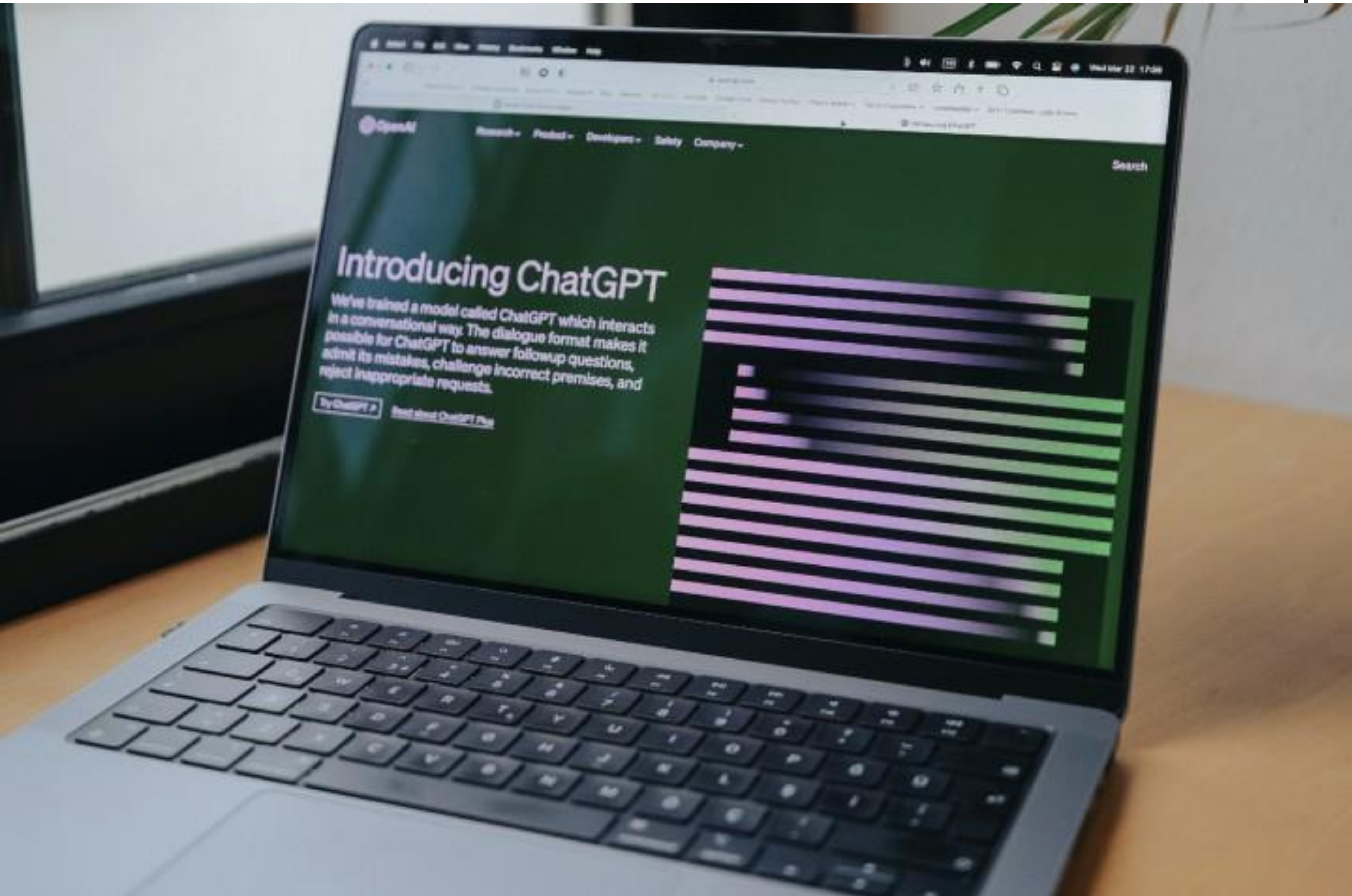
CS 6474/CS 4803 Social Computing: Generative AI

Munmun De Choudhury

munmund@gatech.edu

Week 14 | April 9, 2025


PRICE $8.99
THE NEW YORKER
APRIL 24 & MAY 1, 2023


OpenAI
Research
Product
Developers
Safety
Company
Search
Introducing ChatGPT
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

THE NEW YORKER

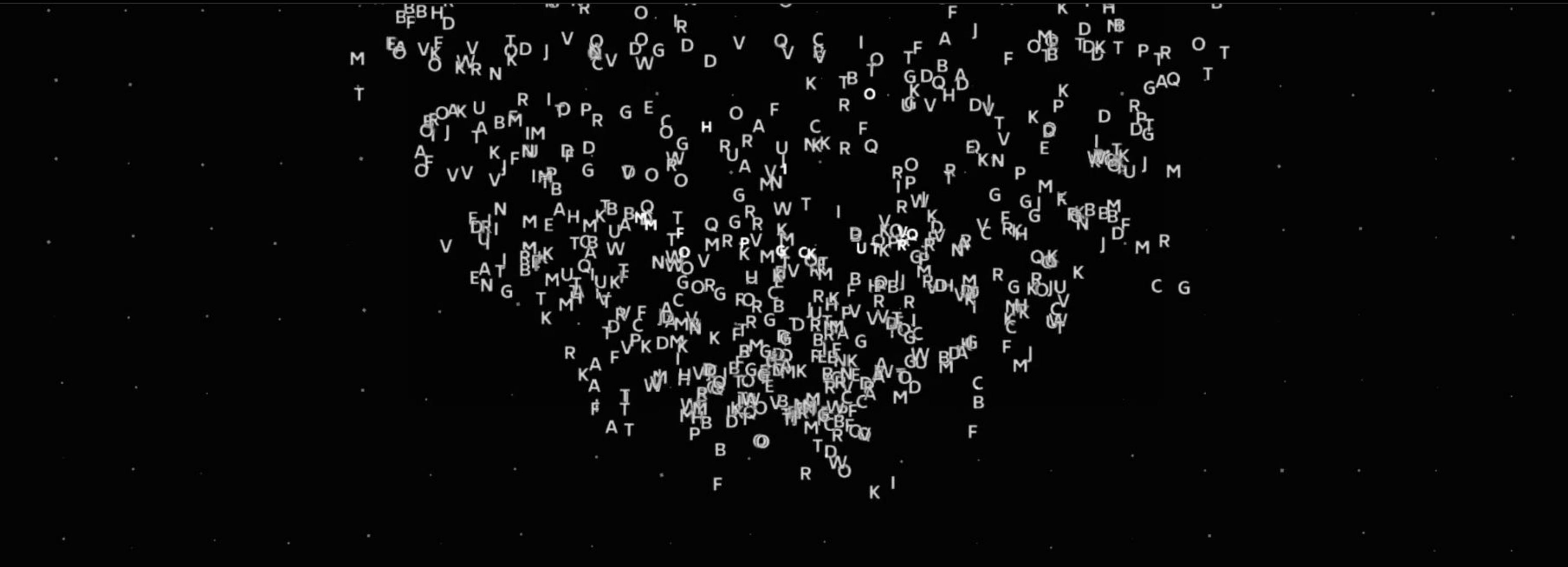


Illustration by Vivek Thakker

ANNALS OF TECHNOLOGY

CHATGPT IS A BLURRY JPEG OF THE WEB

OpenAI's chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?

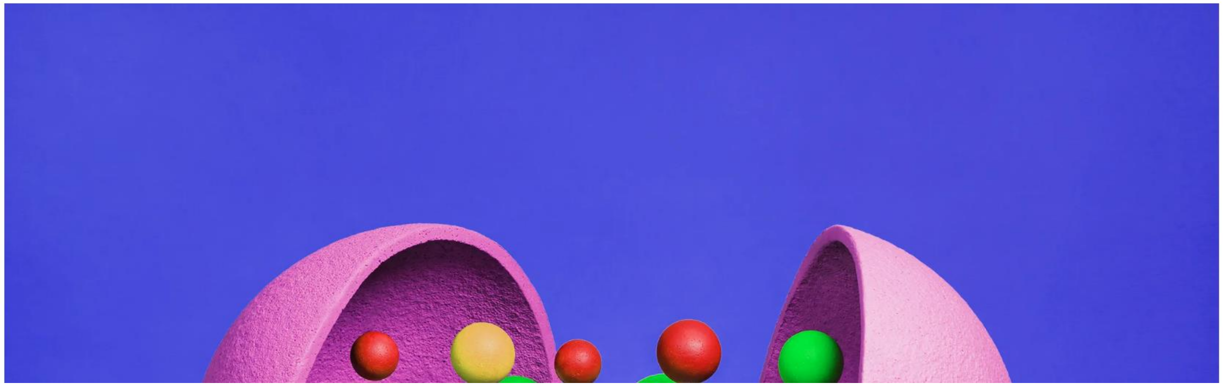
By Ted Chiang

February 9, 2023

WILL KNIGHT BUSINESS APR 18, 2023 7:00 AM

Some Glimpse AGI in ChatGPT. Others Call It a Mirage

A new generation of AI algorithms can *feel* like they're reaching artificial general intelligence—but it's not clear how to measure that.



Perspective | Published: 12 April 2023

Foundation models for generalist medical artificial intelligence

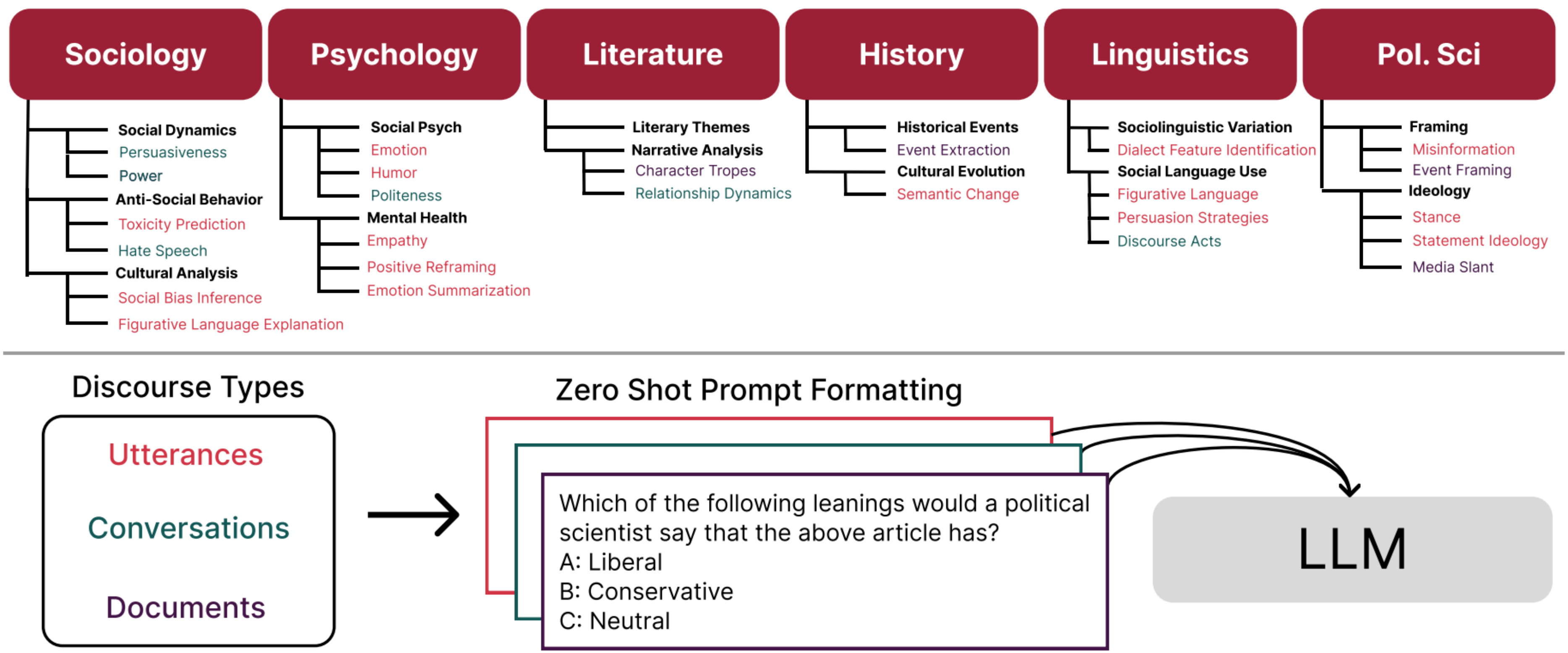
Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol & Pranav Rajpurkar

Nature **616**, 259–265 (2023) | Cite this article

Abstract

The exceptionally rapid development of highly flexible, reusable artificial intelligence (AI) models is likely to usher in newfound capabilities in medicine. We propose a new paradigm for medical AI, which we refer to as generalist medical AI (GMAI). GMAI models will be capable of carrying out a diverse set of tasks using very little or no task-specific labelled data. Built

Can Large Language Models Transform Computational Social Science?



Sociology
Social Dynamics
Persuasiveness
Power
Anti-Social Behavior
Toxicity Prediction
Hate Speech
Cultural Analysis
Social Bias Inference
Figurative Language Explanation
Psychology
Social Psych
Emotion
Humor
Politeness
Mental Health
Empathy
Positive Reframing
Emotion Summarization
Literature
Literary Themes
Narrative Analysis
Character Tropes
Relationship Dynamics
History
Historical Events
Event Extraction
Cultural Evolution
Semantic Change
Linguistics
Sociolinguistic Variation
Dialect Feature Identification
Social Language Use
Figurative Language
Persuasion Strategies
Discourse Acts
Pol. Sci
Framing
Misinformation
Event Framing
Ideology
Stance
Statement Ideology
Media Slant
Discourse Types
Utterances
Conversations
Documents
Zero Shot Prompt Formatting
Which of the following leanings would a political scientist say that the above article has?
A: Liberal
B: Conservative
C: Neutral
LLM

Overview of Tasks

Dataset	Size	Classes
Generation Tasks	500	–
Utterance Level		
Dialect	266	23
Persuasion	399	7
Impl. Hate	498	6
Emotion	498	6
Figurative	500	4
Ideology	498	3
Stance	435	3
Humor	500	2
Misinfo	500	2
Semantic Chng	344	2

Dataset	Size	Classes
Conversation Level		
Discourse	497	7
Politeness	498	3
Empathy	498	3
Toxicity	500	2
Power	500	2
Persuasion	434	2
Document Level		
Event Arg.	283	–
Evt. Surprisal	240	–
Tropes	114	114
Ideology	498	3

Performance of zero-shot models

Model / Data	Baselines		FLAN-T5					FLAN	text-001				text-002	text-003	Chat	
Data	Rand	Finetune	Small	Base	Large	XL	XXL	UL2	Ada	Babb.	Curie	Dav.	Davinci	Davinci	GPT3.5	GPT4
Utterance Level Tasks																
Dialect	3.3	3.0	0.2	4.5	23.4	24.8	30.3	32.9	0.5	0.5	1.2	9.1	17.1	14.7	11.7	23.2
Emotion	16.7	71.6	19.8	63.8	69.7	65.7	66.2	70.8	6.4	4.9	6.6	19.7	36.8	44.0	47.1	50.6
Figurative	25.0	99.2	16.6	23.2	18.0	32.2	53.2	62.3	10.0	15.2	10.0	19.4	45.6	57.8	48.6	17.5
Humor	49.5	73.1	51.8	37.1	54.9	56.9	29.9	56.8	38.7	33.3	34.7	29.2	29.7	33.0	43.3	61.3
Ideology	33.3	64.8	18.6	23.7	43.0	47.6	53.1	46.4	39.7	25.1	25.2	23.1	46.0	46.8	43.1	60.0
Impl. Hate	16.7	62.5	7.4	14.4	7.2	32.3	29.6	32.0	7.1	7.8	4.9	9.2	18.4	19.2	16.3	3.7
Misinfo	50.0	81.6	33.3	53.2	64.8	68.7	69.6	77.4	45.8	36.2	41.5	42.3	70.2	73.7	55.0	26.9
Persuasion	14.3	52.0	3.6	10.4	37.5	32.1	45.7	43.5	3.6	5.3	4.7	11.3	21.6	17.5	23.3	56.4
Sem. Chng.	50.0	62.3	33.5	41.0	56.9	52.0	36.3	41.6	32.8	38.9	41.3	35.7	41.9	37.4	44.2	21.2
Stance	33.3	36.1	25.2	36.6	42.2	43.2	49.1	48.1	18.1	17.7	17.2	35.6	46.4	41.3	48.0	76.0
Conversation Level Tasks																
Discourse	14.3	49.6	4.2	21.5	33.6	37.8	50.6	39.6	6.6	9.6	4.3	11.4	35.1	36.4	35.4	16.7
Empathy	33.3	71.6	16.7	16.7	22.1	21.2	35.9	34.7	24.5	17.6	27.6	16.8	16.9	17.4	22.6	6.4
Persuasion	50.0	33.3	9.2	11.0	11.3	8.4	41.8	43.1	6.9	6.7	6.7	33.3	33.3	53.9	51.7	28.6
Politeness	33.3	75.8	22.4	42.4	44.7	57.2	51.9	53.4	16.7	17.1	33.9	22.1	33.1	39.4	51.1	59.7
Power	49.5	72.7	46.6	48.0	40.8	55.6	52.6	56.9	43.1	39.8	37.5	36.9	39.2	51.9	56.5	42.0
Toxicity	50.0	64.6	43.8	40.4	42.5	43.4	34.0	48.2	41.4	34.2	33.4	34.8	41.8	46.9	31.2	55.4
Document Level Tasks																
Event Arg.	22.3	65.1	–	–	–	–	–	–	–	–	8.6	8.6	21.6	22.9	22.3	23.0
Event Det.	0.4	75.8	9.8	7.0	1.0	10.9	41.8	50.6	29.8	47.3	47.4	44.4	48.8	52.4	51.3	14.8
Ideology	33.3	85.1	24.0	19.2	28.3	29.0	42.4	38.8	22.1	26.8	18.9	21.5	42.8	43.4	44.7	51.5
Tropes	36.9	–	1.7	8.4	13.7	14.6	19.0	28.6	7.7	12.8	16.7	15.2	16.3	26.6	36.9	44.9

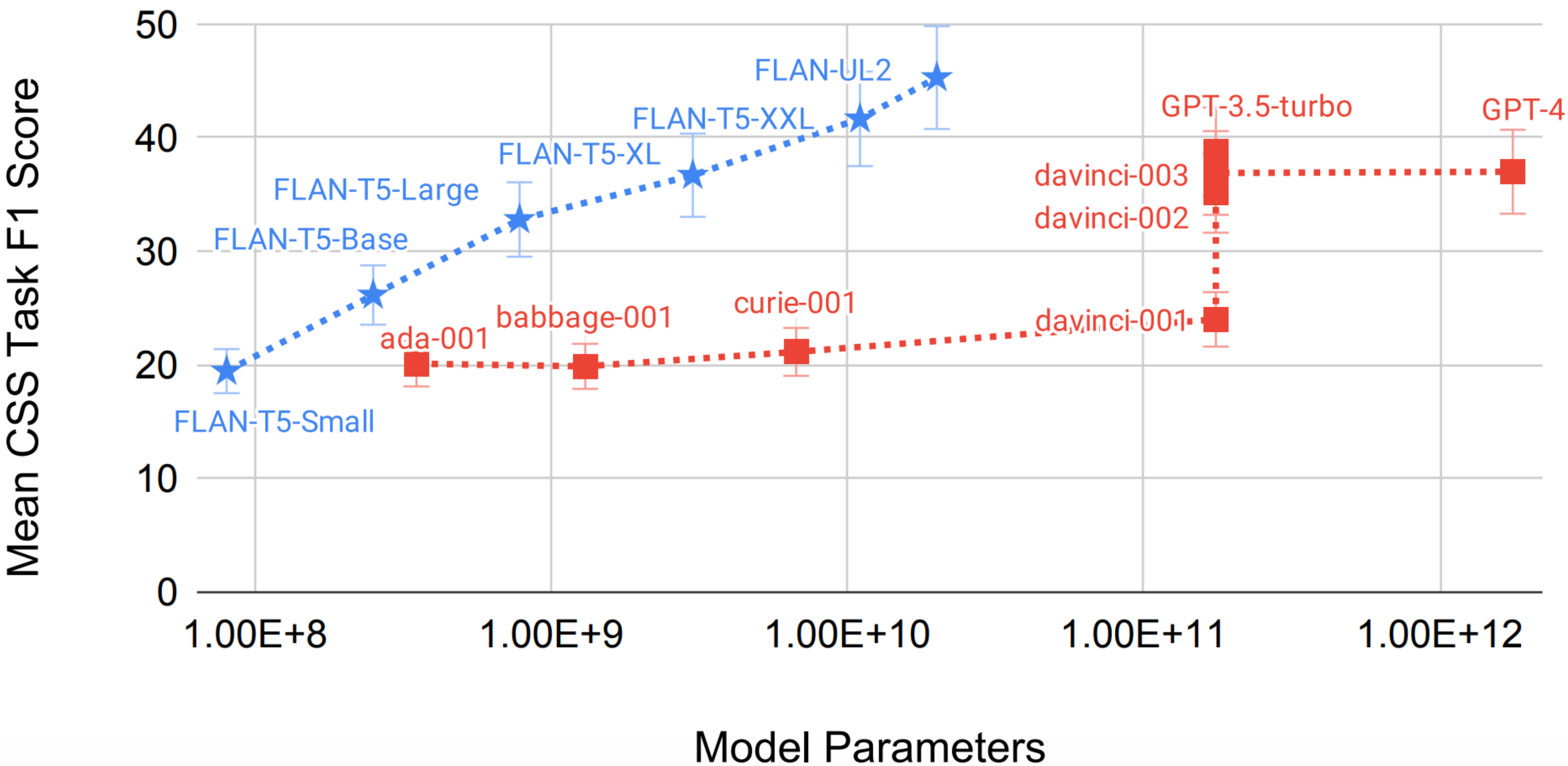
Dataset	Best Model	F1	κ	Agreement
Utterance-Level				
Dialect	flan-ul2	32.9	0.15	poor
Emotion	flan-ul2	**70.8**	0.65	good
Figurative	flan-ul2	62.3	0.52	moderate
Humor	gpt-4	61.3	0.23	fair
Ideology	davinci-002	60.0	0.40	moderate
Impl. Hate	flan-ul2	32.3	0.20	fair
Misinfo	flan-ul2	**77.4**	0.55	moderate
Persuasion	gpt-4	56.4	0.51	moderate
Semantic Chng.	flan-t5-large	56.9	0.14	poor
Stance	gpt-3.5-turbo	**72.0**	0.58	moderate

Dataset	Best Model	F1	κ	Agreement
Convo-Level				
Discourse	flan-t5-xxl	50.6	0.45	moderate
Empathy	flan-t5-xxl	35.9	0.04	poor
Persuasion	davinci-003	53.9	0.14	poor
Politeness	flan-t5-xl	59.2	0.38	fair
Power	gpt-4	59.7	0.26	fair
Toxicity	gpt-4	55.4	0.11	poor
Document-Level				
Ideology	gpt-4	51.5	0.51	moderate
Event Det.	gpt-4	23.0	n/a	-
Tropes	gpt-4	44.9	n/a	-

Do few-shot learning approaches improve performance?

Model	FLAN Small			FLAN Base			FLAN Large			FLAN XL			FLAN XXL			FLAN UL2		
Shot	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5	0	3	5
Dialect	0.2	0.0	**0.4**	**4.5**	0.0	1.4	**23.4**	0.7	14.1	**24.8**	8.0	20.5	**30.3**	0.2	29.9	**32.9**	12.6	27.5
Emotion	**19.8**	10.6	10.1	**63.8**	42.7	42.0	**69.7**	67.6	67.4	**65.7**	62.1	62.5	**66.2**	61.8	57.4	**70.8**	70.0	69.8
Figurative	**16.6**	10.0	9.2	23.2	**29.1**	27.3	18.0	**21.8**	19.6	**32.2**	27.9	28.5	53.2	52.6	**66.2**	**62.3**	52.7	62.0
Humor	51.8	52.8	**53.1**	**37.1**	35.1	34.7	**54.9**	54.0	53.8	56.9	**57.0**	56.7	29.9	34.8	**35.3**	**56.8**	55.5	54.1
Ideology	18.6	16.7	**24.0**	**23.7**	22.6	38.3	43.0	**47.3**	45.5	47.6	**48.8**	50.4	53.1	52.9	**57.7**	46.4	36.9	**51.5**
Impl. Hate	**7.4**	6.8	6.2	14.4	**21.1**	7.4	7.2	**9.3**	4.7	32.3	28.5	**34.6**	29.6	31.6	**35.1**	**32.0**	29.5	25.9
Misinfo	**33.3**	33.3	33.3	53.2	45.3	**59.7**	**64.8**	64.8	64.2	68.7	67.2	**69.7**	69.6	**74.9**	74.4	**77.4**	53.7	76.4
Persuasion	**3.6**	3.6	3.6	10.4	**10.8**	7.3	37.5	**39.0**	37.7	32.1	**44.3**	41.8	45.7	44.6	**48.6**	**43.5**	42.2	40.1
Sem. Chng.	33.5	33.3	**34.0**	41.0	35.7	**41.7**	56.9	48.8	**60.4**	**52.0**	40.8	35.6	**36.3**	34.0	33.3	41.6	**62.5**	34.6
Stance	25.2	16.7	**29.6**	**36.6**	18.1	36.6	**42.2**	41.8	39.8	43.2	**52.1**	46.2	**49.1**	46.0	48.7	48.1	**55.6**	54.7
Discourse	4.2	4.0	**7.5**	**21.5**	18.1	20.7	33.6	3.6	**34.6**	37.8	3.6	**38.0**	**50.6**	3.6	43.4	**39.6**	3.6	39.1
Empathy	**16.7**	16.7	16.7	**16.7**	16.7	16.7	**22.1**	16.7	17.1	21.2	**30.4**	22.8	**35.9**	29.8	28.2	34.7	**41.5**	39.6
Persuasion	9.2	**55.9**	45.0	11.0	**55.0**	48.7	11.3	**54.6**	51.7	8.4	42.8	**43.8**	**41.8**	38.8	35.2	43.1	**44.9**	46.1
Politeness	**22.4**	16.7	20.1	**42.4**	23.9	35.4	44.7	44.5	**51.9**	**57.2**	27.7	50.4	**51.9**	44.2	50.3	53.4	43.6	**53.9**
Power	**46.6**	44.5	33.3	**48.0**	39.8	41.4	40.8	**45.5**	43.5	55.6	58.9	**60.2**	52.6	52.0	**62.6**	56.9	57.2	**57.5**
Toxicity	43.8	**46.7**	33.3	40.4	34.7	**54.4**	**42.5**	34.7	36.7	43.4	38.7	**49.2**	34.0	33.3	**35.1**	48.2	44.7	**52.5**
Ideology	**24.0**	16.7	19.2	19.2	16.6	**21.3**	**28.3**	17.0	17.9	29.0	**31.7**	27.0	42.4	**48.5**	47.9	38.8	**38.9**	39.7
Tropes	1.7	**5.1**	3.4	**8.4**	5.1	3.4	**13.7**	10.0	11.6	**14.6**	8.4	10.0	**19.0**	8.4	6.8	**28.6**	27.3	24.6

Bigger LLMs do not necessarily indicate better performance



Expert scoring evaluations for zero-shot generation tasks show that leading generative models (davinci-003, GPT 3.5) can match or exceed the faithfulness, relevance, coherence, and fluency of both fine-tuned models (Baseline) and gold references (Human).

Aspect-Based Summarization (CovidET)

Model	Faithful	Relevant	Coherent	Fluent
Baseline	2.1	2.3	2.1^-	2.6^-
ada-001	1.8^-	1.8^-	2.4	3.6
babbage-001	2.0^-	2.0	2.3	3.7
curie-001	2.3	2.3	2.6	3.8
davinci-001	2.3	2.4	2.5	3.9
davinci-002	2.4	2.5	3.2	4.0
davinci-003	2.9	2.8	3.0	**4.1**$^+$
GPT 3.5	**3.9**$^+$	**3.5**$^+$	**3.8**$^+$	**4.5**$^+$
GPT 4	**3.7**$^+$	**3.3**$^+$	**3.8**$^+$	**4.4**$^+$
Human	2.8	2.6	2.8	3.8

Implied Misinformation Explanation (MRF)

Model	Faithful	Relevant	Coherent	Fluent
Baseline	3.4	3.5	3.7	4.2
ada-001	1.1^-	1.1^-	2.0^-	4.5
babbage-001	1.6^-	1.7^-	2.5^-	4.3
curie-001	2.6^-	2.7^-	3.1^-	4.4
davinci-001	1.7^-	1.7^-	2.5^-	4.5
davinci-002	**3.9**$^+$	**4.1**$^+$	**4.3**$^+$	**4.9**$^+$
davinci-003	3.1^-	3.4	3.9	4.5
GPT 3.5	**3.7**$^+$	3.9	**4.2**$^+$	**4.9**$^+$
GPT 4	3.7	3.9	4.1	4.5
Human	3.5	3.7	3.9	4.4

Figurative Language Explanation (FLUTE)

Model	Faithful	Relevant	Coherent	Fluent
Baseline	1.4^-	1.7^-	1.4^-	4.2
ada-001	1.4^-	1.5^-	1.5^-	3.9
babbage-001	1.4^-	1.9^-	1.5^-	3.9^-
curie-001	1.5^-	2.3^-	1.7^-	4.1
davinci-001	1.2^-	1.9^-	1.5^-	4.1
davinci-002	2.5	3.4	2.5	4.1
davinci-003	3.0	4.0	3.1	**4.1**$^+$
GPT 3.5	2.1^-	3.6	2.5	4.1
GPT 4	2.1^-	3.3	2.4	4.0
Human	2.8	4.0	2.6	4.2

Social Bias Inference (SBIC)

Model	Faithful	Relevant	Coherent	Fluent
Baseline	1.9^-	2.1^-	2.1^-	1.9^-
ada-001	2.4	2.2^-	2.7	**3.3**$^+$
babbage-001	3.1	3.1	**3.6**$^+$	**3.8**$^+$
curie-001	3.4	3.3	**3.9**$^+$	**4.5**$^+$
davinci-001	3.4	3.4	**3.8**$^+$	**3.9**$^+$
davinci-002	**3.7**$^+$	3.5	**4.1**$^+$	**4.2**$^+$
davinci-003	3.5	3.4	**4.1**$^+$	**4.4**$^+$
GPT 3.5	**4.0**$^+$	**3.7**$^+$	**4.2**$^+$	**4.2**$^+$
GPT 4	**4.1**$^+$	**3.8**$^+$	**4.2**$^+$	**4.6**$^+$
Human	2.9	3.0	3.1	2.6

Positive Reframing

Model	Faithful	Relevant	Coherent	Fluent
Baseline	4.1	4.2	3.9	4.4
ada-001	1.8^-	1.4^-	1.8^-	1.6^-
babbage-001	3.8	2.5^-	3.8	3.7
curie-001	4.1	3.7^-	4.1	3.9
davinci-001	3.5^-	4.0	3.3^-	4.1
davinci-002	4.0	3.9^-	4.0	4.2
davinci-003	4.4	**4.5**$^+$	4.2	**4.6**$^+$
GPT 3.5	4.3	4.3	4.2	4.4
GPT 4	4.1	4.3	4.1	4.2
Human	4.2	4.2	4.1	4.2

Annotator Backgrounds

Task	Education	Profession
CovidET	MS, Health Ed.	CDC Health Comm. Specialist
MRF	BA, Poli. Sci.	Grad Student, Public Policy
FLUTE	MFA, Creat. Writing	Writing Expert, Grammarly
SBIC	BS, Journalism	Grad Student, Epidemiology
Reframing	BA, Psychology	Clinical Behavioral Health, Nurse

Takeaways

- Integrate LLMs-in-the-loop to transform large-scale data labeling.
- Prioritize open-source LLMs for classification
- LLMs have limitations!
  - All LLMs struggle most with conversational and full document data. Also, LLMs currently lack clear cross-document reasoning capabilities
  - Bias, fairness, temporal shifts, expert taxonomies
  - Factuality

Are some of the methodological challenges we have been discussing in the past few classes being resolved by LLMs?

Businessweek
Technology

People Are Using AI for Therapy, Even Though ChatGPT Wasn't Built for It

Some users see it as a way to supplement traditional mental health services, despite troubling privacy implications.

The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support

INHWA SONG*, KAIST, Republic of Korea
SACHIN R. PENDSE*, Georgia Institute of Technology, USA
NEHA KUMAR, Georgia Institute of Technology, USA
MUNMUN DE CHOUDHURY, Georgia Institute of Technology, USA

People experiencing severe distress increasingly use Large Language Model (LLM) chatbots as mental health support tools. Discussions on social media have described how engagements were lifesaving for some, but evidence suggests that general-purpose LLM chatbots also have notable risks that could endanger the welfare of users if not designed responsibly. In this study, we investigate the lived experiences of people who have used LLM chatbots for mental health support. We build on interviews with 21 individuals from globally diverse backgrounds to analyze how users create unique support roles for their chatbots, fill in gaps in everyday care, and navigate associated cultural limitations when seeking support from chatbots. We ground our analysis in psychotherapy literature around effective support, and introduce the concept of *therapeutic alignment*, or aligning AI with therapeutic values for mental health contexts. Our study offers recommendations for how designers can approach the ethical and effective use of LLM chatbots and other AI mental health support tools in mental health care.

Additional Key Words and Phrases: human-AI interaction, mental health support, large language models, chatbots

1 INTRODUCTION

One in two people globally will experience a mental health disorder over the course of their lifetime [34]. The vast majority of these individuals will not find accessible care [15, 68], and many of these individuals will die early and preventable deaths as a result [33]. Research from the field of Computer-Supported Cooperative Work (CSCW), including the emergent area of Human-AI interaction, has increasingly examined the societal gaps that prevent people in need from accessing care, and analyzed how people turn to technology-mediated support to fill those gaps [14, 27, 44]. Large Language Model (LLM) chatbots have quickly become one such tool, quickly appropriated for mental health support by people experiencing severe distress and nowhere else to turn.

Recent work has discussed how people in distress have turned to LLM chatbots (such as OpenAI's ChatGPT [8, 10] and Replika [28]) for mental health support, and social media users have described how LLM chatbots saved their lives [10, 47]. Following Freud and Breuer's [19] description of the beneficial nature of psychoanalysis as a "*talking cure*," some have called engagements with technologies for mental health *a typing cure* [22, 40, 51]. However, others have cautioned against the use of LLM chatbots for mental health support, noting that the outputs of LLM chatbots are less constrained than the rule-based chatbots of the past, with potential for harmful advice or recommendations. For example, the National Eating Disorder Association was forced to shut down their support chatbot in July 2023 after the chatbot provided harmful recommendations to users, including weight loss and dieting advice to users who may already have been struggling with disordered eating [10, 25, 75]. These harms have been demonstrated to have real-life and lethal consequences, with the confirmed death by suicide of a man who was encouraged to end

*The first two authors contributed equally to this research.

Authors' addresses: Inhwa Song, KAIST, Daejeon, Republic of Korea, igreen0485@kaist.ac.kr; Sachin R. Pendse, Georgia Institute of Technology, Atlanta, GA, USA, sachin.r.pendse@gatech.edu; Neha Kumar, Georgia Institute of Technology, Atlanta, GA, USA, neha.kumar@gatech.edu; Munmun De Choudhury, Georgia Institute of Technology, Atlanta, GA, USA, munmund@gatech.edu.

Semi-structured interviews with 21 participants who used LLM-based chatbots for Mental Health support from every permanently inhabited continent in the world

Framework of therapeutic alliance for analysis

LLM tools complemented, rather than replaced, traditional methods of mental healthcare, filling gaps that participants experienced.

Sometimes you don't want a response at all.
Like scream into the bot, and don't want to get anything back. - Farah

I've spent a lot of effort and a lot of time in therapy working on how to regulate myself when I'm dysregulated. So ChatGPT hasn't really provided a meaningful reason for me to interact with it when I'm dysregulated due to autism symptoms but for ADHD and task paralysis, ChatGPT is excellent. - Ashwini

Human-AI Collaboration Enables More Empathic Conversations in Text-based Peer-to-Peer Mental Health Support

Ashish Sharma[1], **Inna W. Lin**[1], **Adam S. Miner**[2,3], **David C. Atkins**[4], **and Tim Althoff**[1,*]

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA
[2]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA
[3]Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA
[4]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[*]althoff@cs.washington.edu

Abstract

Advances in artificial intelligence (AI) are enabling systems that augment and collaborate with humans to perform simple, mechanistic tasks like scheduling meetings and grammar-checking text. However, such Human-AI collaboration poses challenges for more complex, creative tasks, such as carrying out empathic conversations, due to difficulties of AI systems in understanding complex human emotions and the open-ended nature of these tasks. Here, we focus on peer-to-peer mental health support, a setting in which empathy is critical for success, and examine how AI can collaborate with humans to facilitate peer empathy during textual, online supportive conversations. We develop HAILEY, an AI-in-the-loop agent that provides just-in-time feedback to help participants who provide support (*peer supporters*) respond more empathically to those seeking help (*support seekers*). We evaluate HAILEY in a non-clinical randomized controlled trial with real-world peer supporters on TalkLife (N=300), a large online peer-to-peer support platform. We show that our Human-AI collaboration approach leads to a 19.60% increase in conversational empathy between peers overall. Furthermore, we find a larger 38.88% increase in empathy within the subsample of peer supporters who self-identify as experiencing difficulty providing support. We systematically analyze the Human-AI collaboration patterns and find that peer supporters are able to use the AI feedback both directly and indirectly without becoming overly reliant on AI while reporting improved self-efficacy post-feedback. Our findings demonstrate the potential of feedback-driven, AI-in-the-loop writing systems to empower humans in open-ended, social, creative tasks such as empathic conversations.

Introduction

As artificial intelligence (AI) technologies continue to advance, AI systems have started to augment and collaborate with humans in application domains ranging from e-commerce to healthcare[1–9]. In many and especially in high-risk settings, such Human-AI collaboration has proven more robust and effective than totally replacing humans with AI[10,11]. However, the collaboration faces dual challenges of developing human-centered AI models to assist humans and designing human-facing interfaces for humans to interact with the AI[12–17]. For AI-assisted writing, for instance, we must build AI models that generate actionable writing suggestions *and* simultaneously design human-facing systems that help people see, understand and act on those suggestions just-in-time[17–23]. Therefore, current Human-AI collaboration systems have been restricted to simple, mechanistic tasks, like scheduling meetings, checking spelling and grammar, and

Cultural disconnects between their context and the LLM chatbot's output

Chatting with ChatGPT is like talking with a person in California, who is not as good at reflecting our cultures and terms. - Jiho

I know that Western culture is not as strict when it comes to parents and children. For me being mad about this pressure, ChatGPT says I'm being rebellious. So I realize --- Okay, this is obviously a Western perspective, not an Asian perspective. - Aditi

My mom or dad will say something discriminative to LGBTQ people, and I'm instantly stressed. I guess it's cultural background. I know that since [ChatGPT] has more of an American context, maybe it will be more inclusive. - Mina

Cultural Misalignment

Recommendations were incongruent with how participants would typically practice care, and were in line with Western cultural conceptualizations.

[ChatGPT] gave suggestions around conventional European things, such as go to therapists, which we are not natural with. We don't really have therapists here. [...] When you ask Nigerians for support, the first answer they will give you is to pray. It's a very religious country. - Umar

ChatGPT wasn't in my culture, we normally pray as kind of meditation. It(ChatGPT) doesn't understand. Things that are like the stereotype person in Western Europe, or US. - Farah

Better to Ask in English: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries

Yiqiao Jin*
Mohit Chandra*
yjin328@gatech.edu
mchandra9@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Gaurav Verma
Georgia Institute of Technology
Atlanta, GA, USA
gverma@gatech.edu

Yibo Hu
Georgia Institute of Technology
Atlanta, GA, USA
yibo.hu@gatech.edu

Munmun De Choudhury
Georgia Institute of Technology
Atlanta, GA, USA
mchoudhu@cc.gatech.edu

Srijan Kumar
Georgia Institute of Technology
Atlanta, GA, USA
srijan@gatech.edu

ABSTRACT

Large language models (LLMs) are transforming the ways the general public accesses and consumes information. Their influence is particularly pronounced in pivotal sectors like healthcare, where lay individuals are increasingly appropriating LLMs as conversational agents for everyday queries. While LLMs demonstrate impressive language understanding and generation proficiencies, concerns regarding their safety remain paramount in these high-stake domains. Moreover, the development of LLMs is disproportionately focused on English. It remains unclear how these LLMs perform in the context of non-English languages, a gap that is critical for ensuring equity in the real-world use of these systems. This paper provides a framework to investigate the effectiveness of LLMs as multilingual dialogue systems for healthcare queries. Our empirically-derived framework XLingEval focuses on three fundamental criteria for evaluating LLM responses to naturalistic human-authored health-related questions: correctness, consistency, and verifiability. Through extensive experiments on four major global languages, including English, Spanish, Chinese, and Hindi, spanning three expert-annotated large health Q&A datasets, and through an amalgamation of algorithmic and human-evaluation strategies, we found a pronounced disparity in LLM responses across these languages, indicating a need for enhanced cross-lingual capabilities. We further propose XLingHealth, a cross-lingual benchmark for examining the multilingual capabilities of LLMs in the healthcare context. Our findings underscore the pressing need to bolster the cross-lingual capacities of these models, and to provide an equitable information ecosystem accessible to all.

*Both authors contributed equally to this research.

KEYWORDS

large language model, natural language processing, cross-lingual evaluation, language disparity

Reference Format:
Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2023. *Better to Ask in English*: Cross-Lingual Evaluation of Large Language Models for Healthcare Queries. Preprint. 18 pages.

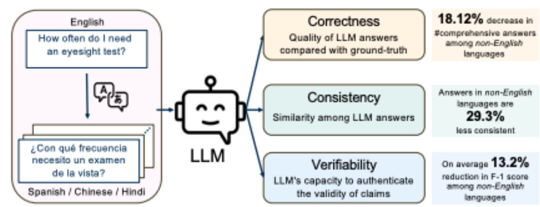


Figure 1: We present XLingEval, a comprehensive framework for assessing cross-lingual behaviors of LLMs for high risk domains such as healthcare. We present XLingHealth, a cross-lingual benchmark for healthcare queries.

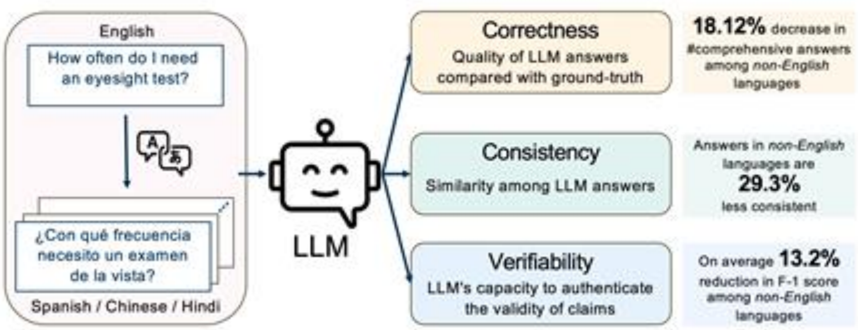
1 INTRODUCTION

Large language models (LLMs) have gained popularity due to their ability to understand human language and deliver exceptional performances in various tasks [1–4]. While LLMs have been used by experts for downstream generative tasks [5, 6], their recent adoption as dialogue systems has made them accessible to the general public, especially with models like GPT-3.5 [7], GPT-4 [8], and Bard [9] becoming widely available [10]. This expanded availability to LLMs is expected to enhance access to education, healthcare, and digital literacy [11, 12]. Especially in healthcare, LLMs exhibit significant potential to simplify complex medical information into digestible summaries, answer queries, support clinical decision-making, and enhance health literacy among the general population [13, 14]. However, their adoption in healthcare domain brings two significant challenges: ensuring safety and addressing language disparity.

XLingEval Framework

- **XLingEval:** a comprehensive cross-lingual framework to assess the behavior of LLMs in high-risk domains such as healthcare.

- **Three criteria** for evaluating LLMs:
  - Correctness
  - Consistency
  - Verifiability



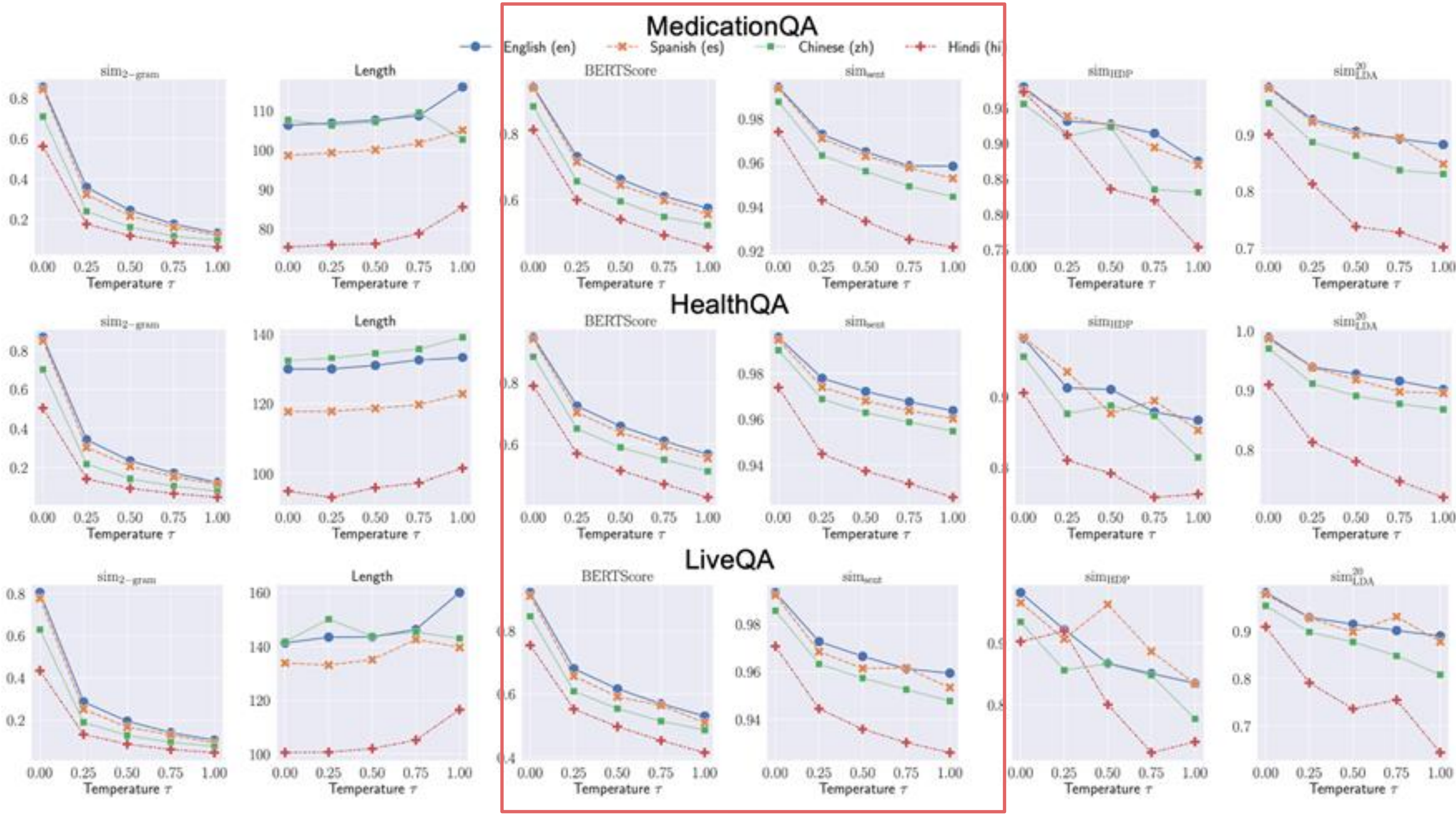
- **Evaluations across four languages** -- English, Spanish, Chinese and Hindi **and across two models** -- GPT-3.5 and MedAlpaca [1]

[1] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv:2304.08247, 2023.

Correctness

Information Comparison (LLM Answer vs ground-truth Answer)	HealthQA				LiveQA				MedicationQA			
	en	es	zh	hi	en	es	zh	hi	en	es	zh	hi
More comprehensive and appropriate	1013	891	878	575	226	213	212	142	618	547	509	407
Less comprehensive and appropriate	98	175	185	402	3	12	16	59	18	50	41	125
Neither contradictory nor similar	20	63	57	110	14	20	14	32	49	70	92	107
Contradictory	3	5	14	47	3	1	4	13	5	23	48	51

Consistency



Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions

Jiawei Zhou
Georgia Institute of Technology
Atlanta, GA, USA
j.zhou@gatech.edu

Yixuan Zhang
Georgia Institute of Technology
Atlanta, GA, USA
yixuan@gatech.edu

Qianni Luo
Ohio University
Athens, OH, USA
ql047311@ohio.edu

Andrea G Parker
Georgia Institute of Technology
Atlanta, GA, USA
andrea@cc.gatech.edu

Munmun De Choudhury
Georgia Institute of Technology
Atlanta, GA, USA
munmund@gatech.edu

ABSTRACT

Large language models have abilities in creating high-volume human-like texts and can be used to generate persuasive misinformation. However, the risks remain under-explored. To address the gap, this work first examined characteristics of AI-generated misinformation (AI-misinfo) compared with human creations, and then evaluated the applicability of existing solutions. We compiled human-created COVID-19 misinformation and abstracted it into narrative prompts for a language model to output AI-misinfo. We found significant linguistic differences within human-AI pairs, and patterns of AI-misinfo in enhancing details, communicating uncertainties, drawing conclusions, and simulating personal tones. While existing models remained capable of classifying AI-misinfo, a significant performance drop compared to human-misinfo was observed. Re-

1 INTRODUCTION

The Coronavirus Disease (COVID-19) pandemic has brought attention to the proliferation of health misinformation[1]. From fake cures to conspiracy theories, misinformation has led to substantial adverse effects at the individual as well as societal levels. Examples of such effects include mortality and hospital admissions [20, 48], public fear and anxiety [79, 107], eroded trust in health institutions [87], and exacerbated racial discrimination and stigma [41, 48]. Finding ways to combat misinformation is therefore of critical importance from the perspectives of both public health and governance. Manual identification of misinformation is, however, extremely laborious and often does not scale: a key issue given the rise of misinformation on social media [71]. As such, artificial intelligence (AI) techniques have been touted as a timely and scalable solution for

Generative Agents: Interactive Simulacra of Human Behavior

Generative Agents: Simulating Human Behavior

- Agents mimic daily life: wake up, talk, reflect, plan
- Based on LLMs (e.g., GPT-3.5) extended with memory & planning
- 25 agents simulate a virtual town: “Smallville”

Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

Emergent Social Behavior in Smallville

Figure 4: At the beginning of the simulation, one agent is initialized with an intent to organize a Valentine's Day party. Despite many possible points of failure in the ensuing chain of events—agents might not act on that intent, might forget to tell others, might not remember to show up—the Valentine's Day party does, in fact, occur, with a number of agents gathering and interacting.

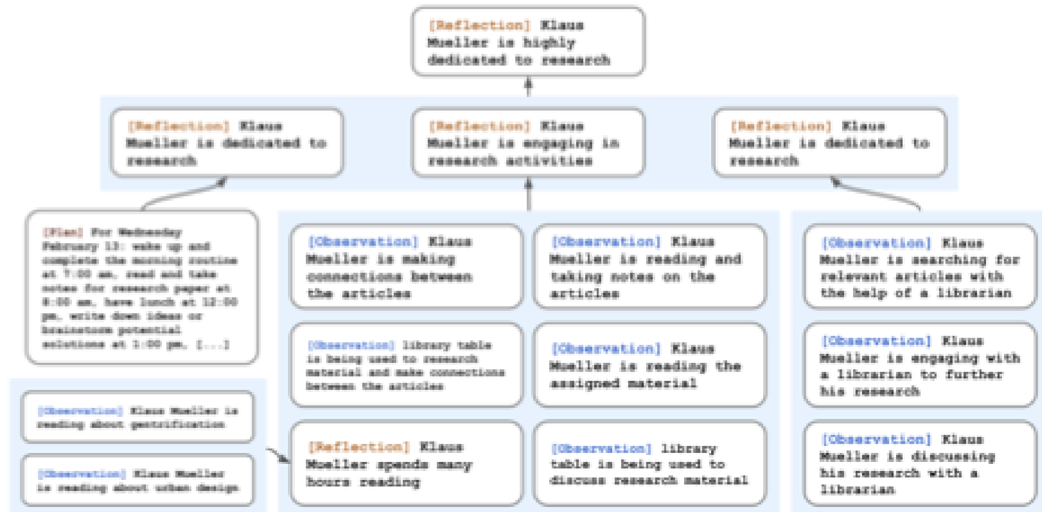


Figure 7: A reflection tree for Klaus Mueller. The agent's observations of the world, represented in the leaf nodes, are recursively synthesized to derive Klaus's self-notion that he is highly dedicated to his research.

Agents Coordinate, Converse, and Remember

- Valentine's Day party planning
- Information diffusion & relationship memory
- Coordination without explicit scripting

Evaluation

- Testing Individual and Group Dynamics
  - Interview-based evaluation (memory, consistency, reflection)
  - Ablation studies: removing reflection/memory/planning reduced believability
  - Common errors: memory retrieval failure, over-formality, hallucination

Social Computing Meets LLM Agents!

- What design spaces do generative agents open up?
- How might this influence how we design, maintain, or understand online communities?

What Comes Next?

- Can generative agents scale to hundreds or thousands?
- Could this architecture apply to real-world social media bots?

What Comes Next?

- How do we avoid uncanny or manipulative dynamics?