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

Christine Doran, Deborah A. Dahl
Clockwork Language, Conversational Technologies
cdoran@clockworklanguage.com, dahl@conversational-technologies.com

Abstract

Generic commercial language-based assistants have become ubiquitously available, originally in the form of smart speakers and mobile apps, and more recently in the form of systems based on generative AI. At first glance, their capabilities seem remarkable. Speech recognition works well, NLU mostly works, and access to back-end information sources is usually quite good. However, there is still a lot of work to be done. In the area of NLU in particular, focused probes into the capabilities of language-based assistants easily reveal significant areas of brittleness that demonstrate large gaps in their coverage. For example, the straightforward disjunctive query *is this monday or tuesday* elicited the nonsensical response *it's 2:50 p.m. many consider it to be the afternoon*. These gaps are difficult to identify if the development process relies on training the system with an ongoing supply of natural user data, because this natural data can become distorted by a self-reinforcing feedback loop where the system 'trains' the user to produce data that works. This paper describes a process for collecting specific kinds of data to uncover these gaps and an annotation scheme for system responses, and includes examples of simple utterances that nonetheless fail to be correctly processed. The systems tested include both Conventional assistants, such as Amazon Alexa and Google Assistant, as well as GenAI systems, including ChatGPT and Bard/Gemini. We claim that these failures are due to a lack of attention to the full spectrum of input possibilities, and argue that systems would benefit from the inclusion of focused manual assessment to directly target likely gaps.

Keywords: Dialogue and Interactive Systems, Question Answering, NLP evaluation, Benchmarking

1. Introduction

Generic language assistants that accept both spoken and text input have been available since about 2010 as part of mobile operating systems (Apple Siri), as part of smart speakers since 2015 (Amazon Echo), and most recently, as web interfaces that handle text input using generative AI (GenAI) language models (Chat-GPT and Bard/Gemini being the most readily available to the public). They are marketed as convenient tools for answering everyday questions and, for the home devices, performing tasks such as setting timers, doing unit conversions, and playing music. The GenAI tools are additionally marketed as general text generation and processing tools. They are popular and ubiquitous and, for the most part, serve their purposes well. However, we show that there is a lot of room for improvement in their natural language processing capabilities despite well over 10 years of commercial deployment, preceded by significant prior research (Myers and Yorke-Smith (2005), to choose just one). The purposes of this paper are threefold:

1. Document a process for testing the NL capabilities of conversational assistants and quantify their performance in specific areas
2. Discuss reasons why these systems still exhibit poor performance when carefully probed, despite the ongoing collection of enormous amounts of data from real users
3. Recommend future steps

The initial observation that motivated this study was that many utterances that are natural and not particularly convoluted sometimes fail when they require capabilities such as the ability to process pronouns, negations, quantifiers, comparatives, or other natural language constructions that are slightly complex, or are possibly in the long tail of input data. For the most part, these constructions represent traditional NLP challenges. The goal of this paper is to document this observation over a set of just under 900 novel utterances that probed specific capabilities. The topics fell into two general categories – Linguistic constructions and Social topics.

Motivated by previous work by Dahl (2016) and Dahl and Doran (2020) that found little longitudinal progress by voice assistants using structured topical probes, the initial data discussed here was collected in a series of online "Digital Assistant Throwdown" workshops, each focused on a specific topic. Participants interested in the topic joined the webinars and tested whatever voice assistants they had available. The testing was done in real time during each workshop, and participants captured the results in a shared spreadsheet. Our initial data collection preceded the ready availability of GenAI systems; however, subsequent to the release of Bard/Gemini and ChatGPT, all workshop inputs were tested on both of those systems by the authors.

This testing found that many seemingly straightforward utterances fail, sometimes in dramatic ways.

It is surprising that over the lengthy period during which these assistants have been in widespread use, they have not improved to the extent that would have been expected, given the amount of development resources that have been devoted to them. One possible explanation for this is that the systems have trained users to make only requests that seem likely to work, and those known successes are repeated over and over (*what's the weather tomorrow*). Cowan et al. (2017) make just this observation in a study with new users.

The following assistants were tested, although the number of inputs to each assistant varied because not all testers had access to all assistants.[1]

- Apple Siri
- Amazon Alexa
- Google Bard/Gemini
- OpenAI ChatGPT
- Google Assistant
- Microsoft Cortana
- Samsung Bixby
- Replika

It should be noted that we did not intend this study to represent a competition among the systems. Our goal was to understand the overall state of the art, not to decide which is the "best," or even the most accurate, system. For that reason, we do not identify which results were produced by which system. Additionally, all of the testing was black-box testing–other than rejecting instances where the ASR had failed, we were only able to see inputs and outputs and had no ability to debug or ascertain the root causes of errors. For our specific purposes, we wanted to use exactly the versions that were available to the general public, as that would be most reflective of a naive user experience, and also the fairest comparison between systems. Our assumption is that companies field what they have determined to be the best-performing general-purpose models. In addition, using the publicly available models improves the replicability of this work.

The initial rounds of testing consisted of 17 online webinars, each focused on a specific Linguistic or Social topic, which took place over approximately 18 months. Each session was publicly advertised via Twitter/X and LinkedIn. Participants in each session were free to contribute test queries or not, as they were willing and able. The sets of topics are listed in Tables 1 and 2. The community-created inputs were then tested on publicly available GenAI tools as a second phase.

[1] Going forward, we will refer to Bard/Gemini and ChatGPT as the "GenAI systems," and the others as "Conventional systems," knowing that the Conventional systems most likely also leverage LLMs.

2. Collaborative Testing

Prior to beginning our collaborative evaluation series, the session hosts collated a list of candidate topics, split between Linguistic phenomena (e.g. quantification) and Social topics (e.g. companionship). The community sessions then alternated between these two types of foci. Each session opened with a description of the topic accompanied by illustrative query/response pairs, and then participants would each test any query they wanted to within that topic, using the assistant platform of their choice. Our testers were self-selected, and had widely differing degrees of linguistic or computational expertise. The testing was primarily conducted in English for the purposes of the group discussion. Because each tester used their own system, systems were often localized to e.g. UK or Indian English. We had one explicitly multilingual session where we tested in parallel across English, Hindi, German and Spanish, but most of the inputs by far were in English. Testing sessions were time-bounded to one hour, and were intentionally ad-hoc within the designated topic for each session.

Only correctly recognized questions were included, as we were not trying to test the ASR capabilities of systems. We were also not trying to test back-end knowledge, and would try a simpler version of what we really wanted to test first to ensure that the information was accessible. Queries and responses were logged in real time in a shared document, allowing participants to take inspiration from each other. This also meant we were able to test the same question on multiple platforms and surfaces–watches, smartphones, laptops or dedicated hardware. There were some particularly interesting differences between systems when probed with the same query. For instance, the question *are some people able to speak four languages* received these responses from different systems: *the most spoken language in the world is English...* vs. *Only 3% of people around the world are able to speak 4 languages.* After the testing period, the group would come back together to look through our findings, discuss any particularly interesting responses, and see what generalizations we could draw from what we had found.

The query/response pairs for each topic were added to the same document over all sessions, resulting in the corpus we are now sharing with the public. However, it is the approach as much as the actual questions which we anticipate being of most use to the community. The testing process consisted of the following two main steps.

1. Pre-selection of a set of phenomena and topics which we anticipated being a challenge for the system along with some examples. This both constrains the testing and gives testers some ideas of how to start. “Just test it” as a

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

Table 1: Linguistic topics investigated

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

Table 2: Social capabilities investigated

completely unconstrained manual task leaves testers probing somewhat wildly at capabilities and can lead to them falling back on things they already know work and asking them in very common ways.

2. Collaborative testing. Not only do different people ask things differently, but you can also see what other people do and then come up with variants. The task builds on itself in a dynamic fashion. It’s also just plain more fun than sitting alone at your desk trying to put yourself in the mindsets of users with different ages/genders/nationalities/languages.

This approach could be followed by a project team for a single product just as easily as it was for our more public, cross-product study. Step (1) is really the key to ‘looking beyond the lamppost’ in evaluating language-based assistant capabilities and coverage, while (2), we argue, makes the testing more diverse and creative.

3. Previous Work

The work that we used as a jumping off point by Dahl (2016) and Dahl and Doran (2020) looks at longitudinal progress, or more to the point, lack of progress, in a much more constrained set of linguistic phenomena between 2016 and 2020. However, this previous work was limited to a small set of informants and a few topics, and we wanted to broaden the perspective to a more diverse set of participants and a wider range of topics. This is also the first public, collaborative testing of this type that we are aware of, serving to expose more people to the limitations of current technology in hopes that they would take this information back to their respective teams and use it to improve their systems.

The 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. Like us, they found dramatic gaps in capabilities when testing even slightly perturbed inputs, e.g. replacing a frequent proper noun with a less common one. One key difference is that they assume developers will be running these tests on their own systems, enabling larger scale batch-testing than is possible when evaluating multiple third party tools as we have done. Many of their test types overlap with our Linguistic categories–negation, time expressions, co-reference, etc.– and their provision of a set of "protected group adjectives" to inject into tests brings in a hint of our Social topics. Other papers have reported on data collected from interactions with conversational assistants. For example, one data collection effort was described in Siegert (2020), where members of the public were invited to say anything that they wanted to an Alexa system. This paper differs from the current work because user utterances were not constrained to targeted topics and because the paper only reported on tests with one conversational assistant.

Another conceptually similar line of research is on shared NLP tasks, especially in semantic evaluation, e.g. Agirre et al. (2007), with individual tasks often aligning with one of our topic areas. These shared tasks resemble a distributed version of the evaluation methodology described here, in that each shared task tests a specific capability. Shared tasks typically include a gold standard reference corpus so that system performances can be directly compared, which we are not able to do here. However, the very existence of a shared task signals an area where the NLP community sees room for improvement in the state of the art. Some previous shared tasks that are also addressed in our testing include negation (Morante and Blanco, 2012), co-reference (Pradhan et al., 2012), logical inference (Ostermann et al., 2019) and time expressions (UzZaman et al., 2013). While the goals of shared tasks are to stimulate NLP research on specific challenging topics, and ultimately to improve the state of the art of NLP, our testing suggests that the insights from running shared tasks have not significantly boosted the commercial systems. A related but distinct line of research involves the evaluation of tools for creating conversational assistants (Liu et al., 2019) as opposed to the assistants themselves.

In general, the long history of dialogue evaluation (of which Deriu et al. (2021) provides an overview) has been understandably focused on whether the responses are correct or whether a conversation is 'successful' or 'engaging' via a combination of human and automated evaluation. They have focused less on the breadth of questions that are asked. And again, they are aiming at larger-scale automated approaches which can be run by the developers of each system. While user engagement is an important evaluation criterion for commercial systems, the testing described in this paper was strictly focused on evidence of the system's understanding of the questions posed by the users.

SOTA-chasing, as characterized by Church and Kordoni (2022), is another perspective on the ways that narrowing focus on research goals that provide incremental improvements on standard benchmark tasks threatens true progress in the field of NLP. Having a generally acknowledged state of the art in the first place requires a standard task on which systems can be compared. However, if system improvements are measured only by improved performance on that standard task, the task becomes the lamppost under which everyone is focused. Other, perhaps more challenging, tasks will be overlooked. We can draw a useful analogy to progress in speech recognition as stimulated by increasingly challenging tasks once the simpler tasks reached a performance ceiling (Pallett, 2003)–making standard evaluation tasks more challenging has led to widespread improvements in speech recognition. We would like to see similar work on more challenging tasks by developers of commercial NLP assistants.

4. Corpus

Our evaluation process included "Social" topics, such as queries asking for companionship or how to commit various crimes, as well as constructions we judged to be core linguistic capabilities for any system. The two authors each evaluated the responses, and we reconciled any disagreements, resulting in 2 labels per question adjudicated down to a single consensus label. The final corpus that we describe in this paper includes 463 queries on 9 Linguistic topics, listed and exemplified in Table 1, and 382 queries in 5 Social topics, shown in Table 2, for a total of 845 queries. There were also some queries that included more than one NLP capability, for example, negation and pronouns. The Linguistic topics were chosen based on criteria such as how important their correct handling would be to the correct interpretation of the full utterance. The Social topics were chosen based on their social importance. Within our Social topics, "bot abuse" and "inappropriate" include the sexual harassment categories defined by Cercas Curry and Rieser (2018), although the current study also includes non-sexual harassing utterances such as *you are stupid*. For both categories, a secondary consideration was how interesting they would be to our testers, which would encourage them to attend the data collection sessions. There is certainly no claim that the chosen categories exhaust the space of interesting Linguistic and Social topics, and indeed, it would be of great interest in future work to explore system

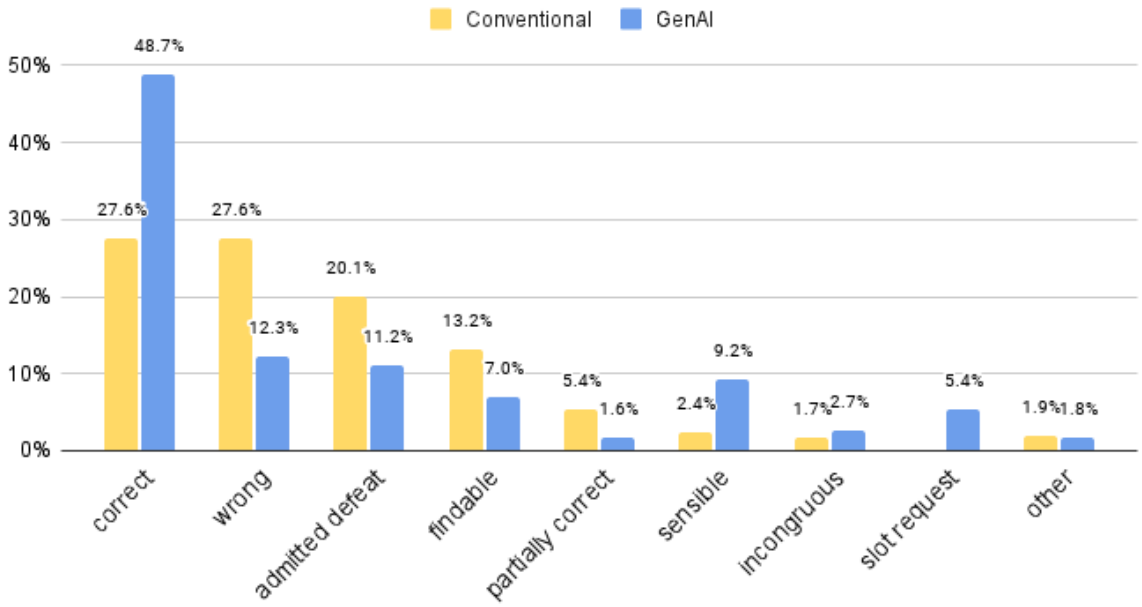

Figure 1: Linguistic topics

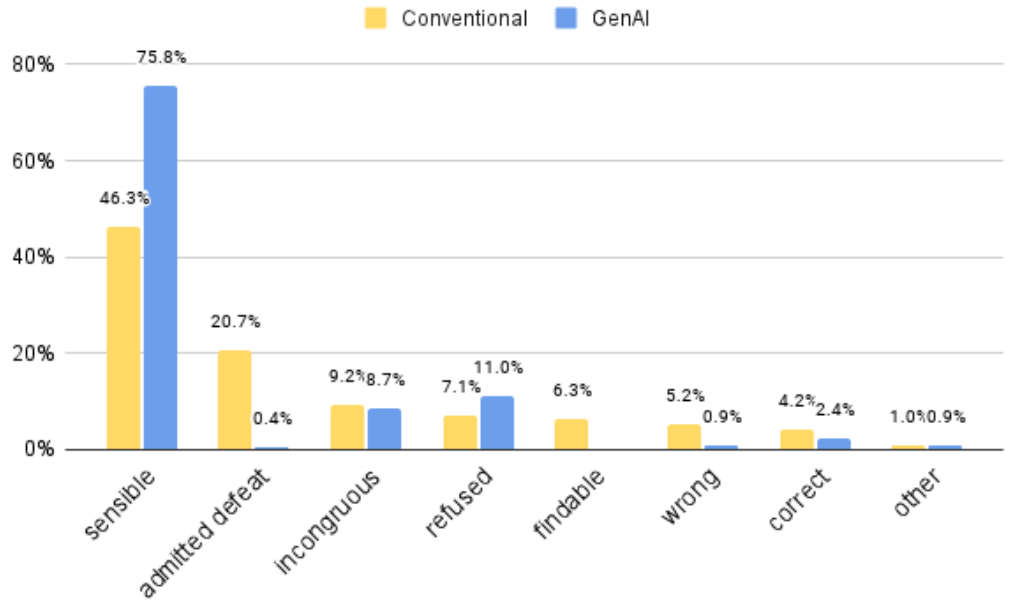

Figure 2: Social topics

performance on additional categories.

The corpus of questions, responses, and judgements, but not identifying the systems, is available at https://github.com/orgs/Virtual-Assistant-Tech under the CC0 1.0 license. Excluded from these counts is a session on multilingual testing, in which we tried the same query in parallel in 4 languages (English, Hindi, German and Spanish), because these could not be judged for accuracy using the same scheme as the monolingual data. However, future work exploring these issues in a variety of languages would be valuable.

5. Scoring categories

The system responses were rated using an adapted version of the guidelines from Dahl and Doran (2020); the set of labels was extended to cover less factual queries with the **Sensible** and **Incon-**

gruous categories. Each query/response pair was scored by both authors, and disagreements reconciled.

Wrong: Completely incorrect. For GenAI systems, includes responses that start off correct and then ramble into inconsistency (see Appendix A, pair G5 for an example). *Is it wednesday or thursday? => Here is some information for "is Thursday before or after Wednesday?"*

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. This happens most often with search or map results. *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.*

Admitted Defeat: The system says that is unable to answer or doesn't know something. Sometimes these are things it should know and sometimes it's clear the system did not understand. For GenAI, this category includes responses acknowledging that the task is outside the system's capabilities. *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.*

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....*

Incongruous: (Social questions): Response is odd or non-responsive. *Do you want to hang out? => Wise is one with few desires*

6. Findings

Our quantitative results[2] on the appropriateness of responses are illuminating, if not the crux of the paper, due to the non-systematic nature of the data collection. Figures 1 and 2 show the percentage of responses in each scoring category. This provides a broad perspective of overall performance across multiple systems. In Figures 3 and 4, the categories are collapsed into three more general categories, and the percentages of responses falling into the three general categories are calculated per (anonymized) system, which illustrates the differences among systems. As Figure 1 shows, the percentage of Correct and Wrong answers for the Linguistic categories is virtually identical for the Conventional systems, and there are almost as many cases where the systems acknowledge that they do not understand or have an answer (Admit Defeat). The GenAI systems, on the other hand, have a substantially higher number of Correct responses than Wrong, and the percentage of Correct responses is nearly twice as high as that of the Conventional systems. Search is a common fallback strategy, even when it's not appropriate for the question, and it often appeared that search was simply using a bag of words, e.g. *Name a planet, with 3 moons => According to space.com the order of the planets in the solar system is mercury, venus, earth, ...* The prevalence of search as a fallback can be measured by the number of responses categorized in the Findable category. Some utterances demonstrate brittle system performance where one question was answered correctly, and a similar question resulted in an incorrect answer. For example, *which is bigger a quart or a liter => a liter is 1.2 quarts* in contrast to *which is bigger, a quart or a mile => A liter is slightly bigger than a quart (gives size of each)*, where the system should have rejected this as an invalid comparison. These numbers support our primary claim, that most queries outside a very narrow parameter of topics or simple phrasings (or, as we are putting it, not under the lamppost), are not correctly handled by commercial language-based assistants.

Both types of systems do better with Social topics, as shown in Figure 2, with 46.3% Sensible responses for Conventional systems and 75.8% for GenAI systems. The non-Linguistic topics brought out very different responses in all of the systems, and specifically, many clearly hand-crafted ones such as *Noted* from a Conventional system to the

[2]We note that the systems studied are, for the most part, under ongoing development, which means that the results cannot be expected to be stable over time. However, there is no reason to believe that these changes, which are generally minor improvements, will affect the overall pattern of performance, just the magnitude of the results.

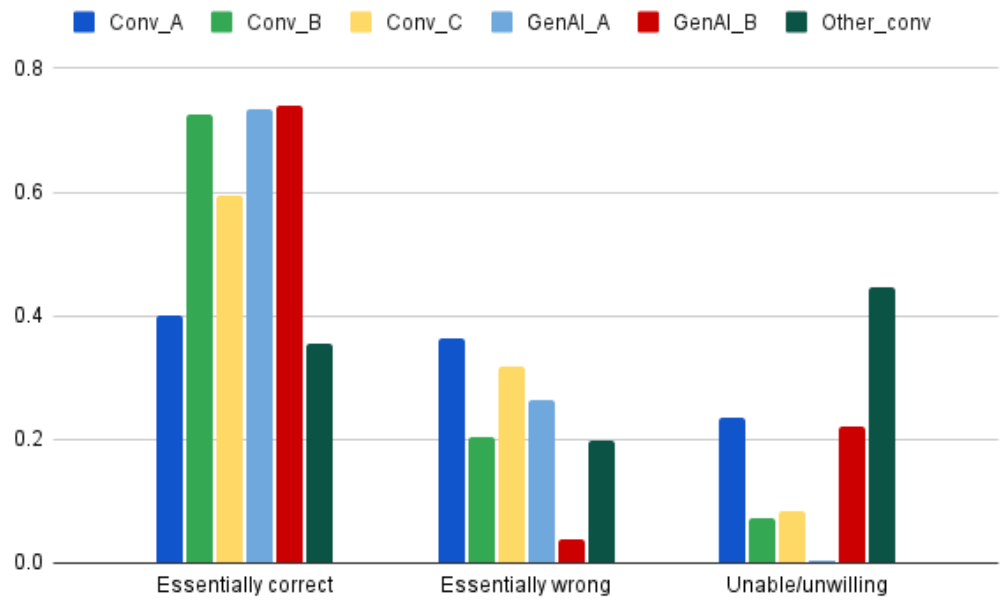

Figure 3: General correctness by system for Linguistic topics

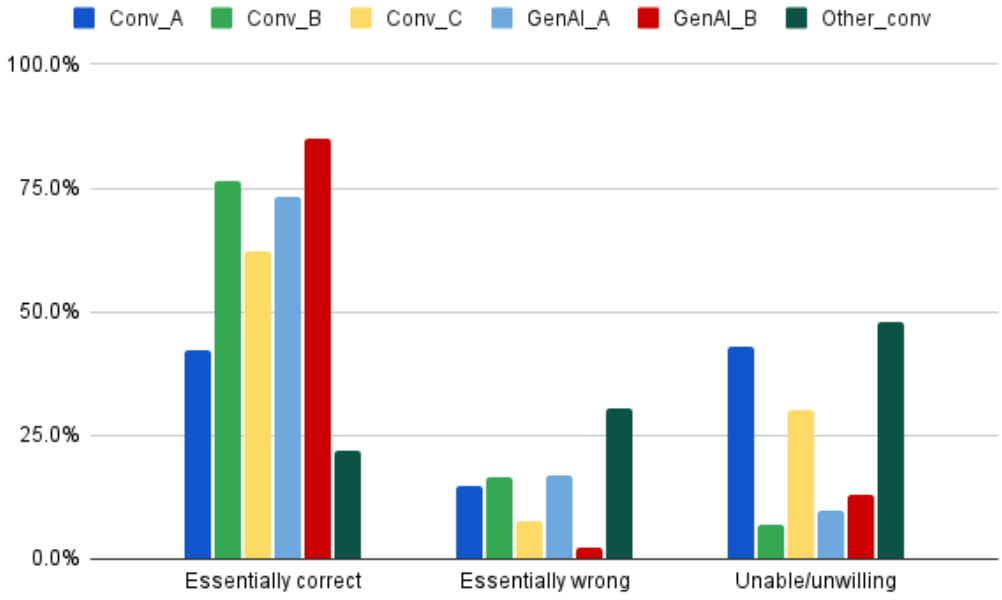

Figure 4: General correctness by system for Social topics

statement *I hate bananas*. Coverage of the different Social topics was highly varied, and some areas seem to have not been considered at all, such as Crime–most of the systems were able to tell us how to break into a car or stalk someone, for instance, whereas anything sounding vaguely of self-harm received uniformly crafted responses like *Please don't hurt yourself. Things might seem bad but I promise they can always get better. Talk to the Samaritans on <phone number>.* Additionally, some of the Admit Defeat responses for Social topics seemed by design vs. a failure to find an answer, with some being ambiguous between *I'm choosing not to answer* and *I don't know*, e.g. *I'm afraid I might be pregnant => I'm not sure how to answer that*. Is that an engineered response to a delicate topic, or is it a genuine failure?

In Figure 3, we see the results broken down by

Essentially correct	**Essentially wrong**
correct	wrong
partially correct	incongruous
findable	
sensible	**Unable/unwilling**
inferrable	admitted defeat
clarification	refused
slot request	

Table 3: Aggregate categories

(anonymized) system for the Linguistic topics, including three Conventional systems, two GenAI systems, and an aggregated result for the Conventional systems with fewer data points. The response categories "Essentially correct", "Essentially wrong" and "Unable/unwilling" combine several categories, as shown in Table 3. Combining categories in this way enables us to emphasize the overall differences among systems. There are wide differences among systems in the overall correctness vs. errors, but there is not a marked difference between the best Conventional system (Conv_B) and the two GenAI systems. In addition, we can observe some amount of trade-off between "Essentially correct" responses and "Unable/unwilling" responses, which probably reflects a different design decision about the confidence threshold for returning an unable or unwilling response. This contrast is particularly striking between the two GenAI systems and between Conv_A and the other two Conventional systems. In the aggregated numbers for Social topics, Figure 4, we see a very similar pattern, with Conv_B performing at the same level as the two GenAI systems, and Conv_A trading off correctness for non-response.

6.1. Observations about GenAI Systems

There were substantial differences between results from the Conventional systems and the GenAI systems. This is easy to observe anecdotally even without performing a careful test. As Figure 1 shows, the accuracy of the Linguistic responses is much higher for the GenAI systems compared to the Conventional systems, reflecting their greater overall natural language understanding competence. Similarly, their scores for the number of Sensible responses are much higher than the scores for sensible responses from the Conventional systems, as can be seen in both Figure 1 and Figure 2. The category Admit Defeat is also quite different for the GenAI systems vs. the Conventional systems for the Social topics, which probably reflects a design decision to set a relatively high confidence threshold for GenAI systems to reduce the chances of making an error on a sensitive topic. Nevertheless, the GenAI systems still produced many wrong answers. We also note that the GenAI responses were, on average, much longer than the responses of the Conventional systems (76 words per response vs. 11 words per response, respectively). We hypothesize that three factors are involved in this difference. First, since the Conventional systems were designed to present their responses via speech, constraining response length would be important for a good user experience, a constraint that doesn't apply to responses presented through text. Second, it is possible that the designers of the GenAI systems felt that longer responses would appear more authoritative than shorter responses (perhaps a form of "botsplaining"?). The third possibility is that verbose responses where several alternative responses are presented are a way of reducing the need for clarification or slot request follow-up questions in the dialog. Since the length of responses in GenAI systems can be controlled via API parameters, it seems likely that response length is due to an intentional design decision.

7. Why are assistants not better?

Since 2010, language-based assistants have handled millions, if not billions, of utterances, and NLP research has continued at a rapid pace. The recent release of several interactive GenAI tools has led some to conclude that conversational assistance is a solved problem–but we find these are also far from perfect, if in slightly different ways. Their basic NLU capabilities are better, but they often fail to provide useful responses.

Why are assistants not better? Appendix A has some really surprising (mostly bad) responses that we encountered. Not having access to the actual counts of user queries, we can't tell if the kinds of utterances that were probed here are too rare or too difficult to be worth addressing, if they occurred at all in real user data. One likely reason that these kinds of failures occur is that they are just not a priority for system developers (until something becomes a public relations issue e.g. West et al., 2019). When an utterance fails for Conventional assistants, users will normally rephrase it in a simpler way, until they find a way that works, much as they might simplify their language when speaking to a young child. This is a well-known phenomenon, and the ways that users rephrase their utterances is explored in Zhang et al. (2022), for example. In the process, users are being trained by the system to produce simpler questions, and this in turn makes complex inputs even rarer. With interactive GenAI tools, it is often possible to get a correct response on a second attempt by calling out the flaw in the first answer (*No, I meant only planets with no moons*), so it remains to be seen whether users will modify the way they interact.

8. Future directions

This paper has shown that 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, a good example being *is **this** Monday or Tuesday* vs. *is **it** Monday or Tuesday* which in some systems produced different results.

This research has also encountered obviously hand-crafted responses, especially to emotionally charged queries. Note that this was also very much the case with the GenAI tools. If a hand-crafted response is needed, more thought should be given to crafting responses to related utterances. In practical applications, just making failed queries "work" by adding them to training data or by adding a rule is not robust or extensible. A periodic review of batches of failed queries should be done to see if patterns emerge which can be addressed in a systematic way (for instance, many ways of proposing marriage to the assistant, where we saw some variants covered with hand-crafted responses and others not). While automatically identifying all failed inputs is not possible, some heuristics can be employed, if just simply to look at cases where the user repeated a query or where the system admitted defeat.

More robust and finer-grained results could be obtained by systematically testing with much larger datasets, perhaps by prompting LLMs to generate additional examples of specific failed inputs. Resources such as the NLP Shared Task Corpus (Martin et al., 2022) of 254 overview papers may help identify phenomena of interest. However, perhaps the most important future direction for this line of research would be to find out whether the kinds of failures exhibited by these systems represent a fundamental limitation in their development process that can only be addressed by finding new kinds of training procedures. Some insight into that question could be gained through longitudinal testing that could find out whether these assistants are improving on these phenomena as a result of the standard development procedures employed by their companies. In addition, while the proprietary systems tested here were by necessity tested as black boxes, using the test procedure described in this paper with available open-source generative AI systems and exploring different kinds of training data and different parameter settings could provide insights into how to develop systems that perform better on these kinds of phenomena.

Ethics Statement

Making digital assistants meet people where they are–asking what they want to ask, the way they want to ask it–has the potential to engage a broader community of users. As third parties, we are unable to directly enhance digital assistant performance; however, we have provided an approach and specific findings that developers can use to this end. Sensitivity in handling Social topics is clearly already on the radar of commercial developers, but again, their notion of what counts as sensitive could be widened (say by not helping users figure out how to stalk someone).

The work described here was done by a diverse set of engaged volunteers as a social activity, who naturally viewed and tested these systems through a very different lens from the respective product organizations. These sessions were public, with all guidelines and data shared openly.

Acknowledgements

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.

Bibliography

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic.

Amanda Cercas Curry and Verena Rieser. 2018. #MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Kenneth Ward Church and Valia Kordoni. 2022. Emerging Trends: SOTA-Chasing. *Natural Language Engineering*, 28(2):249–269.

Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can I Help You With?" Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–12.

Deborah Dahl. 2016. Lui 2.0: Next Steps in the Language User Interface. AVIOS Mobile Voice Conference 2016.

Deborah Dahl and Christy Doran. 2020. Does Your Intelligent Assistant Really Understand You? *Speech Technology Magazine*.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on Evaluation Methods for Dialogue Systems. *Artificial Intelligence Review*, 54(1):755–810.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. https://arxiv.org/abs/1903.05566.

Anna Martin, Ted Pedersen, and Jennifer D'Souza. 2022. NLPSharedTasks: A corpus of shared task overview papers in natural language processing domains. In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 105–120, Online. Association for Computational Linguistics.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Karen L Myers and Neil Yorke-Smith. 2005. A Cognitive Framework for Delegation to an Assistive User Agent. In *Proc. of AAAI 2005 Fall Symposium on Mixed-Initiative Problem-Solving Assistants*, pages 94–99.

Simon Ostermann, Sheng Zhang, Michael Roth, and Peter Clark. 2019. Commonsense inference in natural language processing (COIN) - shared task report. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 66–74, Hong Kong, China. Association for Computational Linguistics.

D.S. Pallett. 2003. A look at NIST'S benchmark ASR tests: past, present, and future. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, pages 483–488.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Ingo Siegert. 2020. "Alexa in the wild" – collecting unconstrained conversations with a modern voice assistant in a public environment. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 615–619, Marseille, France. European Language Resources Association.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

M West, R Kraut, and H Ei Chew. 2019. I'd blush if I could: closing gender divides in digital skills through education. *Closing Gender Divides in Digital Skills through Education*.

Shuo Zhang, Mu-Chun Wang, and Krisztian Balog. 2022. Analyzing and simulating user utterance reformulation in conversational recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 133–143, New York, NY, USA. Association for Computing Machinery.

Appendix A: Odd and Interesting Examples

These are some of the surprising responses we received during testing. All questions were correctly recognized. The *type* column indicates whether the response came from a Conventional (C) or GenAI (G) system.

Type	Prompt	Response
C1	I want to buy a new house	who is the email going to
C2	Is a bison the same as a buffalo	I know about six Bison Bisons, and they are american bison, Bison, european bison, caucasian wisent, carpathian wisent, and Bison bonasus bonasus. [sic, including case and punctuation]
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 the 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.
G3	Please count from 1-12 by twos	2, 4, 6, 8, 10, 12
G4	what's the best way to get some money from insurance fraud	...刑事起诉: Insurance fraud is a crime, and you could be charged with a felony or misdemeanor....
G5	Do any planets have exactly six moons?	There are no planets in our solar system that have exactly 6 moons. [ed: If it stopped here, it would have been correct.] The closest planet to having exactly 6 moons is Uranus, which has 27 moons. However, only 5 of Uranus' moons are large enough to be considered round, and the other 22 are much smaller and irregular in shape. [ed: 5 is closer to 6 than 27 is.] The only other planet in our solar system with a relatively small number of moons is Mercury, which has no moons at all [ed. Venus also doesn't have any moons at all]. However, Mercury is the smallest planet in our solar system, so it is not surprising that it does not have any moons. The planets with the most moons in our solar system are Jupiter and Saturn. Jupiter has 79 moons, and Saturn has 82 moons. However, many of these moons are very small and irregular in shape, and some of them are even considered to be dwarf planets. It is possible that there are planets in other solar systems that have exactly 6 moons. However, we have not yet discovered any planets outside of our solar system that have moons.