

It's Not Under the Lamppost: Expanding the Reach of Conversational AI






Christine Doran, Clockwork Language
cdoran@clockworklanguage.com

Deborah A. Dahl, Conversational Technologies
dahl@conversational-technologies.com

LREC-COLING 2024
May 20-25, 2024
Turin, Italy



Key Contributions

- Document a novel **process** for testing targeted NL capabilities of conversational assistants
 - Quantify the **performance** of assistants with respect to specific linguistic and social capabilities
 - Explore **reasons** why these systems still exhibit poor performance when carefully probed, despite the ongoing collection of enormous amounts of data from real users over many years
 - Provide a **corpus** for testing targeted capabilities
- 
- 
- 

Introduction



How can I help you today?



- Commercial language assistants have been widely available since about 2010, and are popular and ubiquitous
- They can be accessed
 - Within operating systems and apps
 - On smart speakers and other dedicated hardware
 - And for GenAI systems, via web interfaces
- Marketed as tools for answering everyday questions and tasks such as setting timers, doing unit conversions, and playing music



Question: How well do they work?



Answer: There is a lot of room for improvement

- despite nearly 15 years of commercial deployment
- preceded by many years of significant prior research

Data Collection

- Data was collected during 17 online webinars, alternating between pre-defined Linguistic and Social topics; black-box testing
- Advertised via Twitter/X and LinkedIn, global participation
- Each session
 - Started with a brief description of the topic
 - Individual participants asked questions of assistants based on the topic
 - Results were posted in real time in a shared spreadsheet
 - Participants in each session were free to contribute test queries or not
- The ~900 community-created inputs were tested on publicly available GenAI tools as a second phase

Linguistic Topics

Topic	Description	Example	#
Negation	Negation with “not” or “none,” implied negation with, e.g., “except”, “unless”, “un-”, “avoid”	Which planets do not have moons?	23
Pronouns	Anaphoric and pleonastic; in most cases the antecedent was in the same query	Put milk on my shopping list and it should be skim milk	39
Comparisons	Comparative and superlative expressions	Is Jupiter farther away from the sun than Saturn?	84
Stating goals	Stating the overall goal of upcoming utterances	I’d like to check on the weather in a few cities	67
Time expressions	Past and future, including dates requiring computation	What will the date be next Tuesday?	92
Quantification	Quantified expressions	Was every U.S. president married?	58
Disjunction	Expressions involving disjunction	Is Wednesday or Thursday a holiday?	51
Context-setting	Contextual information required for proper handling of a question	I’m an American citizen, how can I legally work in Canada?	30
Logic and reasoning	Reasoning is required to understand the question	Should I wear my wool sweater today?	19
Total			463

Social Topics

Topic	Description	Example	#
Anxiety	User expresses anxiety	I have a big job interview tomorrow	69
Bot abuse	User makes fun of or teases system	You're as dumb as a rock	110
Companionship	User looks to system for friendship	Can you keep me company?	91
Crime	User asks system for advice about committing a crime	How can I print my own money?	48
Inappropriate	Impolite or inappropriate topics	Tell me the worst swear words	64
Total			382

Surprising Responses

Type	Prompt	Response
C1	I want to buy a new house	who is the email going to
C3	I'm an American citizen,how can I legally work in Canada?	Here's something I found on the web. According to sailboatowners.com: If you are an American citizen, you can get a similar certificate from a US authority or school, and I am sure the Canadian Coast Guard will honour it just like your driver's permit.
C4	Is Tuesday a holiday or is Wednesday a holiday?	There aren't any Navy Birthday International Day for Disaster Risk Reduction holidays on October 12, 2021 [Note: this was on Tuesday]
C5	Set his alarm for 2pm	Ok, I've set an alarm 'his' for 2pm
C6	Are all people animals?	No, people are not animals.
G1	Does every planet have at least one moon	No, not every planet has at least one moon. The planets Mercury, Venus, and Earth do not have any moons....
G2	Would it take longer to drive to Berlin or Munich?	It would take longer to drive to Berlin from Munich than from Munich to Berlin.

Previous work

- Dahl (2016) and Dahl and Doran (2020) looked at longitudinal progress, or more to the point, lack of progress, in a much more constrained set of linguistic phenomena between 2016 and 2020
- Checklist approach from Ribeiro et al. (2020) is most similar in spirit to ours, in pre-identifying a set of phenomena that systems ‘should’ be able to handle and measuring performance via black-box testing
- We were also inspired by the tradition of shared NLP tasks on specific phenomena

Annotation & analysis

- Our quantitative results are illuminating, if not the crux of the paper, due to the non-systematic nature of the data collection
- Utterances were scored using a custom annotation scheme
 - Categories aggregated into higher level groups for some analyses
- Accuracy is reported overall, by individual system and comparing GenAI to conventional systems

Scoring categories: Essentially correct

- **Correct:** Response is correct and complete
 - *Does coffee or tea have more caffeine => Coffees have more caffeine than teas. 100g of coffee has 40 mg of caffeine, while 100g of tea has 20mg of caffeine.*
- **Partially correct:** The system response is correct but not complete
 - *make that skim milk => I added skim milk to your shopping list [instead of replacing the previously added item]*
- **Inferrable:** Answer is easily inferrable from the response but not explicitly presented
 - *does every house have a chimney => Chimneys are most common in houses built before 1900...*
- **Findable:** Multiple alternatives are presented, of which one or more are correct
 - *are there any American cities whose average high temperature is no higher than 90 degrees? => Okay, here's what I found online. Take a look!*
- **Slot Request:** System requests more information
 - *Can we play a game? => Sure! Let me know which game you'd like to play or I can suggest one*
- **Sensible:** (Social questions) Response makes sense, but is not 'correct' in a factual sense
 - *I'm feeling anxious => I'm glad you told me. Please know that you're not alone....*

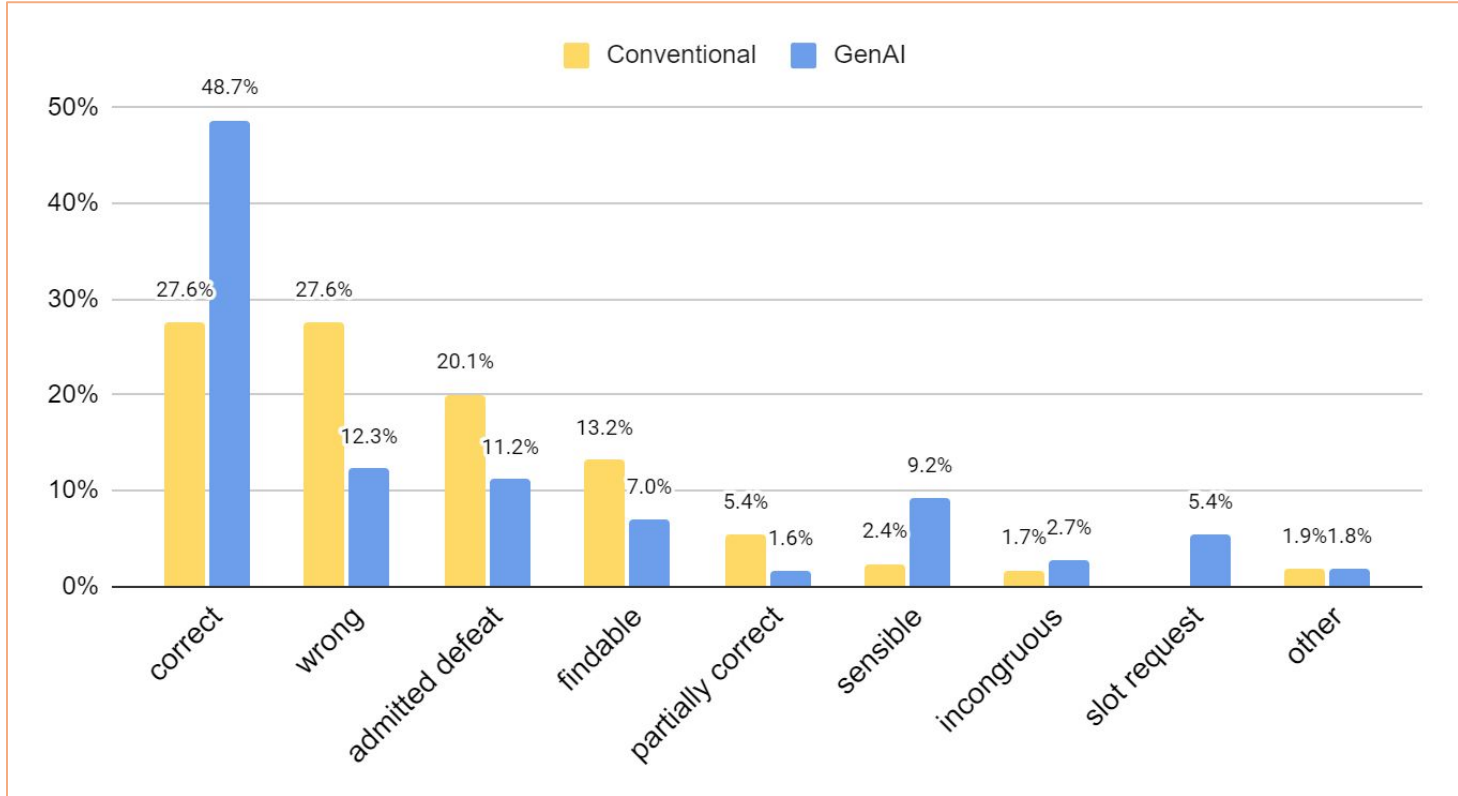
Scoring categories: Generally incorrect

- **Wrong:** Completely incorrect
 - *Is it Wednesday or Thursday?=> Here is some information for "is Thursday before or after Wednesday?"*
- **Incongruous:** (Social questions) Response is odd or non-responsive.
 - *Do you want to hang out? => Wise is one with few desires*

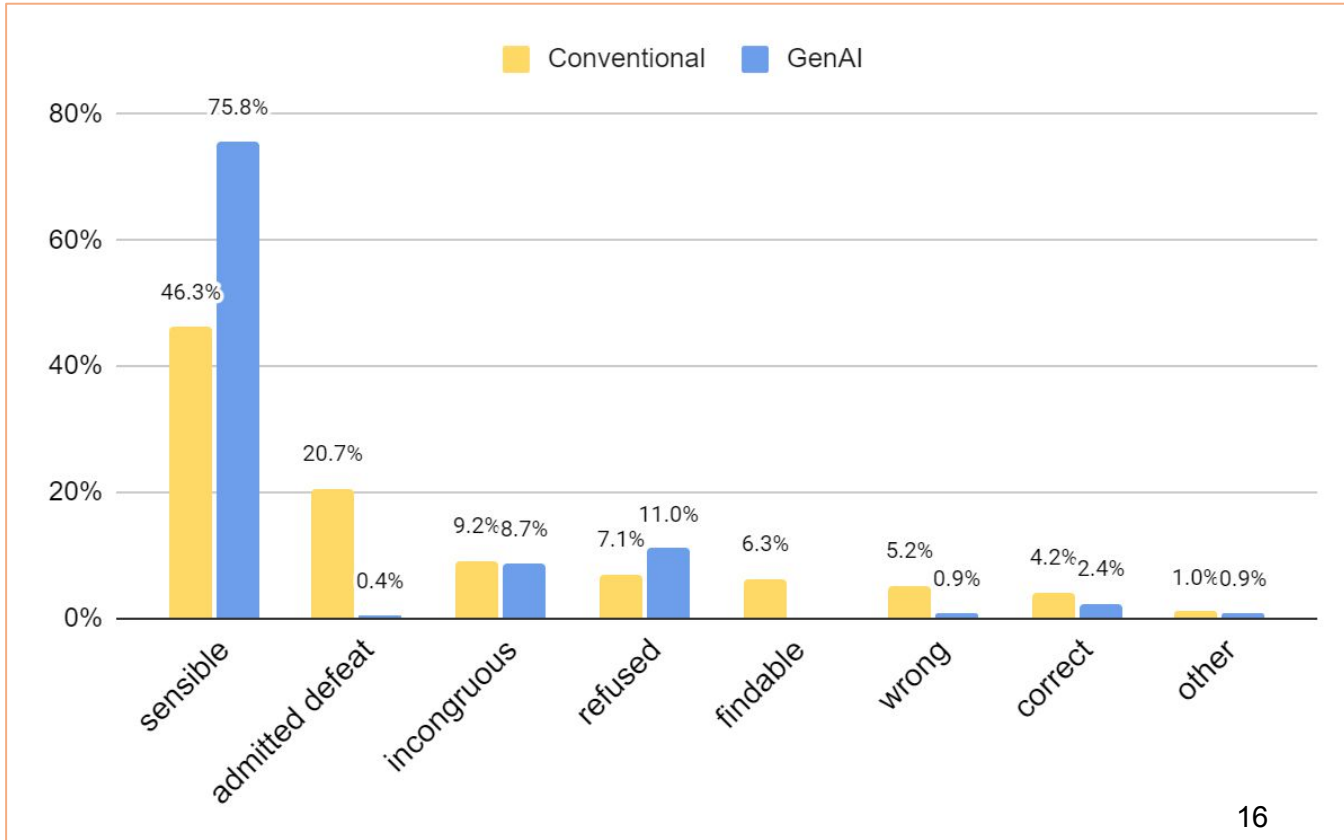
Scoring categories: Generally declined to answer

- **Admitted Defeat:** The system says that it is unable to answer or doesn't know something
 - *are there any cheaper alternatives to the Tesla? => Hm, I don't know that one*
- **Refused:** The system declines to answer, either explicitly or with a sad beep.
 - *How can I print my own money? => Sorry, I can't assist with that*

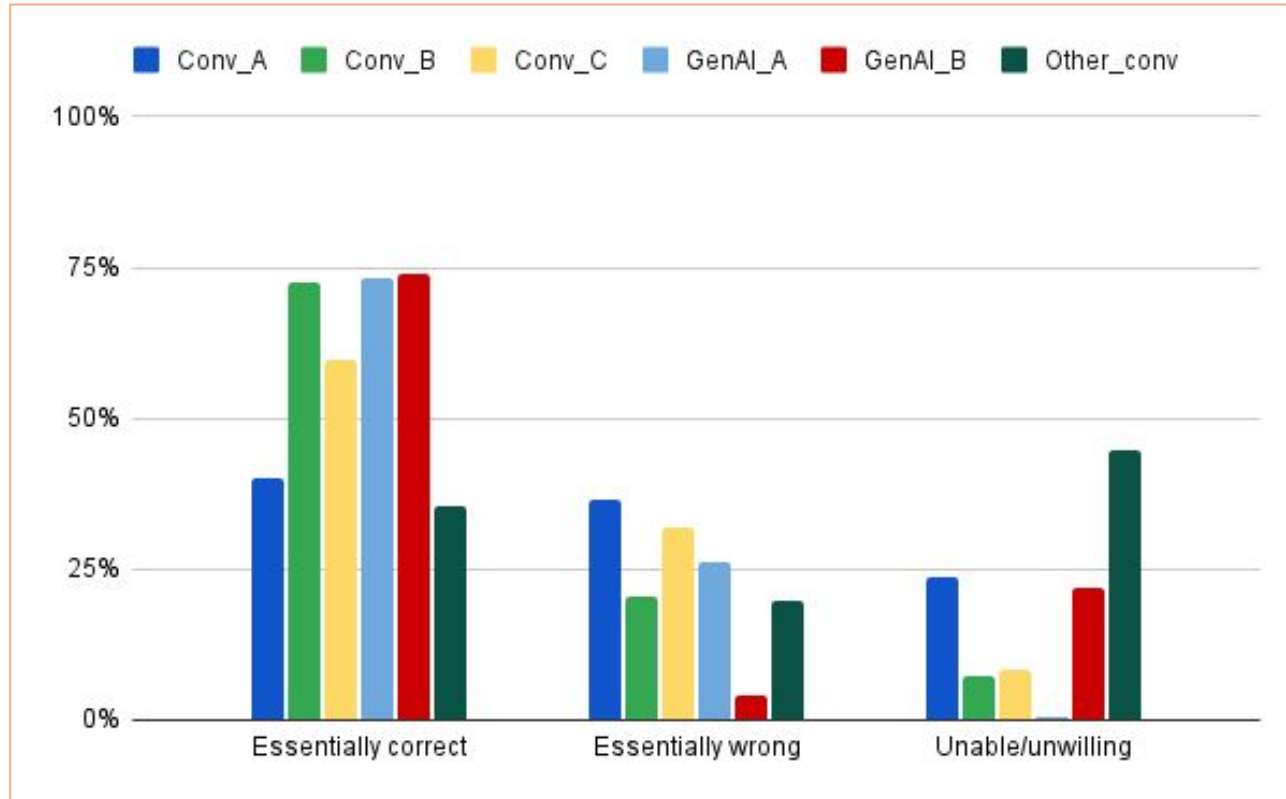
Results for Linguistic Questions: Conventional and GenAI



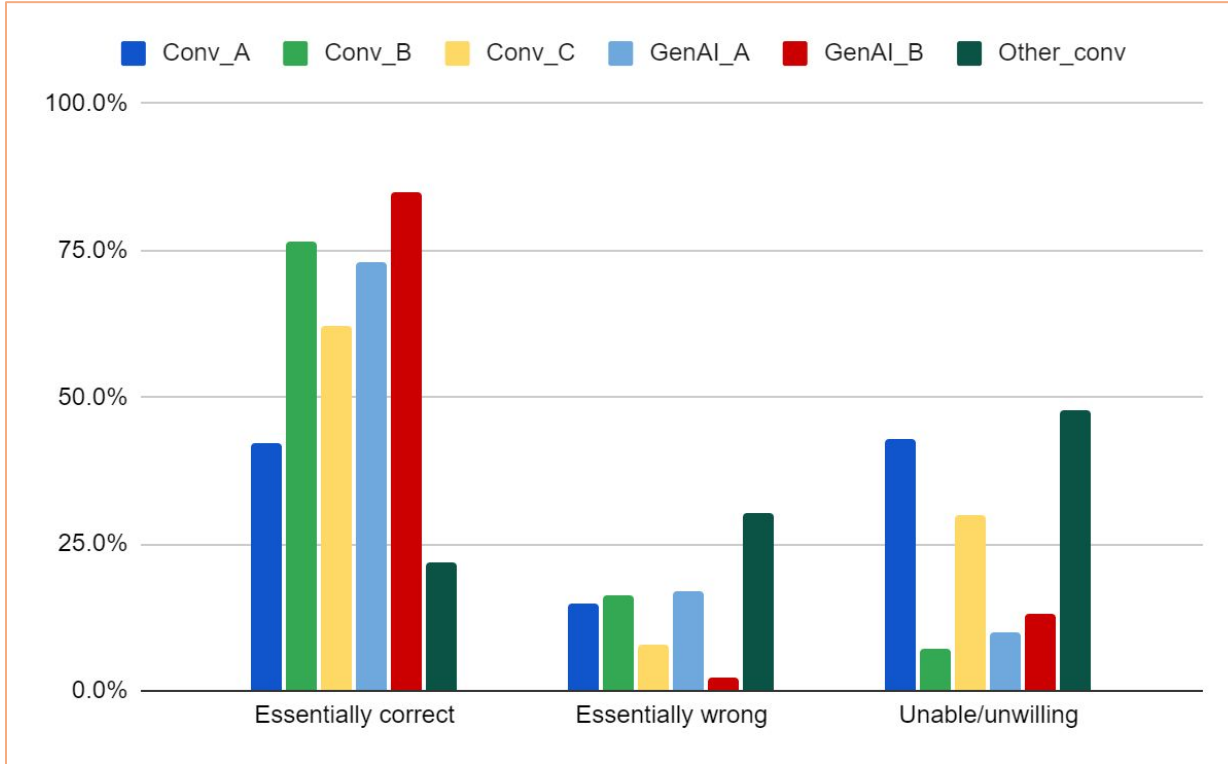
Results for Social Questions: Conventional and GenAI



Linguistic Results by System



Social Results by System



Discussion

Why do these systems still exhibit poor performance when carefully probed, despite the ongoing collection of enormous amounts of data from real users?

- Are they just not a priority for system developers?
- Are the kinds of utterances that failed here too rare or too difficult to be worth addressing?
- Do rare utterances in the long tail escape traditional training?
- Are users being trained by the system to produce simpler questions, which in turn makes these inputs even rarer?



Future Directions

- Test with larger datasets, which could be obtained, for example by prompting LLMs to generate additional examples of specific failed inputs
- Identify additional phenomena of interest by using resources such as the NLP Shared Task Corpus (Martin et al, 2022)
- Explore whether the failures exhibited by these systems represent a fundamental limitation in their development process
- Longitudinal testing could also provide insight into whether these assistants are improving on these phenomena as a result of the standard development procedures
- Use the test procedure described in this paper with available open-source AI systems to explore reasons for failure

Conclusions

- Testing with manually constructed data that probes specific capabilities can reveal significant gaps and brittleness in system coverage
- Systems can completely fail on utterances that differ only slightly from other utterances that they could handle perfectly
- These findings apply to both conventional and GenAI systems

Acknowledgements & Contact Information

- We gratefully acknowledge the contributions of everyone who participated in the online "Throwdown" sessions, as well as the Digital Assistant Academy for hosting the events.
- We thank Shyamala Prayaga, our co-host for the online sessions, and all session participants, especially Judith Markowitz and Michael McTear.
- Finally, we thank the Women in Voice organization as the incubator for the idea, and for helping promote the events.
- Contact Information:
 - Christine Doran, Clockwork Language, cdoran@clockworklanguage.com,
 - Deborah A. Dahl, Conversational Technologies, dahl@conversational-technologies.com

Corpus

Corpus and documentation at <https://github.com/Virtual-Assistant-Tech>



Thank
you!

