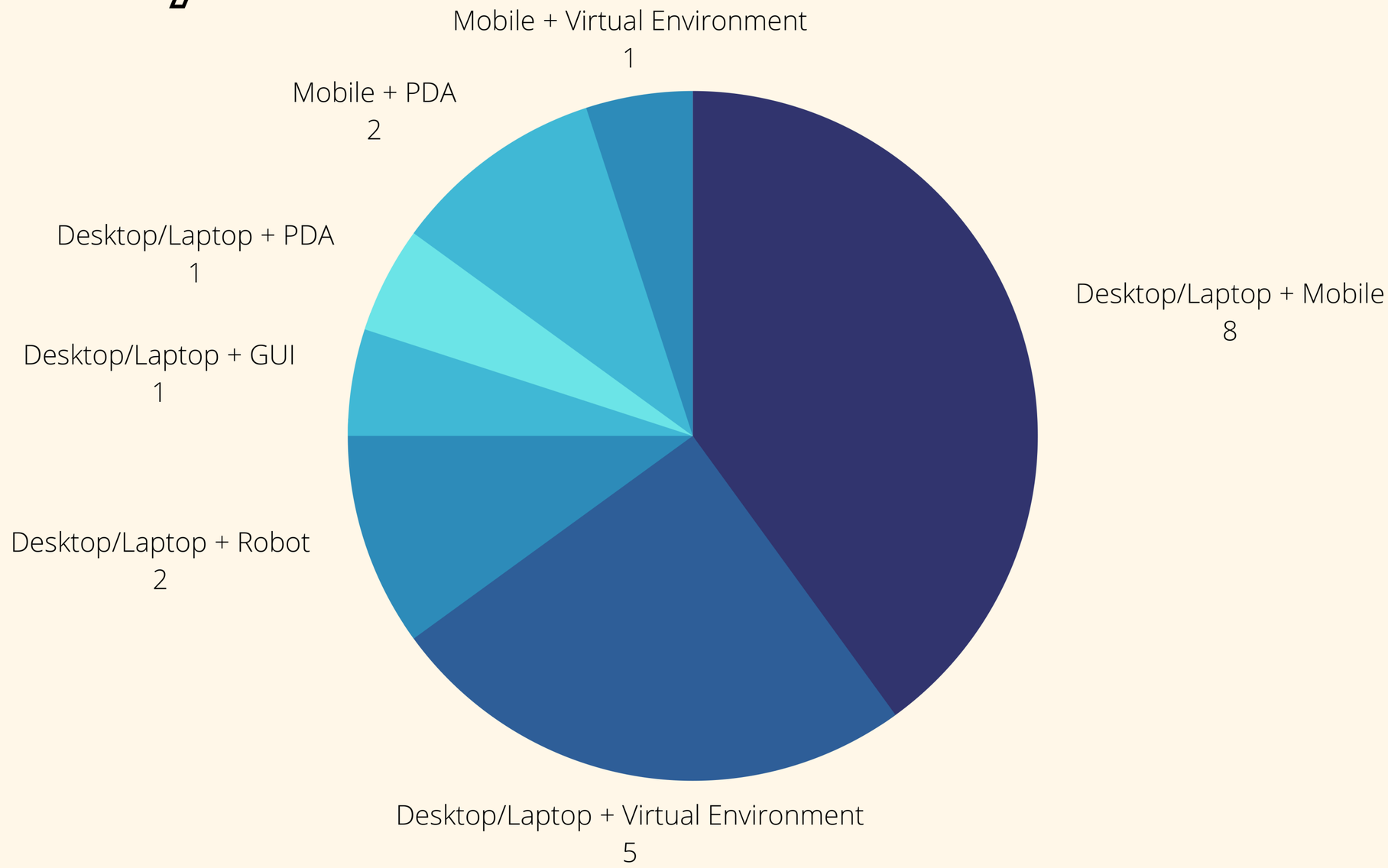
Device Type

Systems may be available on one or more devices:

- Mobile
- Telephone
- Desktop/Laptop
- Car
- PDA
- Robot
- Virtual Environment
- Virtual Reality



How do systems combine devices?



What data was used to train systems?

20

of the 70 papers provided details about the data used.

- 2 provided a link to the dataset
- 18 mentioned the size of the dataset

EVALUATION

Only 10 papers mentioned IRB (or IRB-equivalent) approval for their dataset and/or task.

28

Papers conducted a human evaluation

7

Papers conducted an automated evaluation

9

Papers conducted both human and automated evaluations

The remaining **26** papers didn't describe evaluation.

Human Evaluation

Among papers that described human evaluations:

- Evaluations included an average of 26 (mode=12) participants
- 15 used lab experiments
- 17 used field experiments
- 1 used crowdsourced data
- Most (28) asked users to both interact with the system and rate its dialogue

Across all papers, 33 papers also evaluated system usability.

System Evaluation

Among the 9 systems that reported automated evaluations, many different metrics were used.

Task Completion

Task Performance

Response Correctness

Naturalness

Response Time

Routing Time

OUTLINE

Focus: Task-oriented healthcare dialogue systems, following a recent survey conducted on this topic.

- Current state of task-oriented dialogue systems in the healthcare domain
- Recommendations for the future
- Ongoing work in this area

Currently, system design remains under-explored for these systems.

Most surveyed systems (86%) use a pipeline architecture.

- End-to-end systems offer performance improvements in other domains, but require high-quality data and/or rich external knowledge sources.
- End-to-end systems may also hinder interpretability, which is crucial in many healthcare applications, and be vulnerable to biases and privacy issues.

Many systems (33%) also only allow text interaction, despite evidence that some user groups are more comfortable conversing with dialogue systems via speech.

Currently, system design remains under-explored for these systems.

Many systems were also implemented on mobile phones.

Mobile phones are readily available to most users, but mobile applications are prone to long-term retention issues (Lee et al., 2018).

Most systems (84%) were patient-focused.

Other healthcare stakeholders, such as clinicians and caregivers, stand to benefit from task-oriented dialogue systems but have not often been leveraged as a target audience.

System Design Recommendations

Increase efforts to develop high-quality healthcare dialogue datasets.

In addition to opening the door to new research opportunities, this will facilitate exploration of more data-intensive architectures.

Deploy systems to a wider range of modalities and devices.

This invites participation from a wider range of users, and may facilitate more lasting integration of system use in daily life.

Explore applications targeting non-patient audiences.

Clinician- and caregiver-focused applications may offer broad, high-impact support in understanding, diagnosing, and treating patients' health issues.

Most existing research in this domain is not easily replicable.

Less than a third (29%) of surveyed papers discussed data quantity or characteristics.

Many papers also lacked important implementation details, such as evaluation methods (34%).

Replicability Recommendations

Easy first steps:

- Publish data when permissible, and descriptive statistics to the extent allowable when circumstances prevent data release
- Adopt well-established reproducibility guidelines from other NLP research domains
  - If permissible, release software as a persistent, easily deployable artifact
  - Otherwise, provide implementation and evaluation details in publications

Existing work often lacks adequate coverage of ethical and privacy issues.

- Only 27% of surveyed papers discussed privacy considerations
- Only 14% of papers reported Institutional Review Board (or IRB-equivalent) approval

Researchers should consider potential harms from research studies, data collection, and use or misuse of their systems.

- Researchers should submit their experimental design and protocol for external ethics review
- Researchers should follow guidelines established by the ACM Code of Ethics and/or other best practices when relevant

Deploying research in this domain to more target languages could create substantial broader impacts.

Most surveyed systems (56%) targeted English speakers, which is consistent with linguistic homogeneity elsewhere in NLP.

Systems currently developed for low-resource languages may offer roadmaps for how applications can be adapted to provide better coverage for the diverse, real-world user base.

Researchers should be cognizant of user experience when designing systems.

More than half of surveyed papers did not evaluate user experience or usability, and 60% did not consider user engagement.

- User experience is crucial in healthcare applications, since it directly influences adoption and adherence to healthcare outcomes (Montenegro et al., 2019).
- Researchers should more explicitly consider user experience in their system design, development, and evaluation.

OUTLINE

Focus: Task-oriented healthcare dialogue systems, following a recent survey conducted on this topic.

- ✓ Current state of task-oriented dialogue systems in the healthcare domain
- ✓ Recommendations for the future
- Ongoing work in this area

Our next frontier for task-oriented healthcare dialogue systems....

New(ish) project: Smart and Connected Family Engagement for Equitable Early Intervention Service Design

A smart and connected application to help family caregivers develop early intervention plans for children with rehabilitation needs

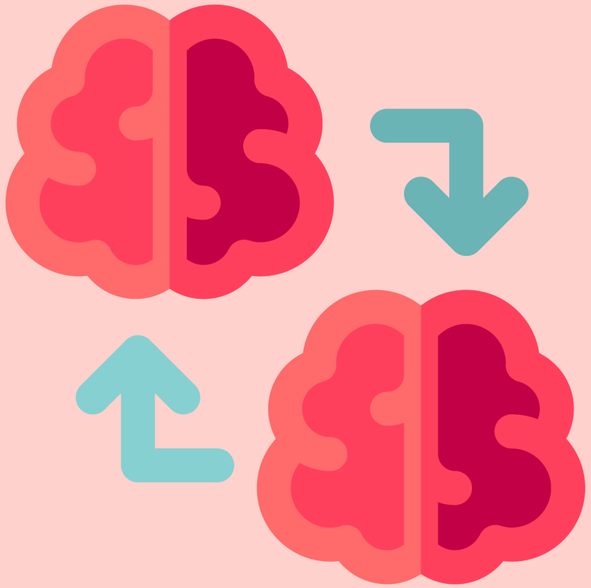
Cultural adaptations to make content more relevant for diverse families, in consultation with expert collaborators

Intelligent strategy exchange feature to suggest relevant caregiving strategies submitted by other families

Adaptive dialogue agent to help caregivers from diverse families navigate the intervention process

Comprehensive evaluations of the intervention with community stakeholders

Intelligent Strategy Exchange



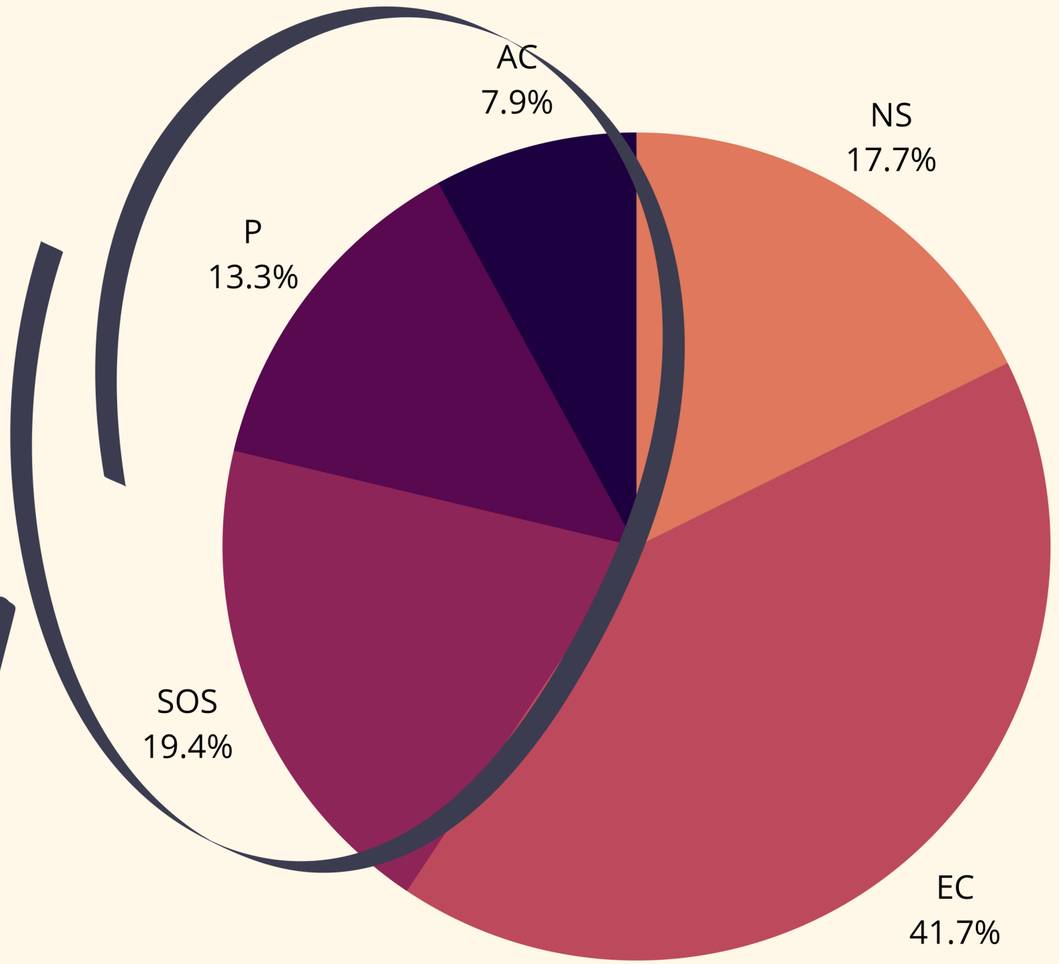
- Class labels based on known drivers of participation (Imms et al., 2017)
  - Environment/context
  - Sense of self
  - Preferences
  - Activity competence

Class Distribution

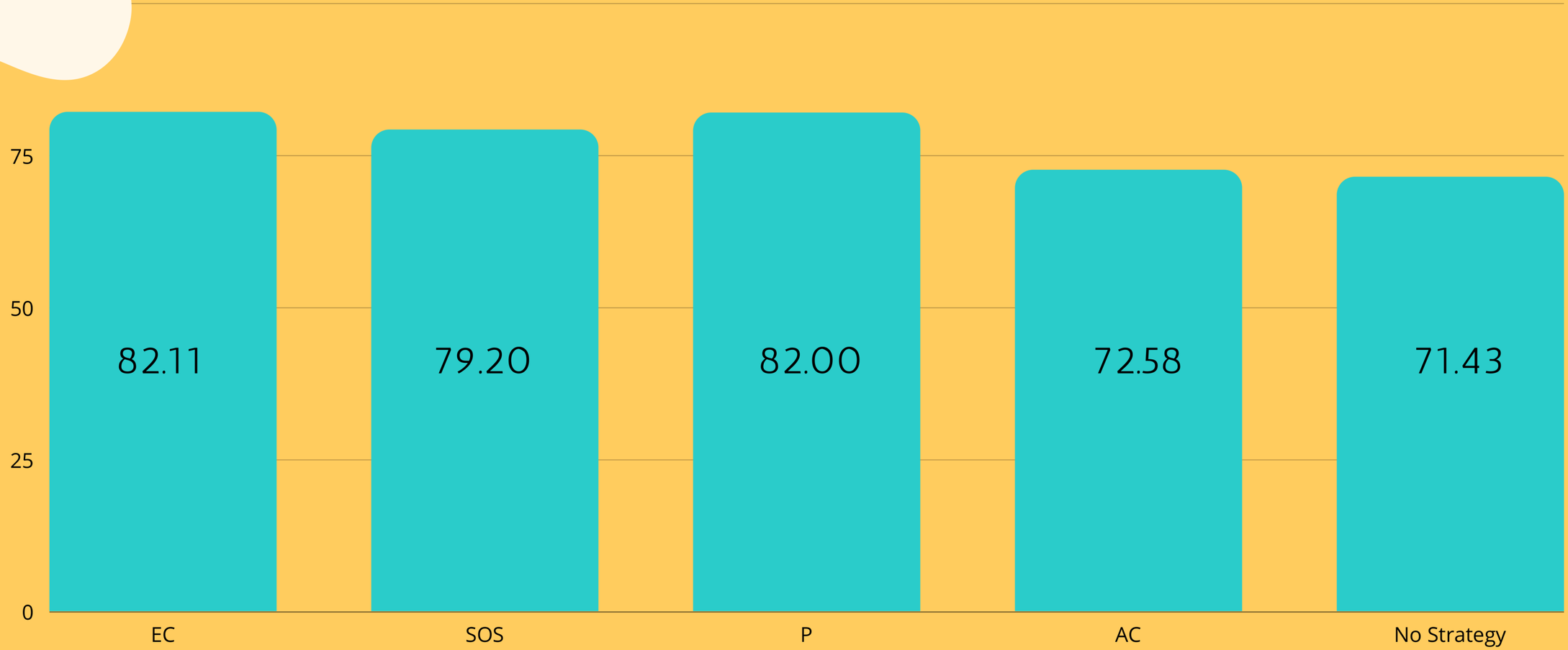
- No Strategy = 138
- Environment/Context (EC) = 325
- Sense of Self (SOS) = 151
- Preferences (P) = 104
- Activity Competence (AC) = 62

- Extrinsic = EC = 325
- Intrinsic = SOS + P + AC = 317

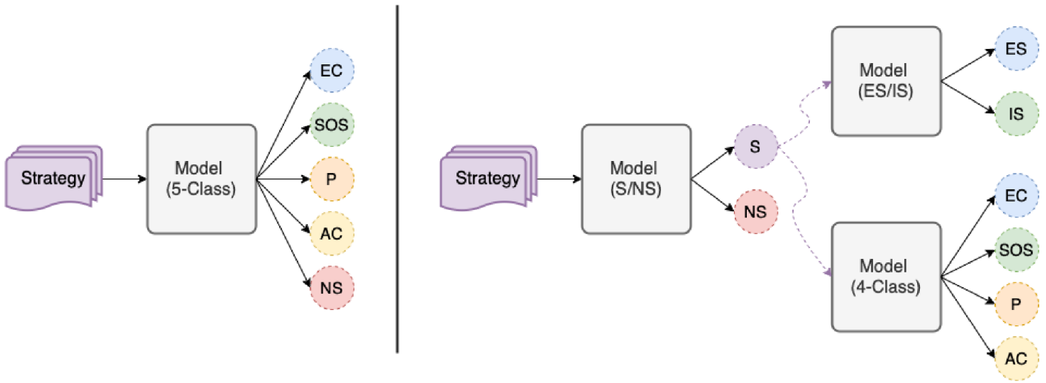
- Strategy = EC + SOS + P + AC = 642
- Total Dataset Size = 780



Inter-Annotator % Agreement



Classification Tasks



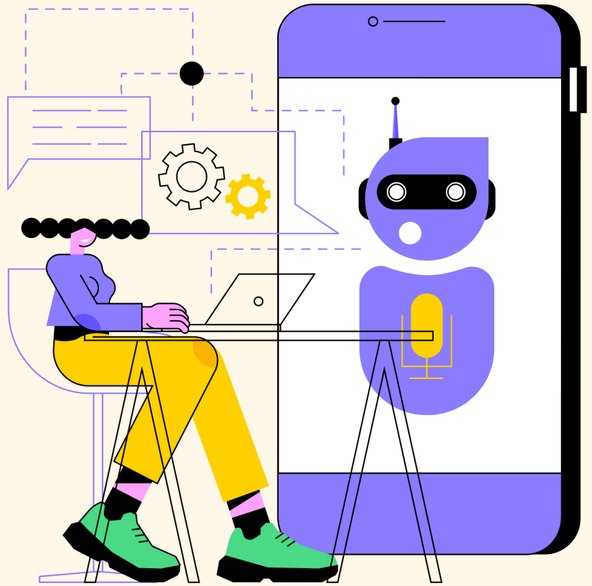
Flat

- Five-Way Classification: Predict one of the class labels or "no strategy"

Pipelined

- Strategy Detection: Predict either "strategy" (environment/context, sense of self, preferences, or activity competence) or "no strategy"
- Extrinsic vs. Intrinsic Classification: Predict either "extrinsic" (environment/context) or "intrinsic" (sense of self, preferences, or activity competence)
- Fine-Grained Classification: Predict one of the four strategy classes

Data Preprocessing



1. Spelling Correction
2. Punctuation Removal
3. Number Replacement
4. Name Replacement
5. Case Normalization
6. (Classical Models) Stopword Removal
7. (Classical Models) Lemmatization

MODELS

Model	Accuracy	Precision	Recall	F1
Baseline	40.78	0.08	0.20	0.11
LR	55.26	0.64	0.40	0.42
NB	52.63	**0.83**	0.36	0.36
BERT	**59.21**	0.58	**0.43**	**0.43**
Bio-ClinicalBERT	53.94	0.40	0.37	0.32

Strategy Representation

- TF-IDF vectorization for classical models

Classification Algorithms

- Most Frequent Class Baseline
- Logistic Regression
- Naïve Bayes
- BERT
- Bio-ClinicalBERT

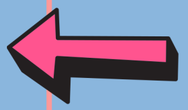
Model Selection

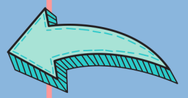
- Compared performance on five-way classification task

Per-Class Accuracy

- At the class level, "environment/context" strategies were easiest to predict, and "preferences" strategies were hardest
- In general, per-class performance largely followed class size

Class	LR	NB	BERT	Bio-ClinicalBERT
EC	87.10	93.55	90.32	83.87
SOS	53.33	53.33	60.00	80.00
P	10.00	10.00	0.00	0.00
AC	16.67	16.67	16.67	0.00
No Strategy	35.71	7.14	50.00	21.42





Task Comparison

Task	Accuracy	Precision	Recall	F1
S/NS	93.43	0.88	0.90	0.89
ES/IS	75.80	0.77	0.75	0.75
Fine-Grained	58.06	0.54	0.57	0.53

Adaptive Dialogue Agent

Under (early stages of) construction!

- Designed incorporating recommendations identified during literature survey

OUTLINE

Focus: Task-oriented healthcare dialogue systems, following a recent survey conducted on this topic.

- Current state of task-oriented dialogue systems in the healthcare domain
- Recommendations for the future
- Ongoing work in this area

Final Thoughts

Task-oriented dialogue systems in healthcare are growing increasingly prevalent.

Existing work covers a wide range of healthcare domains and objectives.

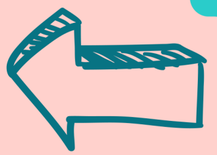
There is still substantial opportunity for growth in this area.

Further investigation of system design, increased attention to user experience, greater consideration of privacy and ethics concerns, and more language diversity are all key areas for improvement.

Caregiver- and clinician-focused applications are high-impact but under-explored.

Our ongoing work investigates caregiver-centered applications more fully.

Thanks!



Literature review:

- Mina Valizadeh and Natalie Parde. The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications. In the *Proceedings of ACL 2022*. Dublin, Ireland, May 22-27, 2022.

Collaborators:

- Mary Khetani, Vera Kaelin, and Vivian Villegas

This work was based in part upon work supported by the National Science Foundation under Grant No. 2125411.

Feel free to contact us with further questions: parde@uic.edu or mvaliz2@uic.edu

References

A. Vaidyam, Hannah Wisniewski, J. Halamka, M. S. Kashavan, and J. Torous. 2019. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. The Canadian Journal of Psychiatry, 64:456 – 464.

William Kearns, Nai-Ching Chi, Yong Choi, Shih-Yin Lin, Hilaire Thompson, and George Demiris. 2019. A systematic review of health dialog systems. Methods of Information in Medicine, 58:179–193.

Liliana Laranjo, Adam Dunn, Huong Ly Tong, A. Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: A systematic review. Journal of the American Medical Informatics Association, 0.

Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. Expert Systems with Applications, 129:56–67.

Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. 2020. Conversational agents in health care: Scoping review and conceptual analysis. J Med Internet Res, 22(8):e17158.

Christine Imms, Mats Granlund, Peter H. Wilson, Bert Steenbergen, Peter L. Rosenbaum, and Andrew M. Gordon. 2017. Participation, both a means and an end: A conceptual analysis of processes and outcomes in childhood disability. Developmental Medicine & Child Neurology, 59(1):16–25.