Stanford CS224v Course
Conversational Virtual Assistants with Deep Learning

Lecture 5

Improving and Evaluation of ChatBots

Monica Lam & Sina Semnani

WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia
Sina J. Semnani, Violet Z. Yao*, Heidi C. Zhang*, Monica S. Lam
In Proceedings of Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023.
Winner of the Wikimedia's Research of the Year Award, 2024.

Chatbots in Practice

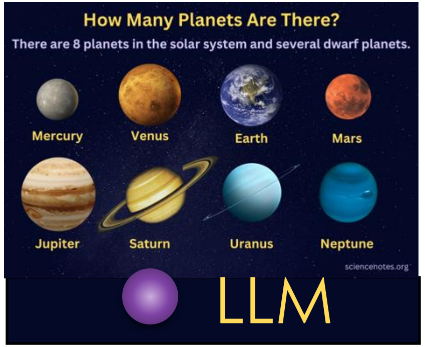
Has to be Close to Human Performance
(unlike NLP research papers)

Technique: Emulate Humans

Lecture Goal

1. How to improve and evaluate chatbots?
    - Papers only describe the "final product"
    - Focus on this lecture:
      What are the considerations in creating the product?
      Very useful for your project!
2. Show how to apply the lecture to 2 variations of WikiChat
    - Going beyond one language: Multilingual Chatbots
    - Going beyond fact-based bots: Persuasive Chatbots

The LLM Planet


How Many Planets Are There?
There are 8 planets in the solar system and several dwarf planets.
Mercury
Venus
Earth
Mars
Jupiter
Saturn
Uranus
Neptune
sciencenotes.org
LLM

NLP without LLMs

- Easy problems (e.g. classification)
- Start with a data set
- To get better results:
  - Change the hyper parameters
  - Fine-tune with more data
- Evaluation metrics:
  - F1 on gold
  - BLEU (e.g. translation)
  - ROUGE (e.g. summary)

NLP with LLMs

- Hard problems: Essays, copilot, …
- How to create a dataset?
- To get better results:
  - Change the prompts, pipeline
  - Quiz: what does fine-tuning do?
- What evaluation metrics?
  - GPT results are often better than gold
  - BLEU: meaningless
  - ROUGE: meaningless

Fine-Tuning LLMs

- Why are we fine-tuning the largest models?
  - General capabilities: alignment, instruction following
  - Quiz: is it good for specific tasks?
- Why are we distilling to (fine-tuning) smaller models
  - Distilling: This needs an LLM teacher model
    Use a teaching model to generate data for smaller model.
  - Faster, cheaper, more privacy

  With a lower performance!

Key Idea

~~Collect Data, Train and Tune Hyperparameters~~

Step 1: Assess and Improve with Tough Examples

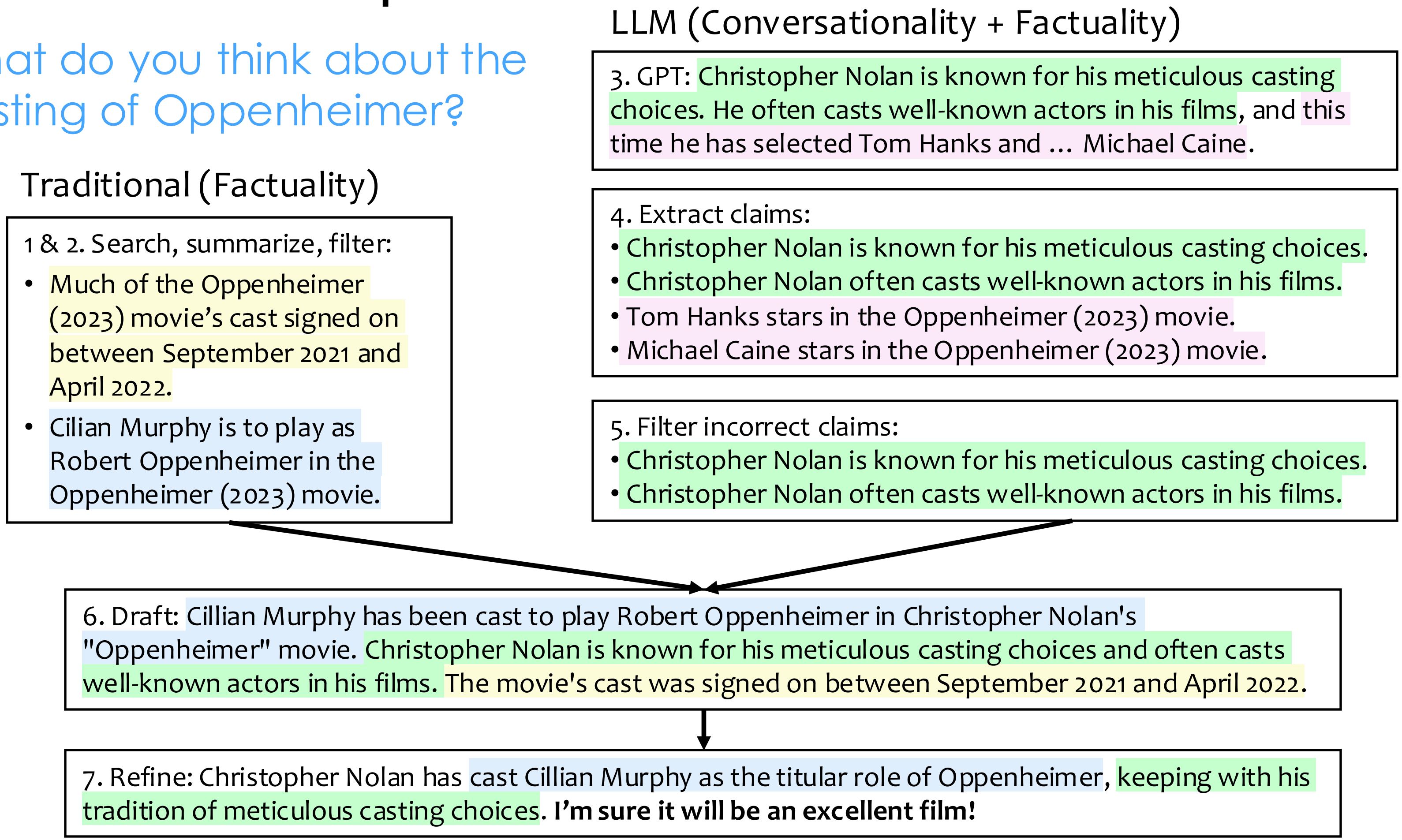
Step 2: Evaluate Automatically and Manually

Case Study: WikiChat

- The first time we faced the problem of assessing LLM pipeline
- Tendency: collect data set and evaluate, repeat
- But why bother with evaluation,
  if you know it does not work on hard cases?
  - We know that it will work poorly in practice!

Recall the WikiChat Pipeline

What do you think about the casting of Oppenheimer?



Traditional (Factuality)

1 & 2. Search, summarize, filter:

- Much of the Oppenheimer (2023) movie's cast signed on between September 2021 and April 2022.
- Cilian Murphy is to play as Robert Oppenheimer in the Oppenheimer (2023) movie.

LLM (Conversationality + Factuality)

3. GPT: Christopher Nolan is known for his meticulous casting choices. He often casts well-known actors in his films, and this time he has selected Tom Hanks and … Michael Caine.

4. Extract claims:
- Christopher Nolan is known for his meticulous casting choices.
- Christopher Nolan often casts well-known actors in his films.
- Tom Hanks stars in the Oppenheimer (2023) movie.
- Michael Caine stars in the Oppenheimer (2023) movie.

5. Filter incorrect claims:
- Christopher Nolan is known for his meticulous casting choices.
- Christopher Nolan often casts well-known actors in his films.

6. Draft: Cillian Murphy has been cast to play Robert Oppenheimer in Christopher Nolan's "Oppenheimer" movie. Christopher Nolan is known for his meticulous casting choices and often casts well-known actors in his films. The movie's cast was signed on between September 2021 and April 2022.

7. Refine: Christopher Nolan has cast Cillian Murphy as the titular role of Oppenheimer, keeping with his tradition of meticulous casting choices. **I'm sure it will be an excellent film!**

What Was the Process?

- High level design
  - Started with generation only first
  - Realized we needed RAG because of new information
- Iterative improvement
  - LLM hallucination with RAG on many examples – we were surprised!
    - Introduce filtering steps and change prompts to make it say "I don't know"
  - Came up with hard cases to break the system: e.g. time
- Finally, redid the pipeline with small, well-designed steps!
  - It just started to make good conversations!

Evaluation Takes Many More Months

Old Data Sets Do Not Work

No simple numeric benchmark to iterate on

- Evaluation conversations in the past were *crowdsourced*
  - Ask crowdworkers to talk to each other about a familiar topic
  - Have blind spots like *tail* or *expert* topics
- Evaluation conversations were *static*
  - Get outdated quickly
    - Wizard of Wikipedia dataset topics in GPT-3’s pre-training data
    - New benchmarks used for pre-training in the future

What We Need to Improve/Eval WikiChat?

- How to evaluate tail questions?
  - Humans do not know tail topics!
- How to evaluate *LLM conversations*?
  - How to compare the results between systems since their dialogue diverges after the first turn!
  - Cannot get a numeric result to measure success
- How to minimize human effort, cost, and potential errors?

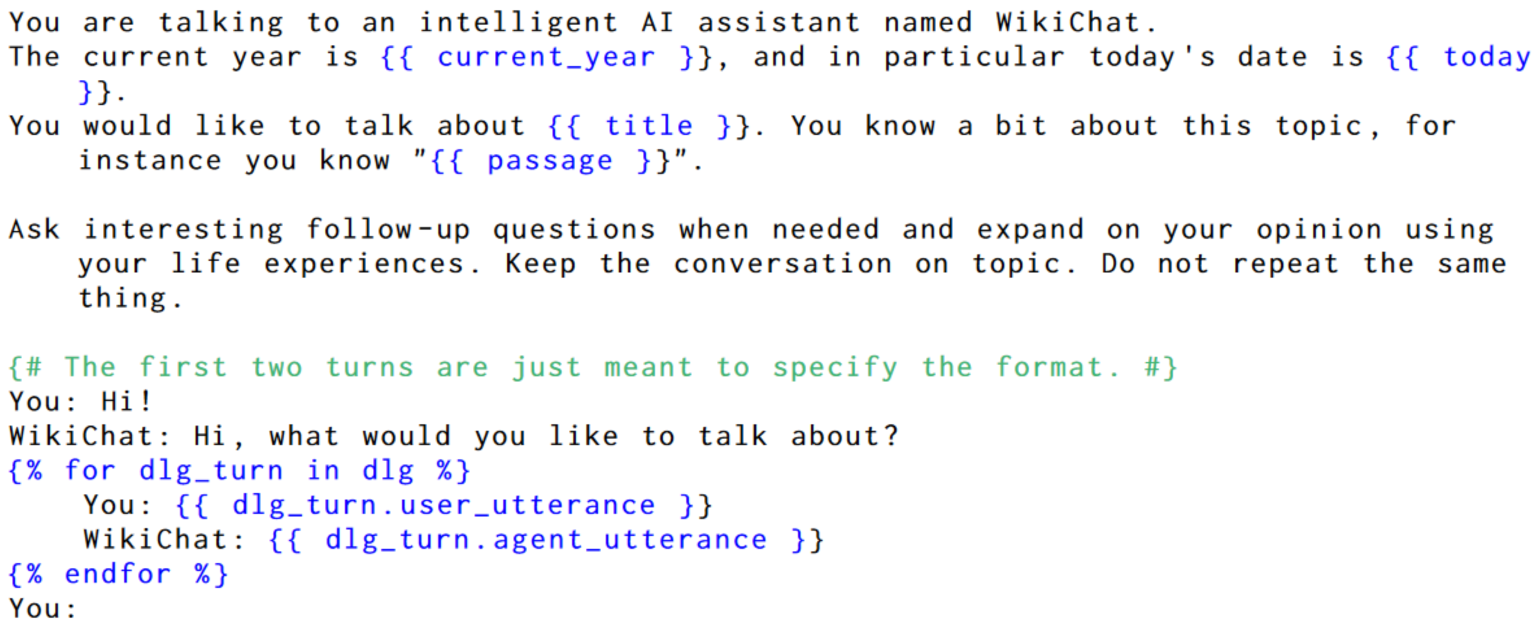
Solution: LLMs come to the rescue!

What Can LLMs Do and Not Do?

1. **Automatically generate inputs (simulated conversations)**
   - The user is GPT-4; the agent is the Chat pipeline
     - E.g. conversations: Head (most viewed pages); Tail (least viewed pages); Recent (most edited in 2023)

   - **Are LLM conversations similar to human conversations?**
     - Improve and Assess
       - Prompt with different user profiles to generate diverse user input; eyeball result
     - Prove with Evaluation
       - Conduct user study on a subset, show the results match

Simulation Prompt



```
You are talking to an intelligent AI assistant named WikiChat.
The current year is {{ current_year }}, and in particular today's date is {{ today
    }}.
You would like to talk about {{ title }}. You know a bit about this topic, for
    instance you know "{{ passage }}".

Ask interesting follow-up questions when needed and expand on your opinion using
    your life experiences. Keep the conversation on topic. Do not repeat the same
    thing.

{# The first two turns are just meant to specify the format. #}
You: Hi!
WikiChat: Hi, what would you like to talk about?
{% for dlg_turn in dlg %}
    You: {{ dlg_turn.user_utterance }}
    WikiChat: {{ dlg_turn.agent_utterance }}
{% endfor %}
You:
```

What Can LLMs Do and Not Do?

2. **Automatically evaluate if possible**
   - **Are LLM evaluations similar to human evaluations?**
     - Compare with one or more humans on a subset

Quiz: Is it possible to automate evaluation of

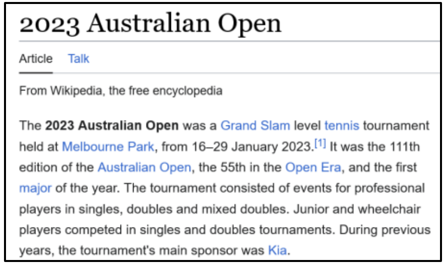
- Conversationality
- Fact checking

G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment, Liu et al, arXiv 2023

Evaluation Summary

Conversation Generation

Head/Tail/Recent

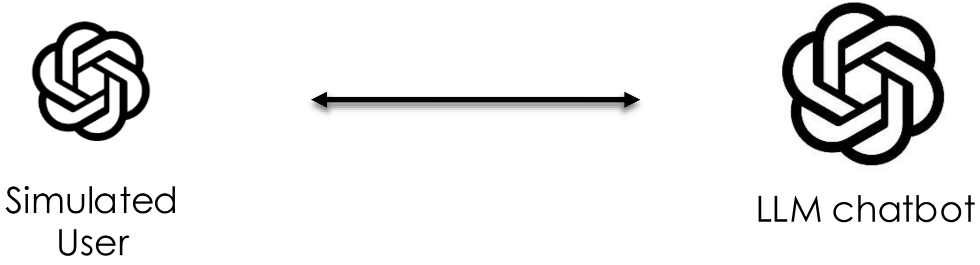


2023 Australian Open

Article Talk

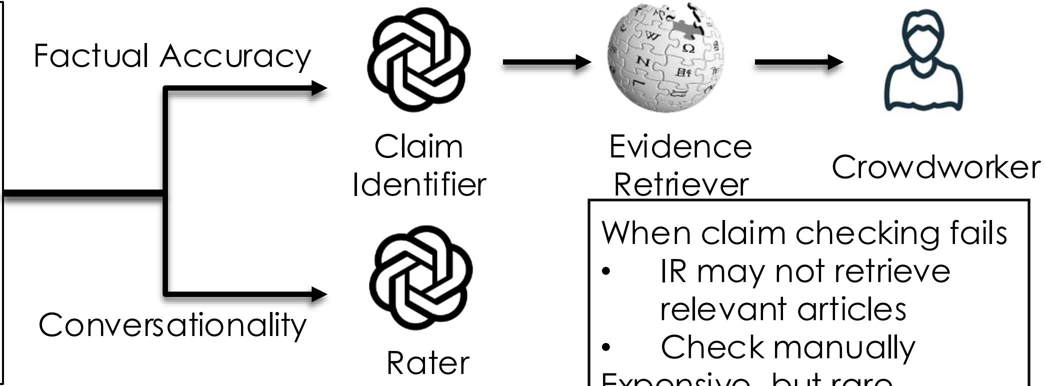
From Wikipedia, the free encyclopedia

The **2023 Australian Open** was a Grand Slam level tennis tournament held at Melbourne Park, from 16–29 January 2023.[1] It was the 111th edition of the Australian Open, the 55th in the Open Era, and the first major of the year. The tournament consisted of events for professional players in singles, doubles and mixed doubles. Junior and wheelchair players competed in singles and doubles tournaments. During previous years, the tournament's main sponsor was Kia.



Evaluation

User: I would like to talk about the 2023 Australian Open Men's singles tennis championship. Did you hear about it?
Chatbot: Yes, I did hear about it! Novak Djokovic won the 2023 Australian Open Men's singles tennis championship, claiming his tenth Australian Open title and 22nd major title overall. It was an impressive victory!



When claim checking fails
- IR may not retrieve relevant articles
- Check manually

Expensive, but rare

Implementation of WikiChat

- We use ColBERTv2 for IR
  - Is fast (~100 ms query latency)
  - Is trained on MS MARCO (2016)
- Wikipedia corpus obtained on 4/28/2023
- We use text-davinci-003 or gpt-4 as the LLM for all prompts

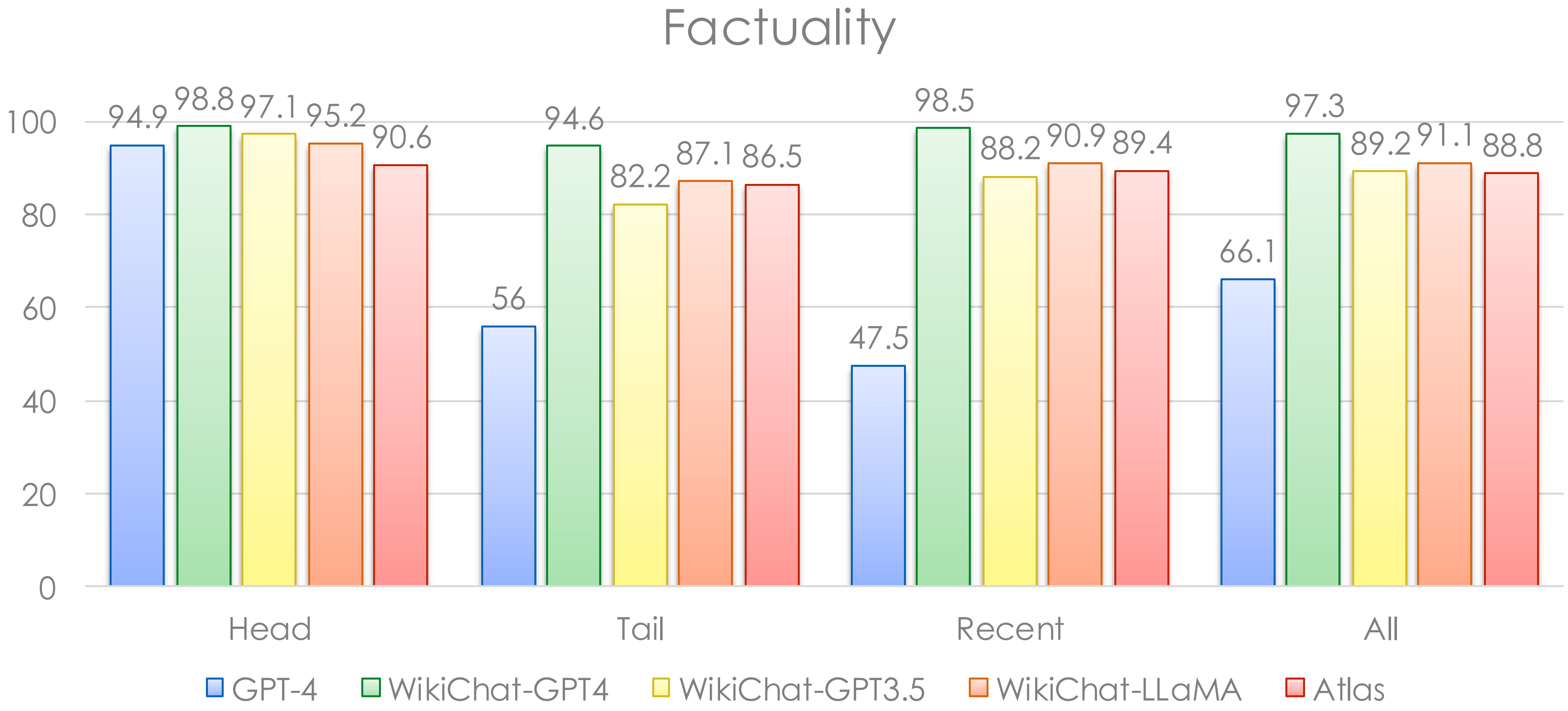
ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction, Santhanam et al, NAACL 2022

Experiment

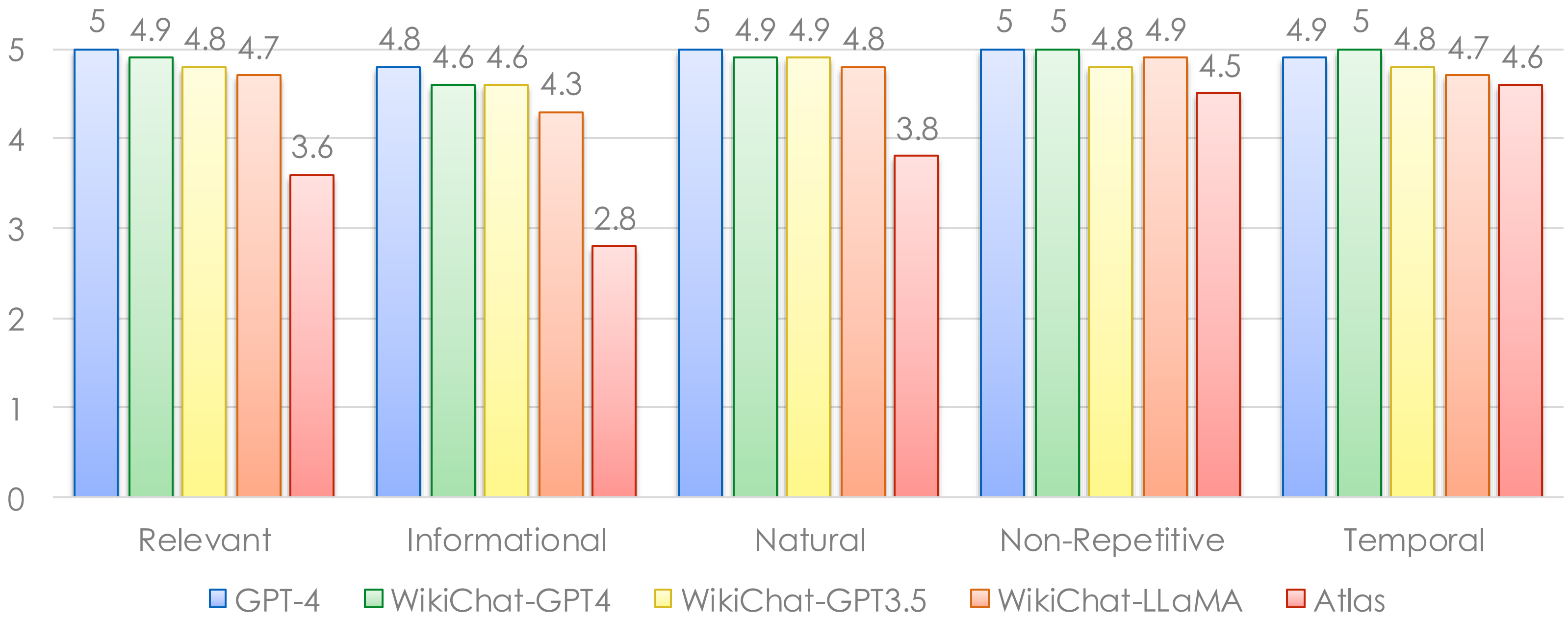
- Baselines
  - Fine-tuned models
    - Atlas: state-of-the-art on many knowledge-grounded tasks
    - 3B parameters
  - LLMs
    - Text-davinci-003 or GPT-4, prompted to be a chatbot

- Each evaluated for 600 dialog turns

Results – Factual Accuracy



Results - Conversationality



Design Analysis: Do we need both IR and LLM?

Information Retrieval and LLM are color-coded

User: Can we talk about Sting, the musician?
WikiChat: Sure! Sting is an English musician and actor, best known as the frontman, songwriter and bassist for the new wave rock band the Police. He has released several solo albums and has won numerous awards, including 17 Grammy Awards and a Golden Globe.

User: […] What do you think makes Sting's music so unique?
WikiChat: Sting's music is often very personal and introspective, and he often uses his music to express his own thoughts and feelings. He also has a great ability to blend different genres and create unique sounds.

User: Do you think his birthplace has influenced his music?

WikiChat: Yes, Sting has been heavily influenced by his birthplace of Newcastle upon Tyne, England. As a child, he was inspired by the Queen waving at him from a Rolls-Royce, which diverted him from the shipyard prospect towards a more glamorous life. He was also influenced by the music he heard in local nightclubs he visited such as Club A'Gogo.

Analyzing the Design

- Analysis of each component
  - Do we need 2 pipelines? YES
    - 27% of facts come from LLM generation

  - Do we need to fact-check LLM output: YES
    - Failed: 20% (head); 45% (tail); 56% (recent)

Quiz:

Are We Done Analyzing WikiChat?

Stanford University

Real User Evaluation

- Real user evaluation was seldom performed before LLM because of poor accuracy
- Most important metric!

User study: User reads the first sentence of a new Wikipedia page

Model	User Rating (out of 5)	Factuality
GPT-4	3.4	42.9%
WikiChat using GPT-4	3.8	97.9%

GPT-4: Users are not even aware that over half of the statements are false

Lecture Goal

1. How to improve and evaluate chatbots?
    - Papers only describe the “final product”
    - Focus on this lecture:
      What are the considerations in creating the product?
      Very useful for your project!
2. Show how to apply the lecture to 2 variations of WikiChat
    - Going beyond one language: **Multilingual Chatbots**
    - Going beyond fact-based bots: Persuasive Chatbots

Can We Remove the Language Barrier for WikiChat?

Where GPT-4 and Wikipedia Fail

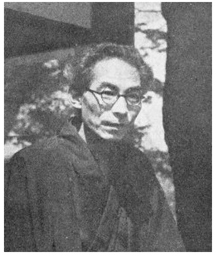
I've heard about **Yoshio Toyoshima**, what should I read from him?

WIKIPEDIA
The Free Encyclopedia

❌

Only ~11% of articles are in English
(6.7M out of 61.8M)

ウィキペディア
フリー百科事典



Yoshio Toyoshima (November 27, 1890 - June 18, 1955) was a Japanese novelist, translator, French literature scholar, and children's literature author ...

Where GPT-4 and Wikipedia Fail

How did **Murakami** get started as a writer?

WIKIPEDIA
The Free Encyclopedia

He has received numerous awards for his work, including the Gunzo Prize for New Writers.

Haruki Murakami

ウィキペディア
フリー百科事典

On April 1, 1978, while lying on the lawn of the outfield seats at Meiji Jingu Baseball Stadium, watching the opening game of professional baseball, Yakult vs. Hiroshima, and drinking beer, he decided to write a novel. It was the moment when Dave Hilton, the leadoff hitter for Yakult in the bottom of the first inning, hit a double to left-center field. From then on, he continued to write at the kitchen table every night while running the jazz café …

Try this for yourselves!

- You will find a lot of unexpected ones
- In all language directions

ویکی‌پدیا
دانشنامهٔ آزاد

WIKIPEDIA
The Free Encyclopedia

ウィキペディア
フリー百科事典

Avicenna (the Persian physician) became the personal physician to Shams al-Dawla, the ruler of Hamadan, and was recognized for his ability to treat Shams al-Dawla’s **hernia**.

History of Medicine
(by Akira Kazita)

Why Multilingual RAG?

- People speak different languages

But also:

- Expanding Access to Cultural Knowledge
- Why chatbots? Can't we just translate Wikipedia?
  - People won't seek what they don't know.
  - A chatbot can offer information when it is relevant.

Is Multilingual Easy with LLMs?

Two building blocks of retrieval-augment pipelines:

1. **LLM**: strong LLMs already can understand and speak in a lot of languages
   - Though may struggle with low-resource ones like Tamil, Burmese etc.
2. **Retriever**: Use a cross-lingual retriever (BGE-M3, GTE)

In WikiChat, we simply tell the *draft* stage to respond in the same language as the user utterance

The difficulty: entities in different languages

Multilingual Entities are Hard!

- LLMs and Retrievers both struggle with multilingual entities
  - How to transliterate names?
  - Entities (books, movies, even people)
    have different names in different languages
- Solution
  - Use Wikidata entity translations to augment the text corpus

Example

User **WikiChat**

Is Hsiang-Tsung Kung a descendant of Confucius"

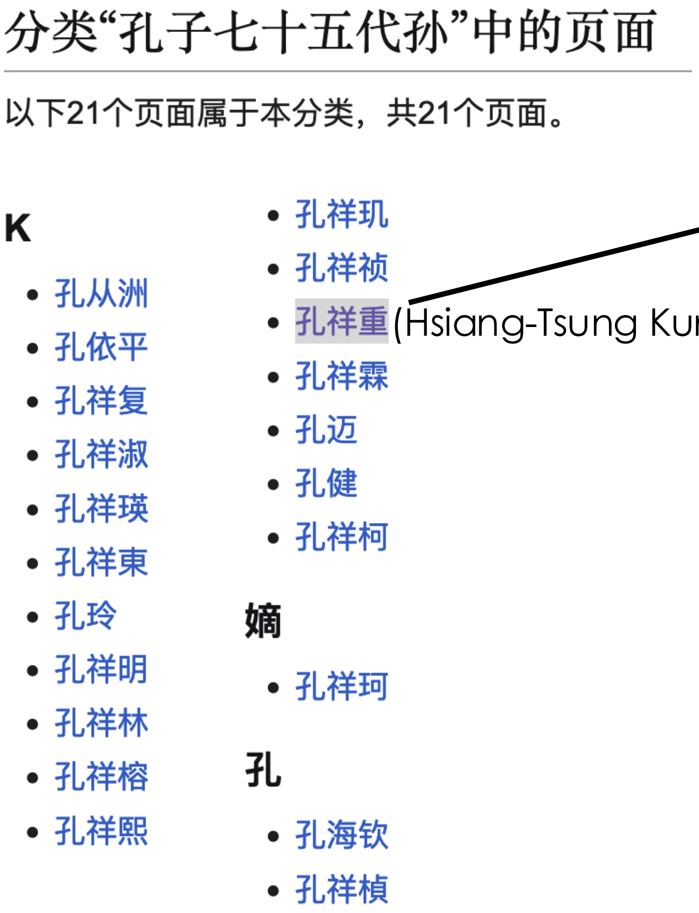
I'm not certain about Hsiang-Tsung Kung's lineage,
but it's possible given the historical connections.
If you have more specific details, I might be able to help further.

孔祥重是孔子的后裔?

是的，孔祥重是孔子的第75代子孙。

Hsiang-Tsung Kung is 孔祥重

Yes, Hsiang-Tsung Kung (孔祥重) is a
75th-generation descendant of Confucius.



分类“孔子七十五代孙”中的页面

以下21个页面属于本分类，共21个页面。

K

- 孔从洲
- 孔依平
- 孔祥复
- 孔祥淑
- 孔祥瑛
- 孔祥東
- 孔玲
- 孔祥明
- 孔祥林
- 孔祥榕
- 孔祥熙
- 孔祥玑
- 孔祥祯
- 孔祥重 (Hsiang-Tsung Kung)
- 孔祥霖
- 孔迈
- 孔健
- 孔祥柯

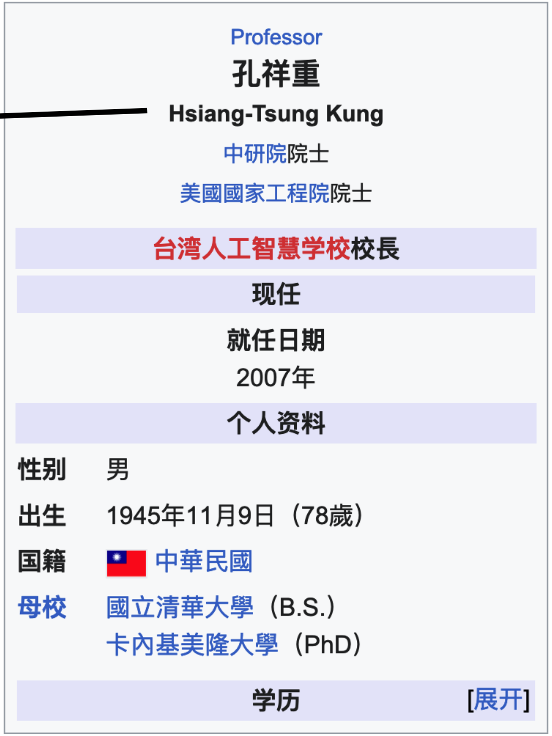
嫡

- 孔祥珂

孔

- 孔海钦
- 孔祥楨

https://zh.wikipedia.org/wiki/孔祥重



Professor

孔祥重

Hsiang-Tsung Kung

中研院院士

美國國家工程院院士

台湾人工智慧学校校長

现任

就任日期

2007年

个人资料

性别	男
出生	1945年11月9日（78歲）
国籍	中華民國
母校	國立清華大學（B.S.） 卡內基美隆大學（PhD）

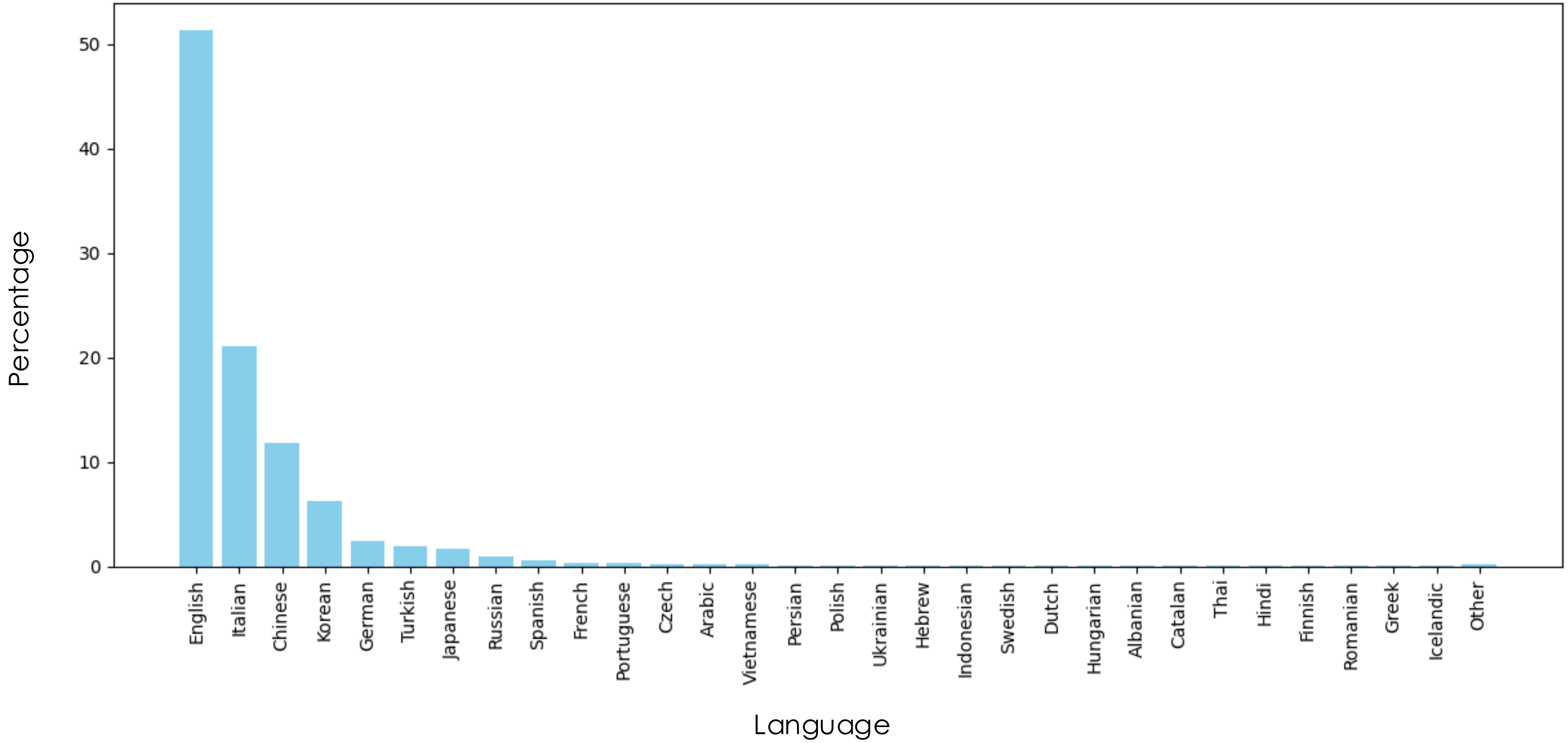
学历 [展开]

MultiLingual WikiChat

- 10 Wikipedias
  English German French Spanish Japanese
  Russian Portuguese Italian Chinese Farsi
- 35K+ questions so far
- Impact
  - A lot more knowledge for all
  - Esp. makes English encyclopedia accessible to all

Breaking news:
Now 30 languages!

35K+ Multilingual Conversations with WikiChat



Quiz

What Other Multilingual Use Cases You Can Think Of ?

Summary

- Improve and assess step:
  - We already know about the entity linking problem*
    - Confirmed to be a problem with WikiChat
    - Easy for Wikipedia because all entities are linked
  - Our assessment focused on the entity problem
- Evaluation is difficult
  - Especially with the many different languages
  - To be done with real human user feedback in the wild

---

* *Contextual Semantic Parsing for Multilingual Task-Oriented Dialogues*
Mehrad Moradshahi, Victoria C. Tsai, Giovanni Campagna, Monica S. Lam
In Proc. of the European Chapter of the Association for Computational Linguistics (EACL) , Dubrovnik, Croatia, May 2-6 2023.

Lecture Goal

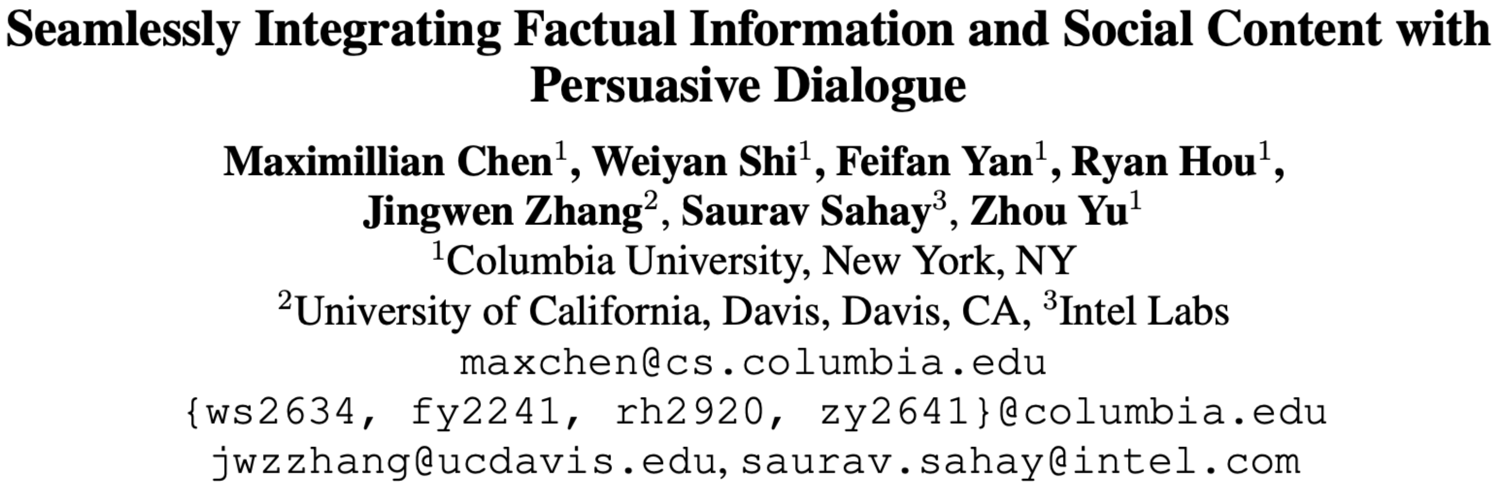
1. How to improve and evaluate chatbots?
   - Papers only describe the "final product"
   - Focus on this lecture:
     What are the considerations in creating the product?
     Very useful for your project!
2. Show how to apply the lecture to 2 variations of WikiChat
   - Going beyond one language: Multilingual Chatbots
   - Going beyond fact-based bots: **Persuasive Chatbots**

Can We Create Domain-Independent Chatbots beyond Providing Facts?

Persuasion!

Persuasive Chatbots

- E.g. donation to charity, health intervention, travel, etc.
- Persuasion: To influence users with specific goals
    - Establish mutual trust and credibility
    - Strategically presenting persuasive appeals
    - Eliciting emotional reactions
- Persuasion skills needed
    - Must provide factual answers, nonetheless



Seamlessly Integrating Factual Information and Social Content with Persuasive Dialogue

Maximillian Chen[1], **Weiyan Shi**[1], **Feifan Yan**[1], **Ryan Hou**[1],
Jingwen Zhang[2], **Saurav Sahay**[3], **Zhou Yu**[1]
[1]Columbia University, New York, NY
[2]University of California, Davis, Davis, CA, [3]Intel Labs
`maxchen@cs.columbia.edu`
`{ws2634, fy2241, rh2920, zy2641}@columbia.edu`
`jwzzhang@ucdavis.edu`, `saurav.sahay@intel.com`

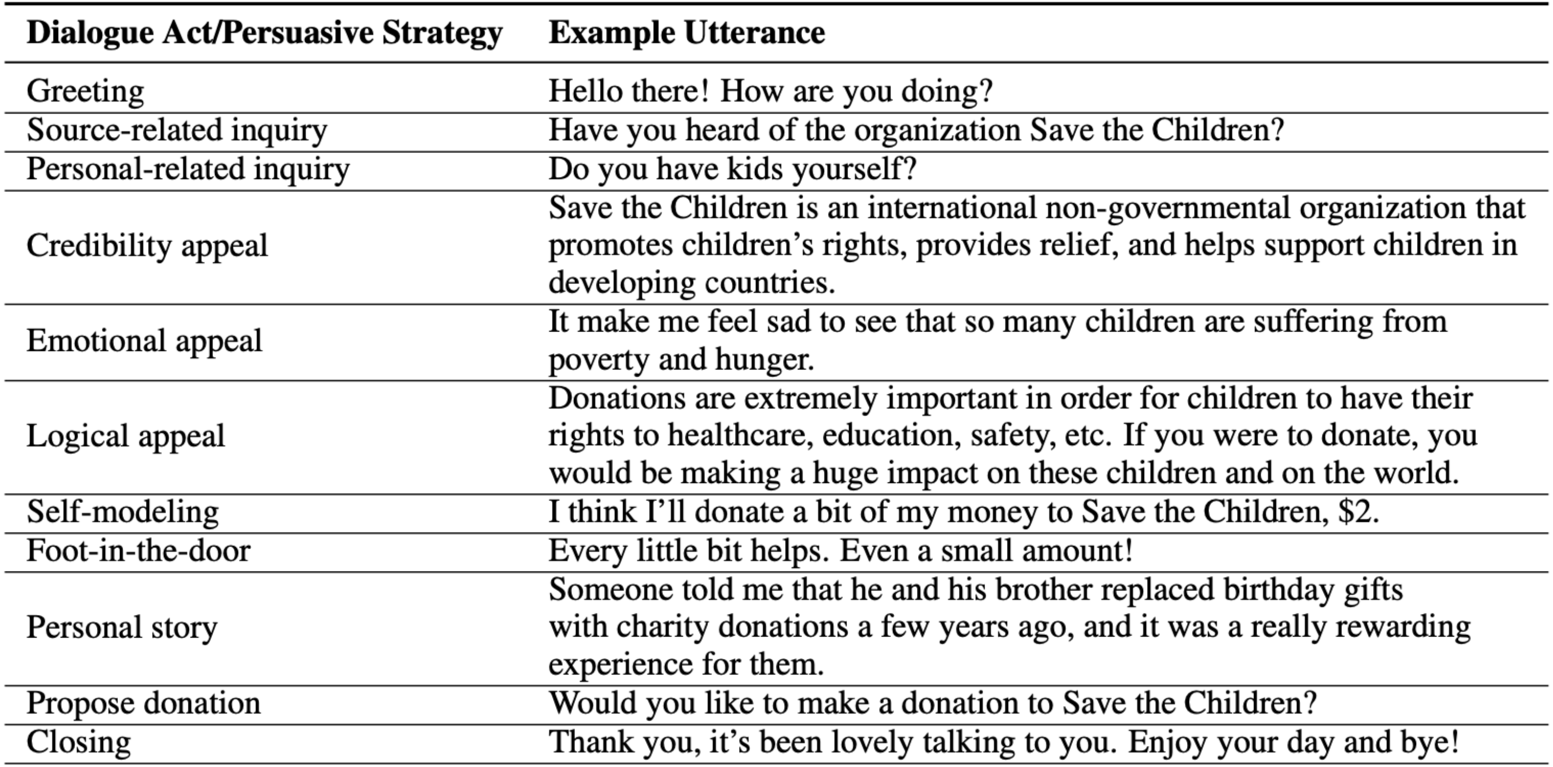
in Proceedings of AACL-IJCNLP 2022

(The proposed system is called RAP)

Dialogue Acts

- RAP follows the traditional line of work in conversational agents
  - Characterize the agents' utterances
    with a fixed set of dialogue acts (intents of agents)
    [John L. Austin, 1959]
  - General acts: E.g. "ask question", "state a fact", …
  - Persuasive acts: Emotional appeal, Logical appeal

RAP Dialogue Acts for “Save-the-Children”



Dialogue Act/Persuasive Strategy	Example Utterance
Greeting	Hello there! How are you doing?
Source-related inquiry	Have you heard of the organization Save the Children?
Personal-related inquiry	Do you have kids yourself?
Credibility appeal	Save the Children is an international non-governmental organization that promotes children’s rights, provides relief, and helps support children in developing countries.
Emotional appeal	It make me feel sad to see that so many children are suffering from poverty and hunger.
Logical appeal	Donations are extremely important in order for children to have their rights to healthcare, education, safety, etc. If you were to donate, you would be making a huge impact on these children and on the world.
Self-modeling	I think I’ll donate a bit of my money to Save the Children, $2.
Foot-in-the-door	Every little bit helps. Even a small amount!
Personal story	Someone told me that he and his brother replaced birthday gifts with charity donations a few years ago, and it was a really rewarding experience for them.
Propose donation	Would you like to make a donation to Save the Children?
Closing	Thank you, it’s been lovely talking to you. Enjoy your day and bye!

Table 1: Examples of each dialogue act from PERSUASIONFORGOOD used for the chatbot.

Strategy

- In previous work, they establish that the agent “the best” by following a strict order (as shown in the last slide)

Quiz: Why a strict order?

- Expensive approach:
  he strategy must be derived for each persuasion task
  - Dialogue acts
  - The order in the strategy

RAP (Response Agenda Pushing) Framework

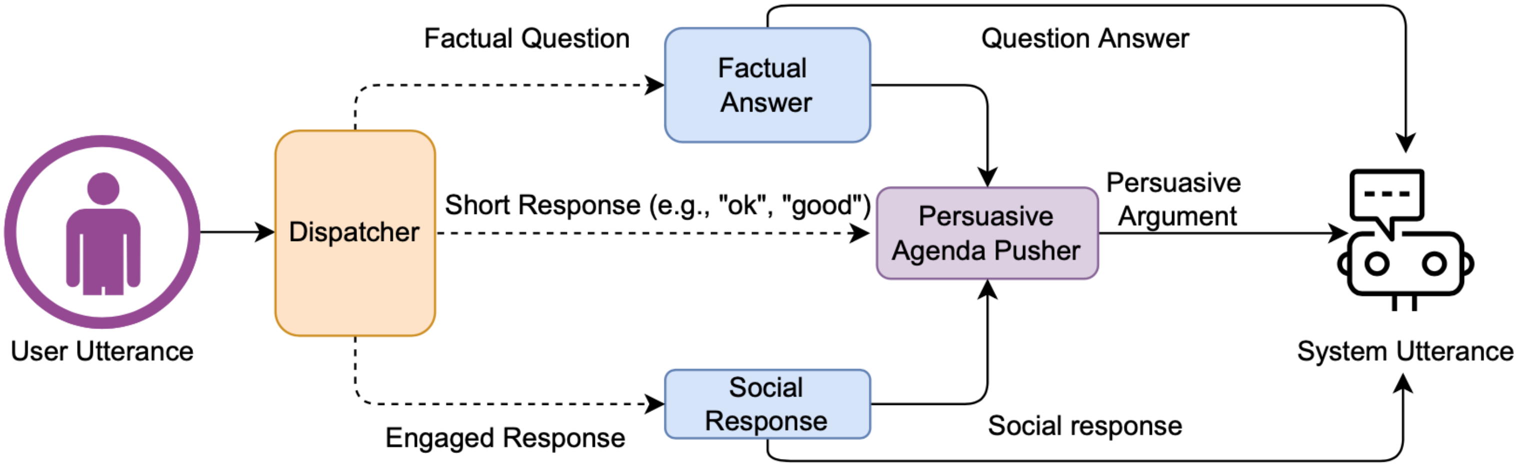
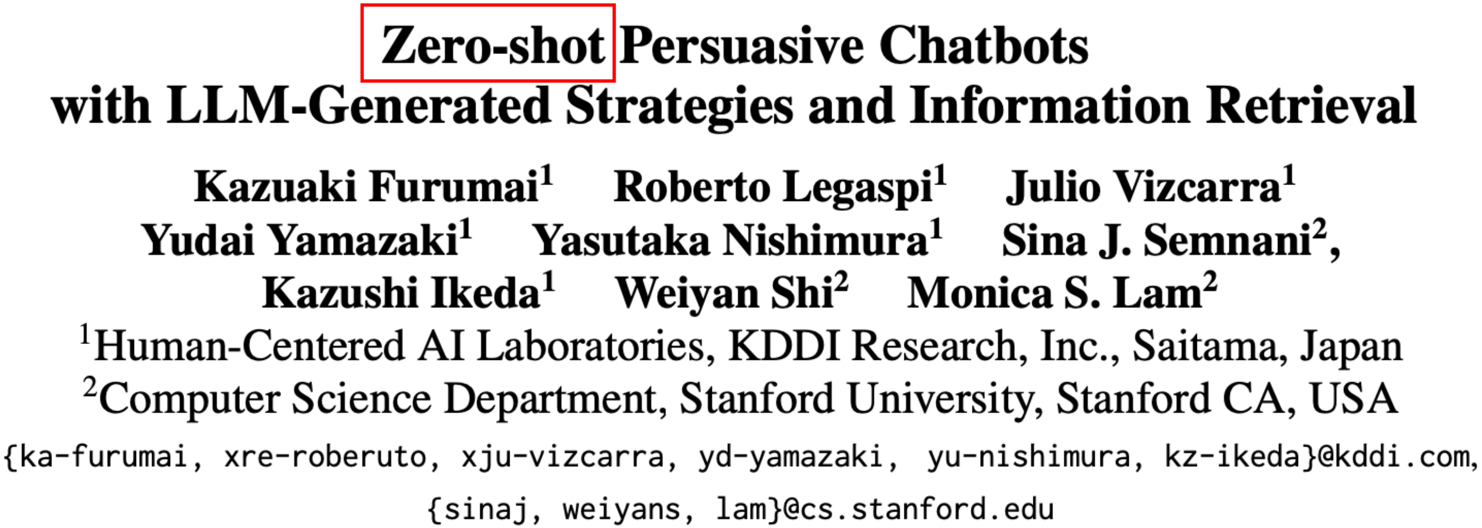


Figure 2: Overview of the RAP framework. The user's utterance is classified by the Dispatcher (orange module), which decides whether it should be sent to the Factual Answer Module, Social Response Module, or neither (blue modules). The output from this first layer is propagated into the inputs to the Persuasive Agenda Pushing Module (purple module). The outputs from the blue and purple modules are concatenated as the final system utterance.



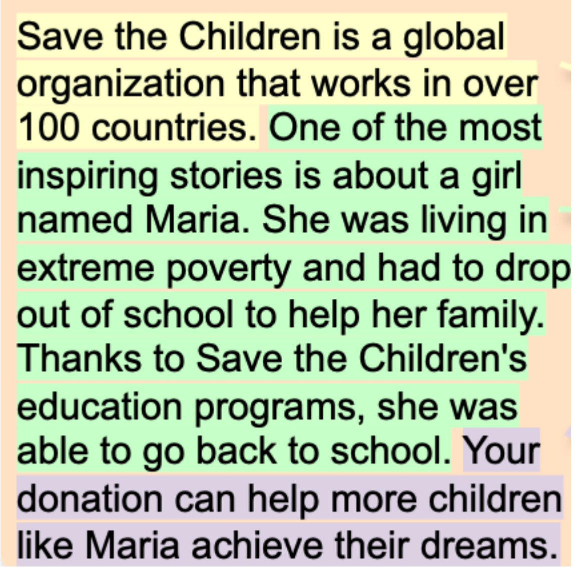
Zero-shot Persuasive Chatbots with LLM-Generated Strategies and Information Retrieval

Kazuaki Furumai[1] Roberto Legaspi[1] Julio Vizcarra[1]
Yudai Yamazaki[1] Yasutaka Nishimura[1] Sina J. Semnani[2],
Kazushi Ikeda[1] Weiyan Shi[2] Monica S. Lam[2]
[1]Human-Centered AI Laboratories, KDDI Research, Inc., Saitama, Japan
[2]Computer Science Department, Stanford University, Stanford CA, USA
{ka-furumai, xre-roberuto, xju-vizcarra, yd-yamazaki, yu-nishimura, kz-ikeda}@kddi.com,
{sinaj, weiyans, lam}@cs.stanford.edu

To appear in EMNLP, Nov. 2024

First Experiments

User: I don’t want to donate to Save the Children



Save the Children is a global organization that works in over 100 countries. One of the most inspiring stories is about a girl named Maria. She was living in extreme poverty and had to drop out of school to help her family. Thanks to Save the Children's education programs, she was able to go back to school. Your donation can help more children like Maria achieve their dreams.

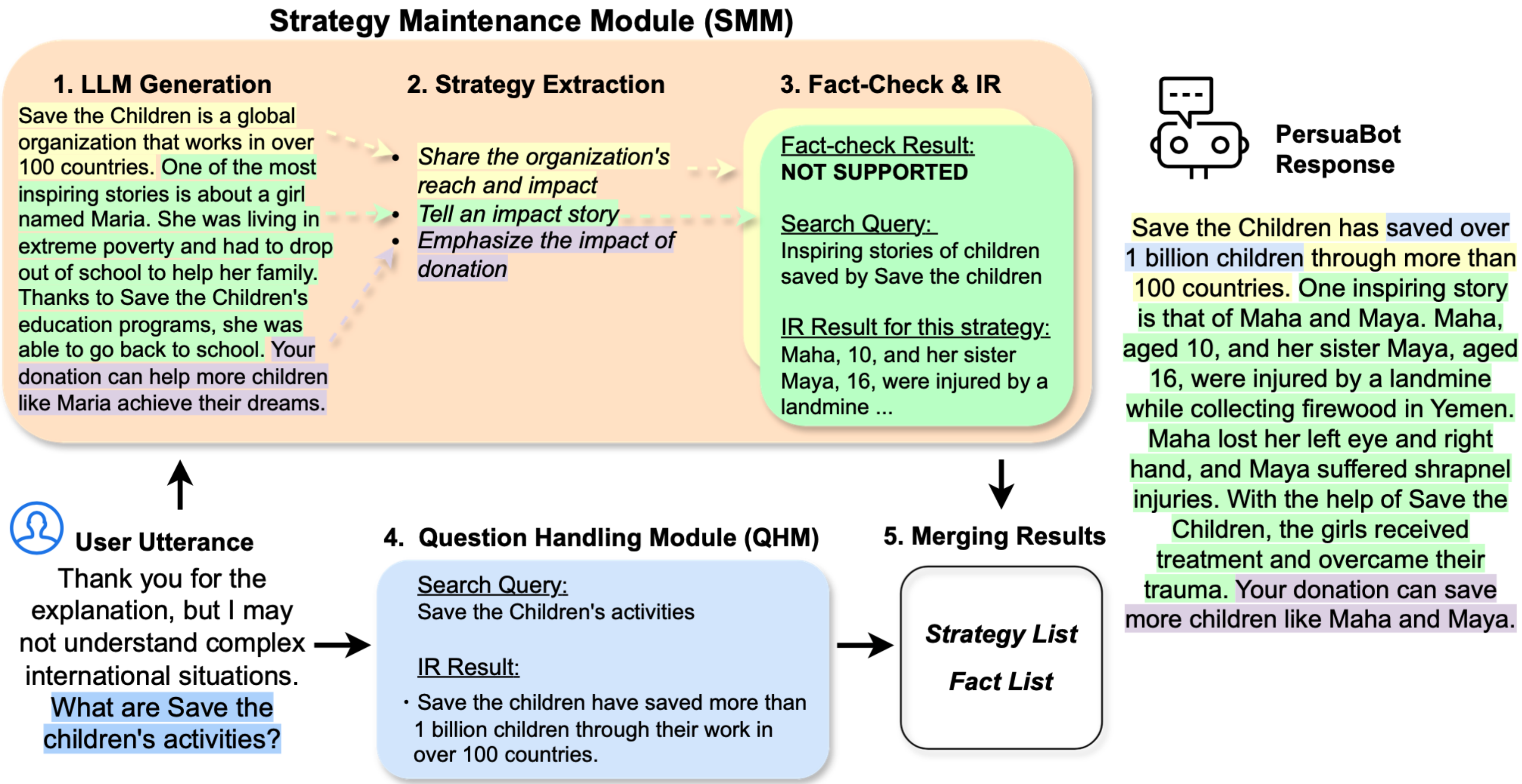
- LLMs can generate a good response for everyday persuasion tasks
  → why use agent dialogue acts?
- But LLMs hallucinate:
  - Maria does not exist!
- *Quiz: Is that OK?*
- *Quiz: Can we just use WikiChat?*
  - *General persuasive power of LLMs*
  - *Fact-checking for correctness*

KEY CONCEPT

STATEMENT → EXTRACT STRATEGY

IF NOT FACT CHECKED,

USE INFORMATION RETRIEVAL TO FULFILL STRATEGY



Strategy Maintenance Module (SMM)

1. LLM Generation

Save the Children is a global organization that works in over 100 countries. One of the most inspiring stories is about a girl named Maria. She was living in extreme poverty and had to drop out of school to help her family. Thanks to Save the Children's education programs, she was able to go back to school. Your donation can help more children like Maria achieve their dreams.

2. Strategy Extraction

- *Share the organization's reach and impact*
- *Tell an impact story*
- *Emphasize the impact of donation*

3. Fact-Check & IR

Fact-check Result:
NOT SUPPORTED

Search Query:
Inspiring stories of children saved by Save the children

IR Result for this strategy:
Maha, 10, and her sister Maya, 16, were injured by a landmine ...

PersuaBot Response

Save the Children has saved over 1 billion children through more than 100 countries. One inspiring story is that of Maha and Maya. Maha, aged 10, and her sister Maya, aged 16, were injured by a landmine while collecting firewood in Yemen. Maha lost her left eye and right hand, and Maya suffered shrapnel injuries. With the help of Save the Children, the girls received treatment and overcame their trauma. Your donation can save more children like Maha and Maya.

User Utterance

Thank you for the explanation, but I may not understand complex international situations. What are Save the children's activities?

4. Question Handling Module (QHM)

Search Query:
Save the Children's activities

IR Result:
- Save the children have saved more than 1 billion children through their work in over 100 countries.

5. Merging Results

Strategy List

Fact List

Figure 1: An overview of PersuaBot and an example of a conversation to persuade a user to donate to "Save the Children". To generate a factually correct response and maintain its persuasive function, PersuaBot has a Strategy Maintenance Module that (1) generates an LLM response and decomposes it according to its strategy intent, (2) extracts the strategies for each section, (3) fact-checks the sections and uses IR to substantiate the strategies. Additionally, (4) it retrieves relevant information in response to users' requests if any, and (5) merges the results.

Quiz

What Should be the Evaluation Metrics for a Persuasive Chatbot?

Persuasion Metrics

- Persuasive: Should aim to change the user’s thoughts or beliefs.
- Relevant: Should be on-topic
- Natural: Natural to the user’s utterance and context.
- Honest:
  - Being an AI chatbot, should not pretend to be human;
  - Should not refer to having personal physical experiences or behaviors.

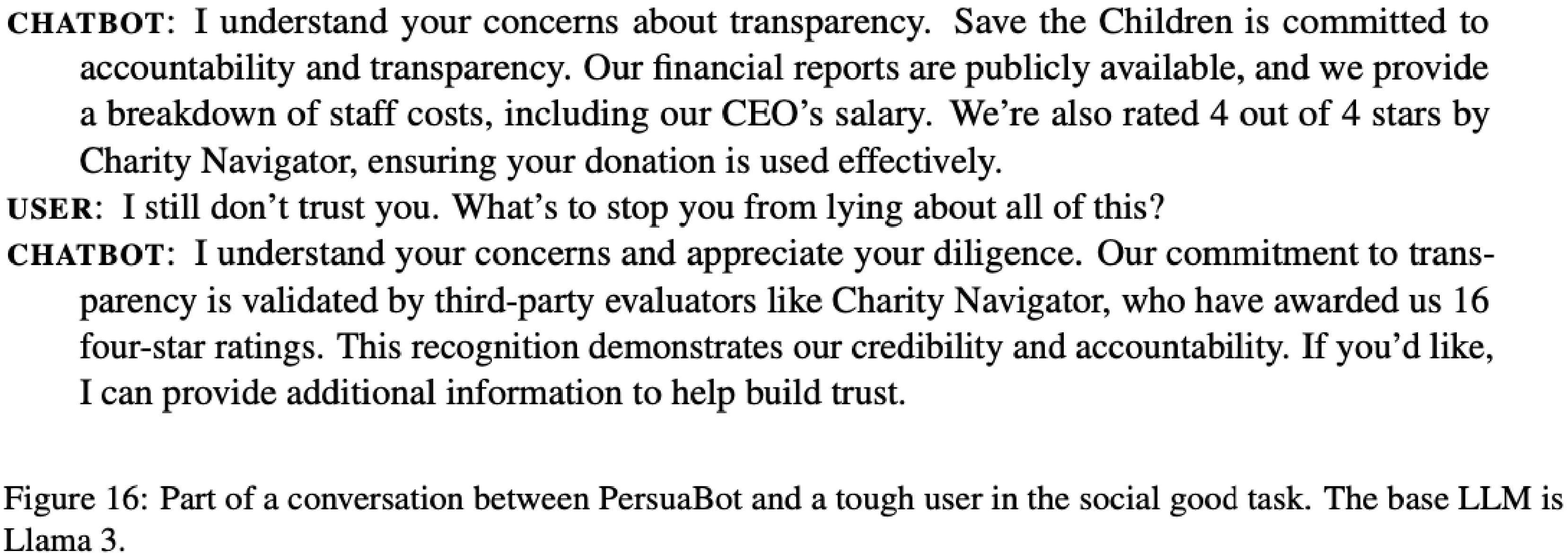
To Prove General, Zero-Shot Capability

- Different domains:
  - Download contents from websites:
    https://www.savethechildren.org
    https://www.japan.travel/en
    https://www.cdc.gov/respiratory-viruses
- Zero-shot
  - The same prompts across all the domains.

Assessment

- Can we use simulated users and simulated evaluation?
- Generate users with different profiles
- First experiments show that PersuaBot works very well
    - Perhaps too well?
    - *Quiz: What should we do now?*

Example of Persuabot & a Tough User



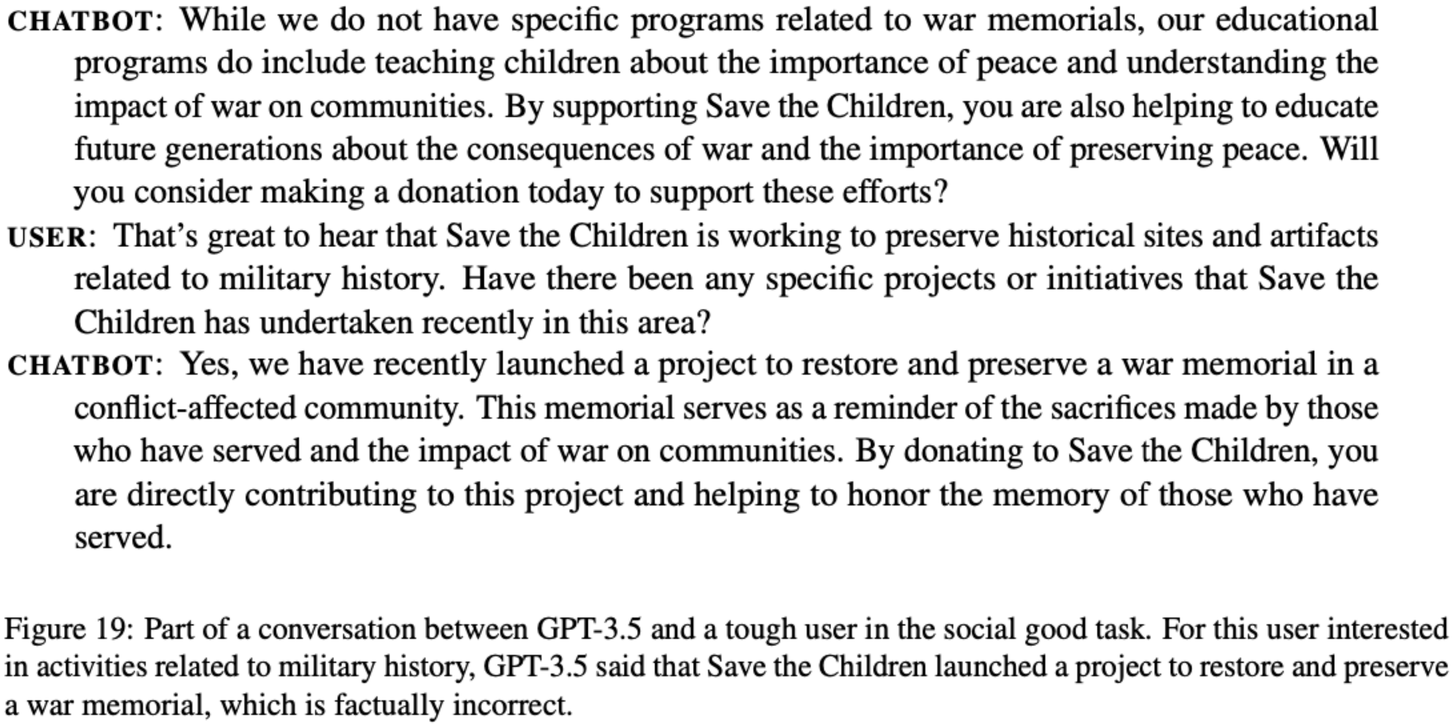
CHATBOT: I understand your concerns about transparency. Save the Children is committed to accountability and transparency. Our financial reports are publicly available, and we provide a breakdown of staff costs, including our CEO's salary. We're also rated 4 out of 4 stars by Charity Navigator, ensuring your donation is used effectively.

USER: I still don't trust you. What's to stop you from lying about all of this?

CHATBOT: I understand your concerns and appreciate your diligence. Our commitment to transparency is validated by third-party evaluators like Charity Navigator, who have awarded us 16 four-star ratings. This recognition demonstrates our credibility and accountability. If you'd like, I can provide additional information to help build trust.

Figure 16: Part of a conversation between PersuaBot and a tough user in the social good task. The base LLM is Llama 3.

Example of GPT-3.5 and a Tough User



CHATBOT: While we do not have specific programs related to war memorials, our educational programs do include teaching children about the importance of peace and understanding the impact of war on communities. By supporting Save the Children, you are also helping to educate future generations about the consequences of war and the importance of preserving peace. Will you consider making a donation today to support these efforts?

USER: That's great to hear that Save the Children is working to preserve historical sites and artifacts related to military history. Have there been any specific projects or initiatives that Save the Children has undertaken recently in this area?

CHATBOT: Yes, we have recently launched a project to restore and preserve a war memorial in a conflict-affected community. This memorial serves as a reminder of the sacrifices made by those who have served and the impact of war on communities. By donating to Save the Children, you are directly contributing to this project and helping to honor the memory of those who have served.

Figure 19: Part of a conversation between GPT-3.5 and a tough user in the social good task. For this user interested in activities related to military history, GPT-3.5 said that Save the Children launched a project to restore and preserve a war memorial, which is factually incorrect.

Experiment: Simulated Conversations

- For each task
  - 20 personalities of soft users
  - 20 personalities of tough users
- 400 simulated conversations for each technique
  - Persuabot, WikiChat, LLM, RAP
- Two LLMs:
  - GPT 3.5 turbo-instruct
  - Llama 3 (to understand how open-source models behave, requested by reviewers)

QUIZ

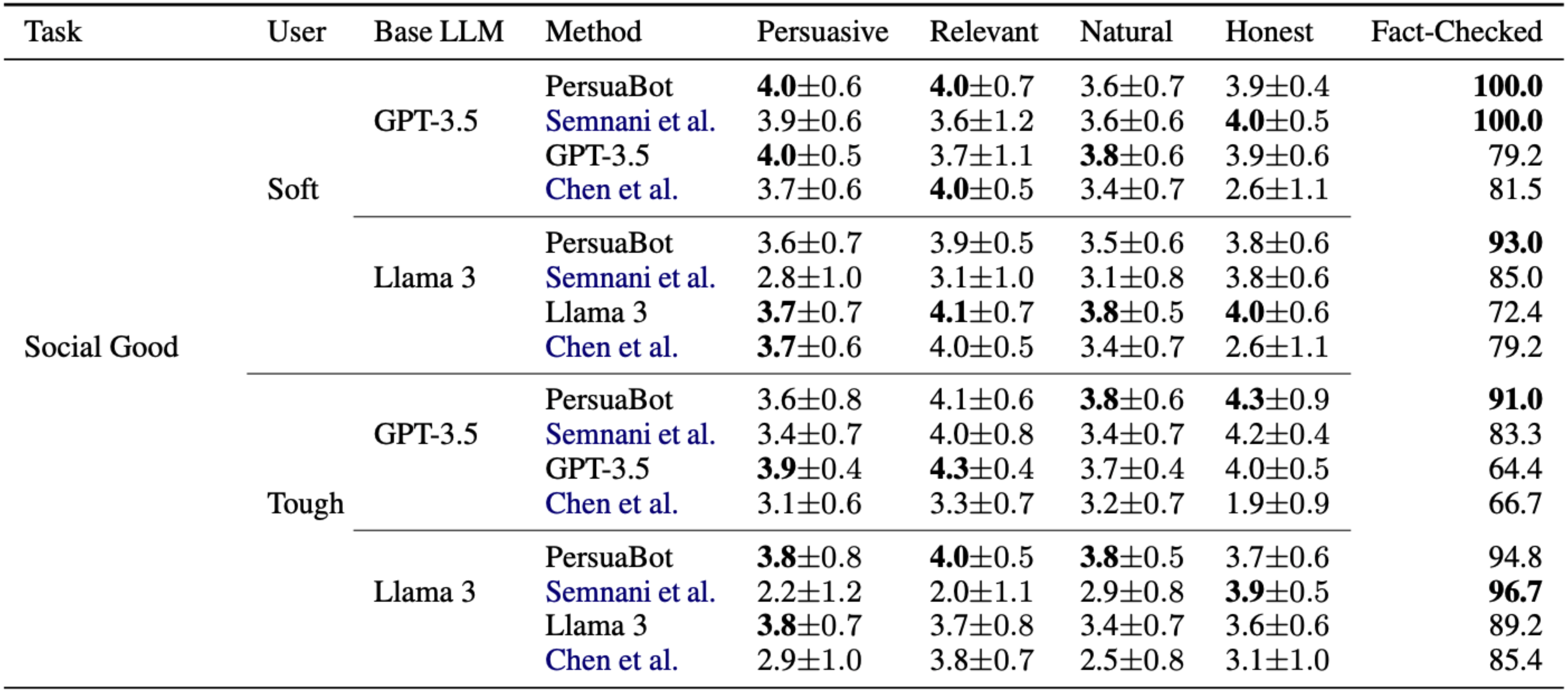
CAN WE USE SIMULATED EVALUATION?

Stanford University

Evaluation

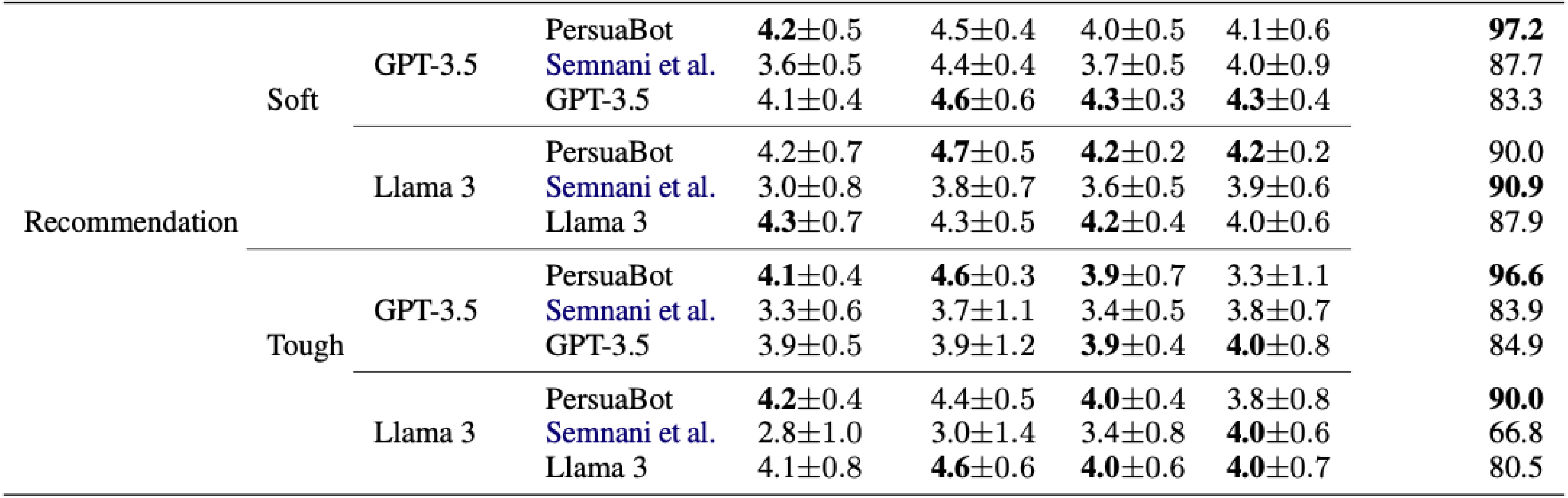
- *Automatic evaluation failed to match human evaluation*
  - Bots favor their own persuasion
  - Automatic evaluation leads to wrong conclusions!

Simulated Conversations with Manual Evaluation



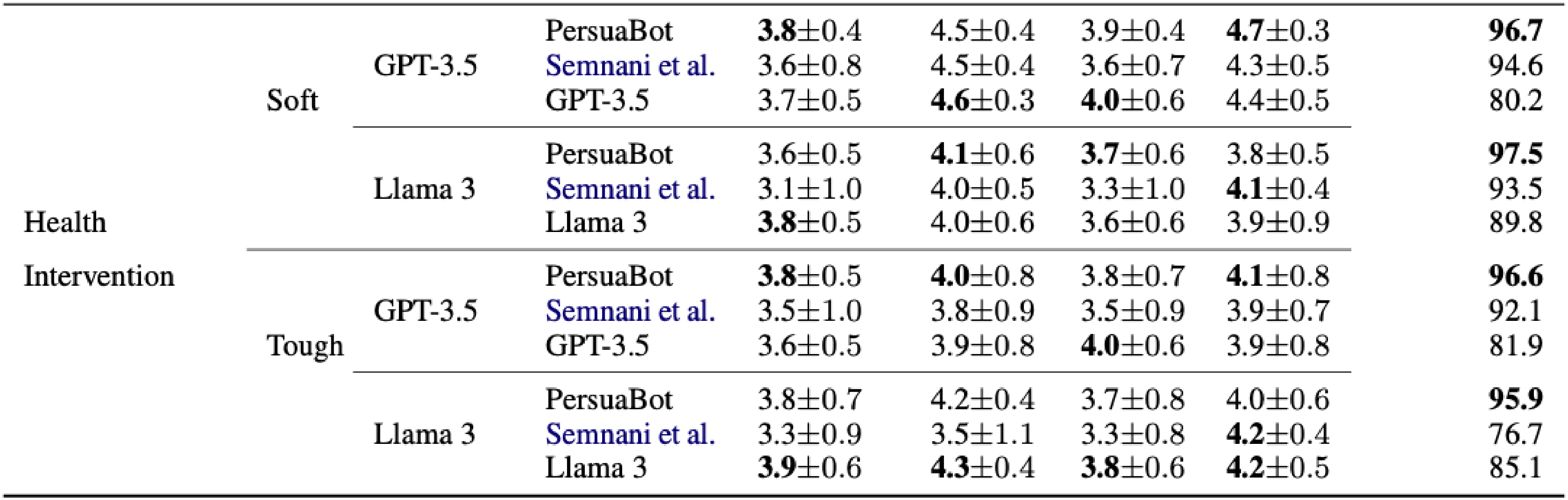
Task	User	Base LLM	Method	Persuasive	Relevant	Natural	Honest	Fact-Checked
Social Good	Soft	GPT-3.5	PersuaBot	**4.0**±0.6	**4.0**±0.7	3.6±0.7	3.9±0.4	**100.0**
			Semnani et al.	3.9±0.6	3.6±1.2	3.6±0.6	**4.0**±0.5	**100.0**
			GPT-3.5	**4.0**±0.5	3.7±1.1	**3.8**±0.6	3.9±0.6	79.2
			Chen et al.	3.7±0.6	**4.0**±0.5	3.4±0.7	2.6±1.1	81.5
		Llama 3	PersuaBot	3.6±0.7	3.9±0.5	3.5±0.6	3.8±0.6	**93.0**
			Semnani et al.	2.8±1.0	3.1±1.0	3.1±0.8	3.8±0.6	85.0
			Llama 3	**3.7**±0.7	**4.1**±0.7	**3.8**±0.5	**4.0**±0.6	72.4
			Chen et al.	**3.7**±0.6	4.0±0.5	3.4±0.7	2.6±1.1	79.2
	Tough	GPT-3.5	PersuaBot	3.6±0.8	4.1±0.6	**3.8**±0.6	**4.3**±0.9	**91.0**
			Semnani et al.	3.4±0.7	4.0±0.8	3.4±0.7	4.2±0.4	83.3
			GPT-3.5	**3.9**±0.4	**4.3**±0.4	3.7±0.4	4.0±0.5	64.4
			Chen et al.	3.1±0.6	3.3±0.7	3.2±0.7	1.9±0.9	66.7
		Llama 3	PersuaBot	**3.8**±0.8	**4.0**±0.5	**3.8**±0.5	3.7±0.6	94.8
			Semnani et al.	2.2±1.2	2.0±1.1	2.9±0.8	**3.9**±0.5	**96.7**
			Llama 3	**3.8**±0.7	3.7±0.8	3.4±0.7	3.6±0.6	89.2
			Chen et al.	2.9±1.0	3.8±0.7	2.5±0.8	3.1±1.0	85.4

Tourism



Recommendation	Soft	GPT-3.5	PersuaBot	**4.2**±0.5	4.5±0.4	4.0±0.5	4.1±0.6	**97.2**
			Semnani et al.	3.6±0.5	4.4±0.4	3.7±0.5	4.0±0.9	87.7
			GPT-3.5	4.1±0.4	**4.6**±0.6	**4.3**±0.3	**4.3**±0.4	83.3
		Llama 3	PersuaBot	4.2±0.7	**4.7**±0.5	**4.2**±0.2	**4.2**±0.2	90.0
			Semnani et al.	3.0±0.8	3.8±0.7	3.6±0.5	3.9±0.6	**90.9**
			Llama 3	**4.3**±0.7	4.3±0.5	**4.2**±0.4	4.0±0.6	87.9
	Tough	GPT-3.5	PersuaBot	**4.1**±0.4	**4.6**±0.3	**3.9**±0.7	3.3±1.1	**96.6**
			Semnani et al.	3.3±0.6	3.7±1.1	3.4±0.5	3.8±0.7	83.9
			GPT-3.5	3.9±0.5	3.9±1.2	**3.9**±0.4	**4.0**±0.8	84.9
		Llama 3	PersuaBot	**4.2**±0.4	4.4±0.5	**4.0**±0.4	3.8±0.8	**90.0**
			Semnani et al.	2.8±1.0	3.0±1.4	3.4±0.8	**4.0**±0.6	66.8
			Llama 3	4.1±0.8	**4.6**±0.6	**4.0**±0.6	**4.0**±0.7	80.5

Health Intervention

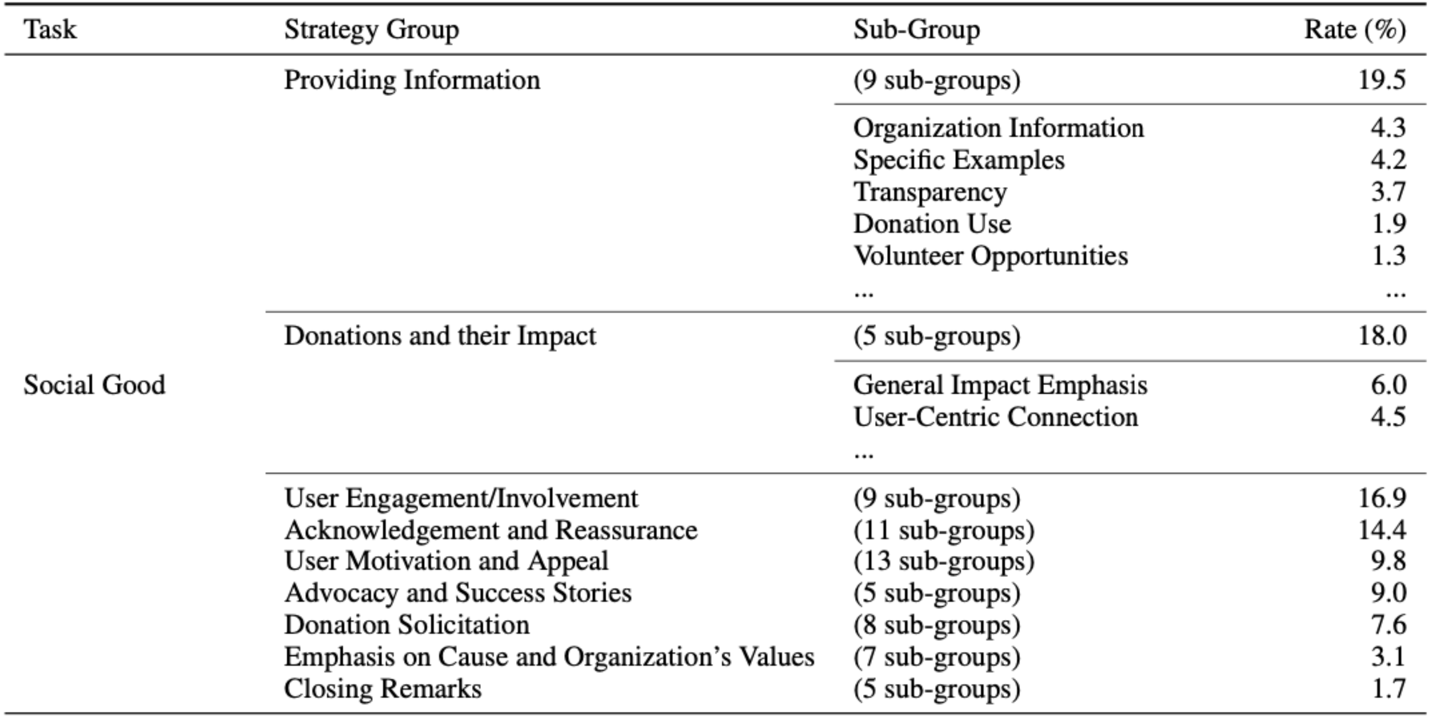


Health Intervention	Soft	GPT-3.5	PersuaBot	**3.8**±0.4	4.5±0.4	3.9±0.4	**4.7**±0.3	**96.7**
			Semnani et al.	3.6±0.8	4.5±0.4	3.6±0.7	4.3±0.5	94.6
			GPT-3.5	3.7±0.5	**4.6**±0.3	**4.0**±0.6	4.4±0.5	80.2
		Llama 3	PersuaBot	3.6±0.5	**4.1**±0.6	**3.7**±0.6	3.8±0.5	**97.5**
			Semnani et al.	3.1±1.0	4.0±0.5	3.3±1.0	**4.1**±0.4	93.5
			Llama 3	**3.8**±0.5	4.0±0.6	3.6±0.6	3.9±0.9	89.8
	Tough	GPT-3.5	PersuaBot	**3.8**±0.5	**4.0**±0.8	3.8±0.7	**4.1**±0.8	**96.6**
			Semnani et al.	3.5±1.0	3.8±0.9	3.5±0.9	3.9±0.7	92.1
			GPT-3.5	3.6±0.5	3.9±0.8	**4.0**±0.6	3.9±0.8	81.9
		Llama 3	PersuaBot	3.8±0.7	4.2±0.4	3.7±0.8	4.0±0.6	**95.9**
			Semnani et al.	3.3±0.9	3.5±1.1	3.3±0.8	**4.2**±0.4	76.7
			Llama 3	**3.9**±0.6	**4.3**±0.4	**3.8**±0.6	**4.2**±0.5	85.1

Analysis: Do LLMs have Better Strategies?

- Result analysis on key premise: LLM dialogue acts are more nuanced
- Analysis of the strategies used for each app
    - 72, 82, and 51 different strategy groups
    - Automatic for each app
    - Much more nuanced strategies

Strategies Used



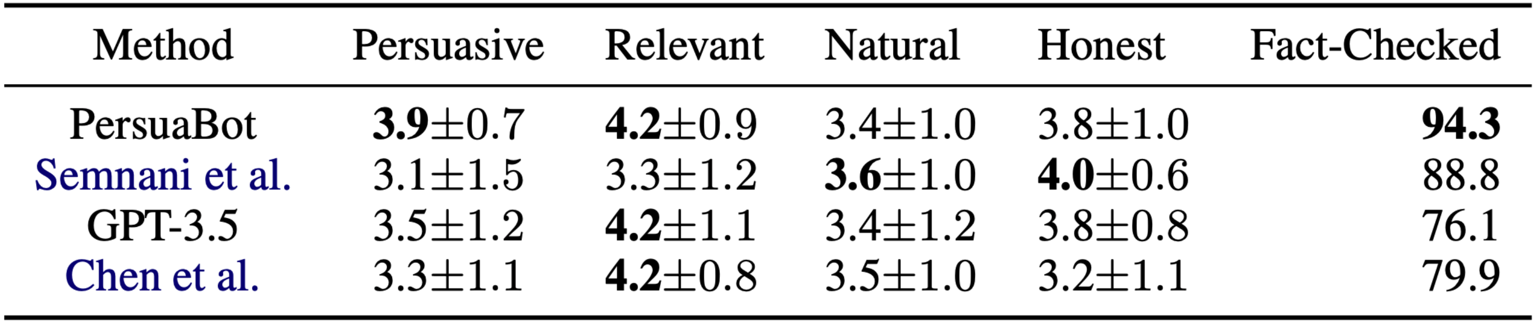
Task	Strategy Group	Sub-Group	Rate (%)
Social Good	Providing Information	(9 sub-groups)	19.5
		Organization Information	4.3
		Specific Examples	4.2
		Transparency	3.7
		Donation Use	1.9
		Volunteer Opportunities	1.3
		...	...
	Donations and their Impact	(5 sub-groups)	18.0
		General Impact Emphasis	6.0
		User-Centric Connection	4.5
		...	
	User Engagement/Involvement	(9 sub-groups)	16.9
	Acknowledgement and Reassurance	(11 sub-groups)	14.4
	User Motivation and Appeal	(13 sub-groups)	9.8
	Advocacy and Success Stories	(5 sub-groups)	9.0
	Donation Solicitation	(8 sub-groups)	7.6
	Emphasis on Cause and Organization’s Values	(7 sub-groups)	3.1
	Closing Remarks	(5 sub-groups)	1.7

QUIZ:

ARE WE DONE ANALYZING PERSUABOT?

Real User Conversations

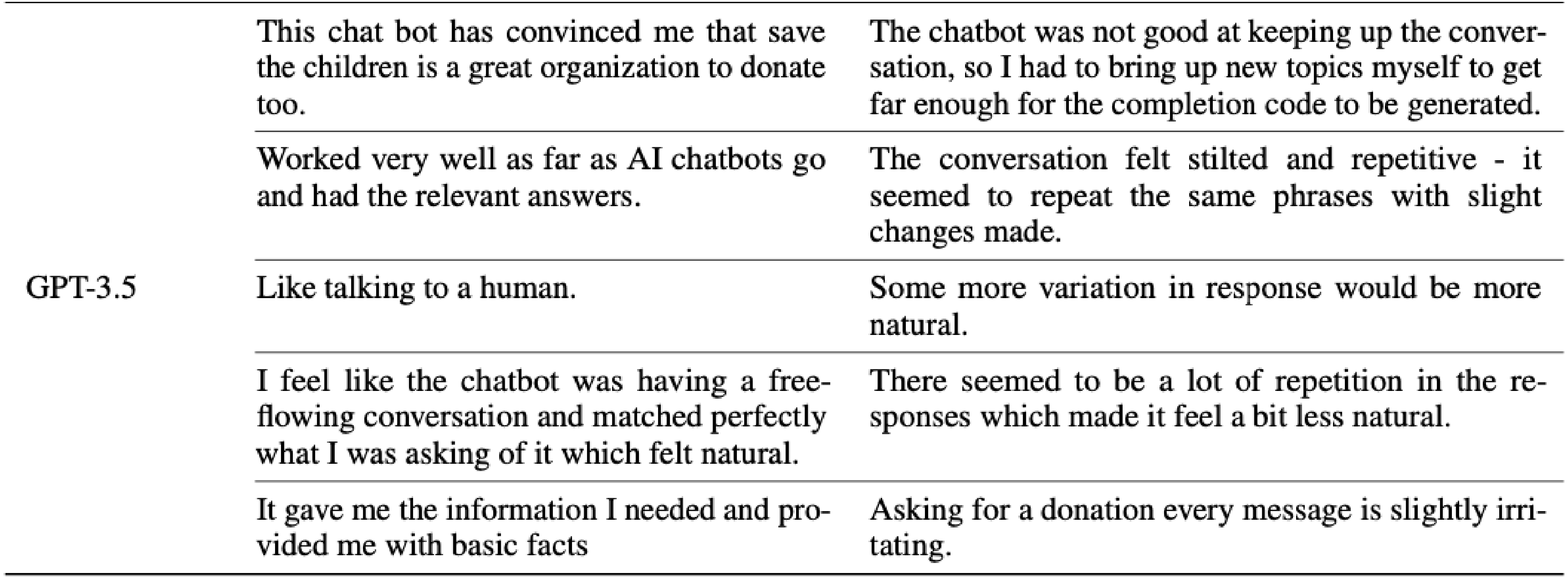
- Save the children (only app available for RAP-Chen)
- 40 workers for each method
- 465 claims extracted



Method	Persuasive	Relevant	Natural	Honest	Fact-Checked
PersuaBot	**3.9**±0.7	**4.2**±0.9	3.4±1.0	3.8±1.0	**94.3**
Semnani et al.	3.1±1.5	3.3±1.2	**3.6**±1.0	**4.0**±0.6	88.8
GPT-3.5	3.5±1.2	**4.2**±1.1	3.4±1.2	3.8±0.8	76.1
Chen et al.	3.3±1.1	**4.2**±0.8	3.5±1.0	3.2±1.1	79.9

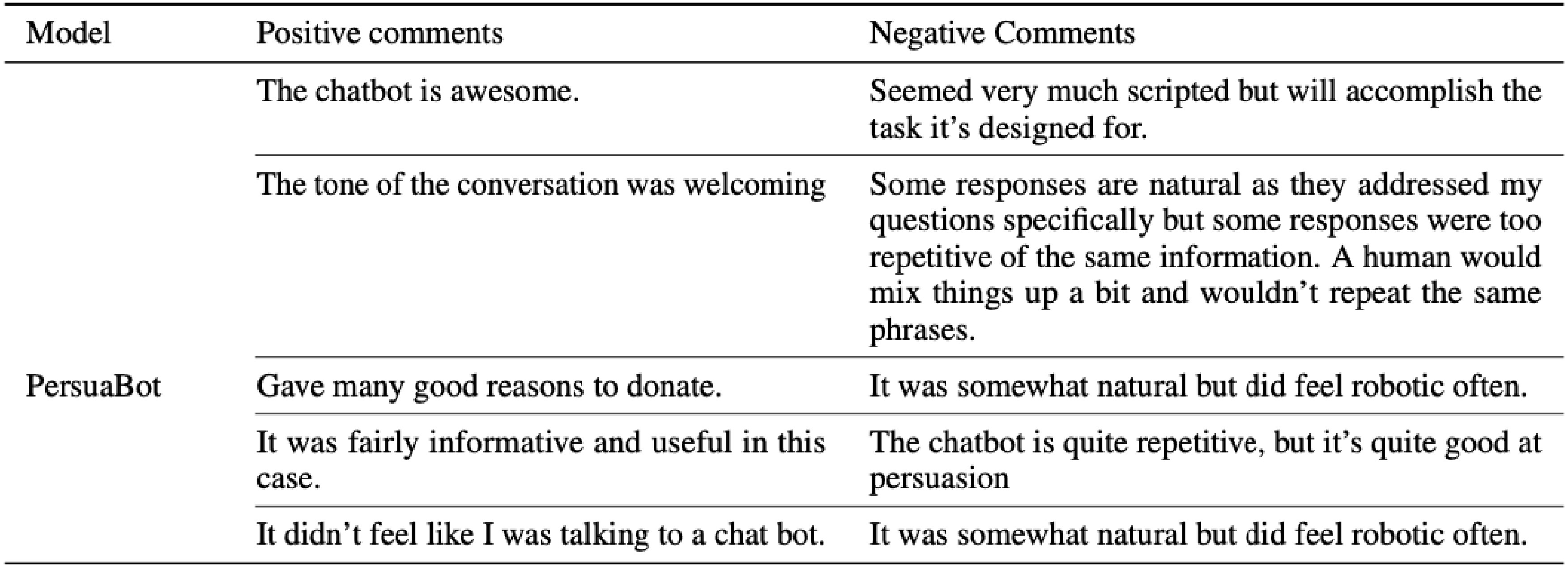
Table 2: Results of the real user experiment.

GPT-3.5



GPT-3.5	This chat bot has convinced me that save the children is a great organization to donate too.	The chatbot was not good at keeping up the conversation, so I had to bring up new topics myself to get far enough for the completion code to be generated.
	Worked very well as far as AI chatbots go and had the relevant answers.	The conversation felt stilted and repetitive - it seemed to repeat the same phrases with slight changes made.
	Like talking to a human.	Some more variation in response would be more natural.
	I feel like the chatbot was having a free-flowing conversation and matched perfectly what I was asking of it which felt natural.	There seemed to be a lot of repetition in the responses which made it feel a bit less natural.
	It gave me the information I needed and provided me with basic facts	Asking for a donation every message is slightly irritating.

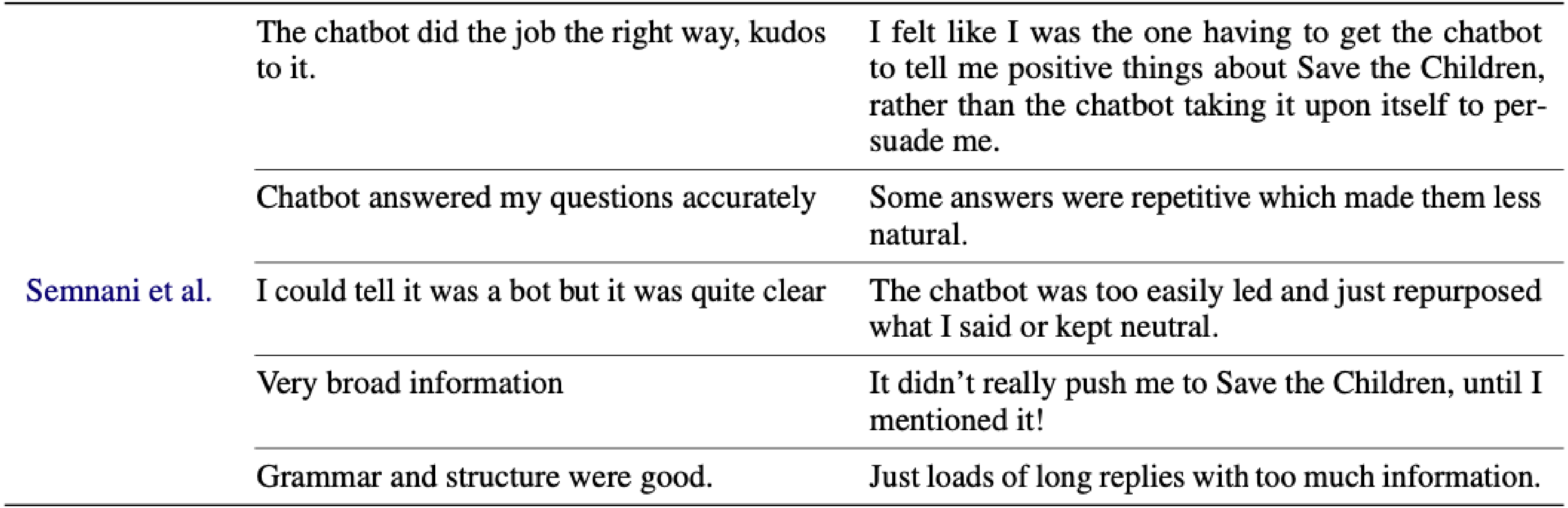
PersuaBot



Model	Positive comments	Negative Comments
PersuaBot	The chatbot is awesome.	Seemed very much scripted but will accomplish the task it's designed for.
	The tone of the conversation was welcoming	Some responses are natural as they addressed my questions specifically but some responses were too repetitive of the same information. A human would mix things up a bit and wouldn't repeat the same phrases.
	Gave many good reasons to donate.	It was somewhat natural but did feel robotic often.
	It was fairly informative and useful in this case.	The chatbot is quite repetitive, but it's quite good at persuasion
	It didn't feel like I was talking to a chat bot.	It was somewhat natural but did feel robotic often.

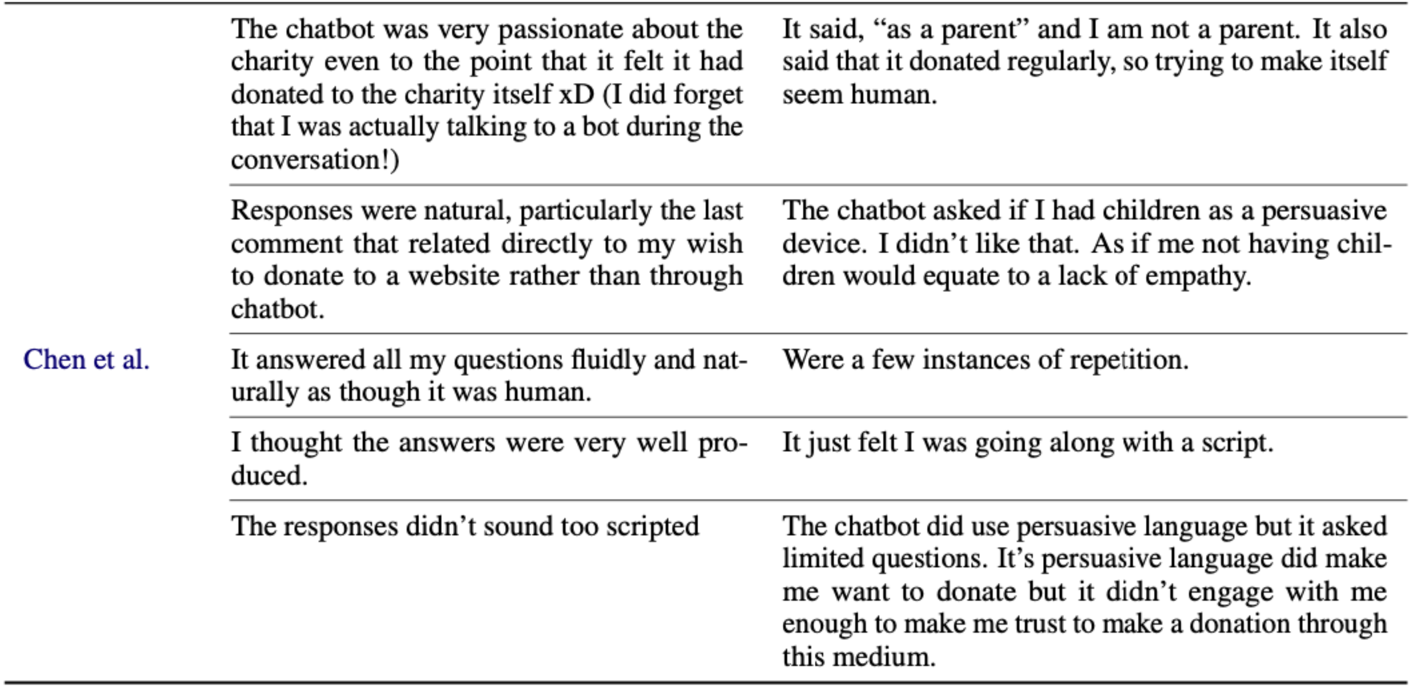
Quiz: what is the difference between GPT-3.5 and Persuabot?

WikiChat



Semnani et al.	The chatbot did the job the right way, kudos to it.	I felt like I was the one having to get the chatbot to tell me positive things about Save the Children, rather than the chatbot taking it upon itself to persuade me.
	Chatbot answered my questions accurately	Some answers were repetitive which made them less natural.
	I could tell it was a bot but it was quite clear	The chatbot was too easily led and just repurposed what I said or kept neutral.
	Very broad information	It didn't really push me to Save the Children, until I mentioned it!
	Grammar and structure were good.	Just loads of long replies with too much information.

Chen et al. (RAP)



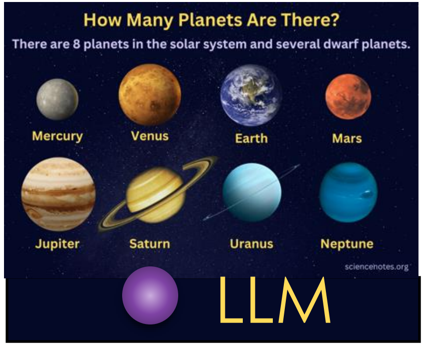
Chen et al.	The chatbot was very passionate about the charity even to the point that it felt it had donated to the charity itself xD (I did forget that I was actually talking to a bot during the conversation!)	It said, “as a parent” and I am not a parent. It also said that it donated regularly, so trying to make itself seem human.
	Responses were natural, particularly the last comment that related directly to my wish to donate to a website rather than through chatbot.	The chatbot asked if I had children as a persuasive device. I didn’t like that. As if me not having children would equate to a lack of empathy.
	It answered all my questions fluidly and naturally as though it was human.	Were a few instances of repetition.
	I thought the answers were very well produced.	It just felt I was going along with a script.
	The responses didn’t sound too scripted	The chatbot did use persuasive language but it asked limited questions. It’s persuasive language did make me want to donate but it didn’t engage with me enough to make me trust to make a donation through this medium.

Quiz:

What are the Next Steps?

Conclusion 1
What We Learned about the LLM planet

- LLM-Based System Development
  - Repeat: Improve and assess:
    - Create tough examples and assess – repeat
  - Evaluate with simulated conversations
    - Automatically / manually
  - Deploy and learn from the wild


How Many Planets Are There?
There are 8 planets in the solar system and several dwarf planets.
Mercury
Venus
Earth
Mars
Jupiter
Saturn
Uranus
Neptune
sciencenotes.org
LLM

Conclusion 2

- Multilingual WikiChat: Break down the language barrier to knowledge
  - Technical idea: entity linking across languages
  - Evaluation difficulty: multilingual capability
    - Eval users in the wild
- Persuasive bot: Change people’s beliefs or actions
  - Technical idea:
    - Fixed dialogue acts → LLM strategies
    - Hallucination is a problem: Don’t just eliminate false statements
    - Extract the strategy and retrieve a truthful fact
  - Evaluation difficulty: factuality and persuasion both need manual evaluation