Do Explanations Help Users Detect Errors in Open-Domain QA? An Evaluation of Spoken vs. Visual Explanations

Ana Valeria González[*1], **Gagan Bansal**[2], **Angela Fan**[3,4], **Yashar Mehdad**[3], **Robin Jia**[3], and **Srinivasan Iyer**[3]

[1]University of Copenhagen, [2]University of Washington, [3]Facebook AI, [4]LORIA
ana@di.ku.dk
bansalg@cs.washington.edu
{angelafan, mehdad, robinjia, sviyer}@fb.com

Abstract

While research on explaining predictions of open-domain QA systems (ODQA) is gaining momentum, most works do not evaluate whether these explanations improve user trust. Furthermore, many users interact with ODQA using *voice*-assistants, yet prior works exclusively focus on *visual* displays, risking (as we also show) incorrectly extrapolating the effectiveness of explanations across modalities. To better understand the effectiveness of ODQA explanations strategies in the wild, we conduct user studies that measure whether explanations help users correctly decide when to accept or reject an ODQA system's answer. Unlike prior work, we control for explanation *modality*, *i.e.*, whether they are communicated to users through a spoken or visual interface, and contrast effectiveness across modalities. We show that explanations derived from retrieved evidence can outperform strong baselines across modalities but the best explanation strategy varies with the modality. We show common failure cases of current explanations, emphasize end-to-end evaluation of explanations, and caution against evaluating them in proxy modalities that differ from deployment.

1 Introduction

Despite copious interest in developing explainable AI, there is increasing skepticism as to whether explanations (of system predictions) are useful to end-users in downstream applications. For instance, for assisting users with classifying sentiment or answering LSAT questions, Bansal et al. (2021) observed no improvements from giving explanations over simply presenting model confidence. Similarly, Chu et al. (2020) observed that visual explanations fail to significantly improve user accuracy or trust. Such negative results present a cautionary tale for explainability and emphasize the need to evaluate explanations using careful user studies.

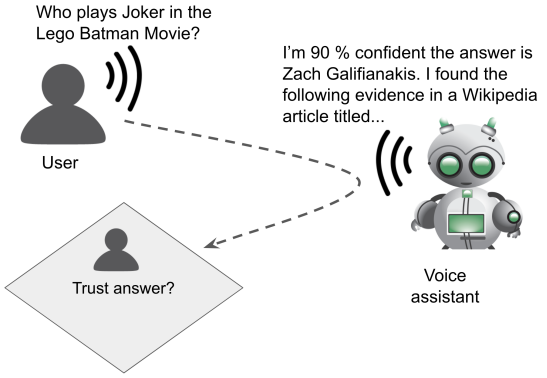

Figure 1: Using end-to-end user studies, we evaluate whether explanation strategies of open-domain QA assistants help users decide when to trust (or reject) predicted answers.

We explore the effectiveness of explanations for ***O**pen-**D**omain **Q**uestion **A**nswering* models, which involves answering users' questions (*e.g.*, *"Who plays the Joker in the Lego Batman movie?"*) using a large corpus (*e.g.*, Wikipedia). Such models are increasingly deployed not only in visual modalities (*e.g.*, Web search) but also in spoken ones (voice assistants).[1] Spoken interfaces for ODQA are also important because they make ODQA systems more accessible for users with visual impairments. Despite improvements in accuracy, deployed ODQA models remain imperfect. This motivates the need to provide users with mechanisms (*e.g.*, estimates of uncertainty or explanations) that can help improve *appropriate reliance* (Lee and See, 2004), *e.g.*, by allowing users to detect erroneous answers. We henceforth refer to a user's ability to distinguish correct and incorrect answers as *error-detectability*, and ask *Does explaining the system's reasoning, help improve error-detectability?* (Figure 1).

Alongside recent negative results (Bansal et al., 2021), Lamm et al. (2020) showed that visually complex "QED" explanations that communicate

[*] Work done while at Facebook AI.

[1] https://www.perficient.com/insights/research-hub/voice-usage-trends

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1103–1116
August 1–6, 2021.

coreference and entailment information along with evidence marginally improve error-detectability. However, the study lacks the recommended baseline (Amershi et al., 2019; PAIR, 2019) of communicating model confidence which has been shown to be effective on other domains (Bansal et al., 2021). Also, the transferability of complex visual explanations to the spoken modality remains unclear. Although Feng and Graber (2019) compare visual explanations with presenting model confidence on a different QA task, i.e., answering timed, multi-clue trivia questions, it was unclear whether explanations led to appropriate reliance (Bansal et al., 2021); thus the effectiveness of explanations for end users of QA systems still remains unclear. In this paper, we set out to evaluate the ability of NL explanations in both visual and spoken modalities, to improve error-detectability for the task of ODQA for non-expert users over strong baselines.

However, explaining ODQA systems in the spoken modality may pose unique challenges, *e.g.*, because the same information content can impose higher cognitive demands when communicated by voice than visually (Sweller, 2011; Leahy and Sweller, 2016); potentially reducing effectiveness of longer, more complex explanations (*e.g.*, QED) in the *spoken* modality. Thus we also ask, *Can the most useful explanation strategy change with the modality?* In summary:

1. We present user studies evaluating how well explanations for ODQA help users detect erroneous answers (error-detectability). Unlike prior work, we evaluate explanations in both visual and spoken interfaces, and compare against calibrated confidence.
2. Our experiments with over 500 MTurk users confirm significant improvements in error-detectability for ODQA over showing confidence. To the best of our knowledge, our work is the first to show statistically significant improvements in appropriate reliance through NL explanations for non-expert users. (Section 4.1)
3. We show that the best explanation approach can change with the modality: while longer explanations (evidence paragraphs) led to the highest error-detectability in the visual modality, shorter explanations (evidence sentence) performed best in the spoken modality. We connect our observations with prior work on cognitive science and identify failure cases for ODQA explanations (Section 4.3).

2 Background

Open-domain QA. ODQA involves answering questions using a large, broad corpus of unstructured documents (*e.g.*, Wikipedia or the Web). More specifically, the questions are factoid and the target answer is present as a span in one of the documents (Voorhees et al., 1999).

Recent models for ODQA use a pipelined approach and contain two components: a document *retriever* that finds a subset of the most relevant documents from the corpus, and a *reader* that selects an answer span from the retrieved documents (Chen et al., 2017; Lee et al., 2019a). We use a state-of-the-art ODQA model and a benchmark dataset that contains questions asked by real lay users (Section 3.3).

An ODQA model's prediction can be explained by providing a justification in natural language, *e.g.*, by extracting snippets of text from the retrieved documents (*rationales*) or more generally by generating new text (*abstractive explanation*). For example, rationales can explain a text classifier using phrases in the input text that are relevant to the prediction (Lei et al., 2016). However, for some tasks, such as NLI (Camburu et al., 2018) and common-sense reasoning (Rajani et al., 2019), the input text alone may not contain enough meaningful justification for the prediction. For other tasks, the evidence can be spread across documents (*e.g.*, HotpotQA; Yang et al. (2018)). In such cases, abstractive explanations become more useful. While we primarily focus on extractive explanations, we also experiment with abstractive explanations (Section 3.1).

Evaluating explanations. One important reason to explain ODQA models is to improve error-detectability (Figure 1).[2] Many prior works evaluate the quality of NL explanations by comparing similarity with human-written explanations for tasks such as NLI, common-sense reasoning and fact verification (DeYoung et al., 2020; Paranjape et al., 2020; Swanson et al., 2020; Camburu et al., 2018; Rajani et al., 2019). Other works conduct user studies but rely on *proxy* tasks, *e.g.*, whether explanations allow users to anticipate the model predictions (Hase and Bansal, 2020; Nguyen, 2018). However, evaluating explanations on such

[2]Note that there exists other downstream applications of explanations, such as debugging models (Koh and Liang, 2017).

proxy tasks and metrics, which differ from the actual deployment setting, risks drawing misleading conclusions about the effectiveness of explanations in practice (Buçinca et al., 2020). Thus, we focus on directly evaluating explanations using user studies on error-detectability.

Feng and Graber (2019) found that explanations improve player accuracy when answering Quizbowl questions. Our task differs from Quizbowl in many respects: 1) Unlike Quizbowl where the users are trivia enthusiasts or experts, ODQA users are non-experts, lay people who ask questions to satisfy their information needs. Thus, ODQA users have no or very limited expertise in answering these questions. 2) While Quizbowl questions comprise multiple clues (incrementally revealed) for a single answer, ODQA questions typically contain a single clue. 3) Feng and Graber (2019) observed improvements from the explanations when their QA model was considerably more accurate than their users, outperforming the best trivia players. In contrast, we carefully design our user study so that on our study sample, users cannot achieve high performance by simply trusting the model (Section 3.3).

Visual vs. spoken modalities. We hypothesize that differences in processing of spoken and written information can substantially impact the effectiveness of NL explanations in ODQA. For example, Flowerdew et al. (1994) observed that one of the main differences in processing spoken versus written information is linearity. When listening, information progresses naturally, as opposed to reading, where people are able to jump back and forth in the text (Buck, 1991; Lund, 1991). This results in differences in recall of information across the two modalities (Osada, 2004). Although it is possible to repeat spoken information, Lund (1991) found that for some listeners, listening to information again was not as effective as re-reading. Another difference is the effect on concentration across modalities. Thompson and Rubin (1996) found that the heavier cognitive load imposed by listening to information can make people lose concentration more easily. Our experimental setup is the first to evaluate explanation effectiveness across these two modalities.

3 Experimental Setup

We evaluate explanation effectiveness for ODQA by varying the *type* of explanation and *modality* of communication. We combine variations of each factor to obtain explanation conditions (Section 3.1) for a state-of-the-art ODQA model (Section 3.3). We then deploy these conditions on Amazon Mechanical Turk (MTurk) to validate our hypotheses about the effectiveness of improving error-detectability (Section 3.2). We justify various design choices made to ensure quality in Section 3.4.

3.1 Explanation Types and Conditions

ODQA models can justify their predictions by pointing to *evidence* text containing the predicted answer (Das et al., 2018; Lee et al., 2019a; Karpukhin et al., 2020). We experiment with two types of *extractive* explanations:

- EXT-SENT: Extracts a sentence containing the predicted answer as evidence.
- EXT-LONG: Extracts a longer, multi-sentence paragraph containing the answer as evidence.

While extractive explanations are simpler to generate, we also evaluate a third explanation type that has potential to more succinctly communicate evidence spread across documents (Liu et al., 2019).

- ABS: (Abstractive) Generates a new text snippet to justify the predicted answer.

Explanation conditions. For the *spoken modality*, we test five conditions (two baselines and three explanation types): (1) BASE: present only the top answer, (2) CONF, a stronger baseline that presents the top answer along with the model's certainty, (3) ABS, (4) EXT-LONG, and (5) EXT-SENT. In the *visual modality*, we test two explanation types: EXT-LONG and EXT-SENT. We implemented these two types to contrast their effectiveness effectiveness across modalities (Figure 2, Section 3.4). For all explanation conditions, we show confidence, mention that the answer was obtained from Wikipedia and provide the source article. [3]

3.2 Hypotheses

We investigated five (pre-registered) hypotheses about the relative performance of various explanation conditions at improving the accuracy of error-detectability (Section 3.4):

H1 CONF will improve accuracy over BASE.

H2 *Spoken* EXT-SENT will improve accuracy over CONF— the explanation would provide additional context to help validate predicted answers.

[3] Appendix A shows more examples.

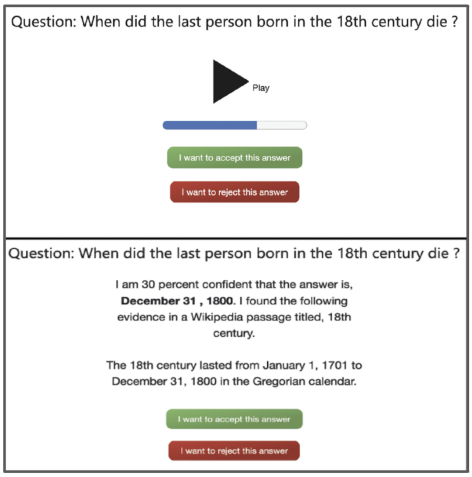

Figure 2: **U**I for spoken (top) and visual modalities (bottom) for the EXT-SENT explanation type. Users either read or hear an explanation and decide whether to trust or discard the QA system's prediction.

H3 *Spoken* EXT-SENT will lead to higher accuracy than *Spoken* EXT-LONG. Since the spoken modality may impose higher cognitive limitations on people (Section 2), concise explanations may be more useful despite providing less context.

H4 ABS will improve accuracy over *Spoken* EXT-SENT. ABS contains more relevant information than EXT-SENT (same length), which may help users make better accept/reject decisions.

H5 *Visual* EXT-LONG will lead to higher accuracy than *Spoken* EXT-LONG.

3.3 Implementation Details for Conditions

Dataset. We used the Natural Questions (NQ) corpus (Kwiatkowski et al., 2019). NQ is composed of anonymized queries posed by real users on the Google search engine. The answers are human-annotated spans in Wikipedia articles. As factoid web search is typically done through both spoken and visual modalities, this data is an ideal choice for our evaluation setup. To simplify the study, we restrict to questions with short target answers (< 6 tokens) (Lee et al., 2019b). This subset contains 80k training examples, 8,757 examples for development, and 3,610 examples for testing.

Model. We train the current (extractive) state-of-the-art model on NQ: Dense Passage Retrieval and Reader (DPR) (Karpukhin et al., 2020). Similar to Karpukhin et al. (2020), we split documents (Wikipedia articles), into shorter passages of equal lengths (100 tokens). To answer an input question, DPR uses two separate *dense* encoders $E_Q(\cdot)$ and $E_P(\cdot)$ to encode the question and all passages in the corpus into vectors. It then retrieves k most similar passages, where *passage similarity* to a question is defined using a dot product: $sim(q, p) = E_Q(q)^\intercal E_P(p)$. Given the top k passages, a neural reader (Section 2) assigns a passage selection score to each passage, and a *span score* to every answer span. [4]

Generating explanations. Extractive explanations use the passage associated with DPR's answer— EXT-SENT uses the sentence containing the answer whereas EXT-LONG uses the entire passage. Since DPR does not generate abstractive explanations, we simulate ABS by manually creating a single sentence that captures the main information of EXT-SENT and adds additional relevant information from EXT-LONG, whilst remaining the same length as EXT-SENT. To improve transparency, all explanation conditions also inform the source to the users, by providing them the *title* of the article. Figure 2 shows an example of the final EXT-SENT explanation condition. To convert text to speech, we use a state-of-the-art TTS tool. When spoken, questions in our final ABS and EXT-SENT conditions were on average 15 seconds long, EXT-LONG was between 30-40 seconds.

Confidence calibration. Confidence scores generated by neural networks (*e.g.*, by normalizing softmax scores) often suffer from poor calibration (Guo et al., 2017).To alleviate this issue and to follow best practices (Amershi et al., 2019; Bansal et al., 2021), we calibrate our model's confidence using *temperature scaling* (Guo et al., 2017), which is a *post hoc* calibration algorithm suitable for multi-class problems. We calibrate the top 10 outputs of the model. We defer additional details of calibration to Appendix B.

3.4 User Study & Interface

We conduct our experiments using Amazon Mechanical Turk. For each of the 7 conditions we hire 75 workers, and present each with 40 questions (this amounts to a total of 21,000 data samples) one-by-one, while showing them the model's answer (along with other condition-dependant information, such as confidence or explanation) and ask them to

[4] We re-score each answer using the product of the passage and span score and use the highest-scored answer as the prediction— Our initial analysis showed that this re-scoring improved exact match scores of predicted answers.

either *accept* the model's prediction if they think it is correct or *reject* it otherwise. Figure 2 shows an example. Additional details about the platform and participants can be found in Appendix D.

Question selection. We sample a set of questions on which the model's aggregate (exact-match) accuracy is 50%; thus any improvements in error-detectability, beyond random, must be a result of users making optimal assessment about the model's correctness. To improve generalization, we average results over three such mutually exclusive sets of 40 questions. Before sampling the questions, we removed questions that were ambiguous or questions where the model was correct but the explanations failed to justify the answer. Appendix C contains additional details on question selection.

Incentive scheme. In addition to providing a fixed upfront pay of $10 for participating in the task, to encourage workers to engage, we also used a bonus-based strategy (Bansal et al., 2019) — When users accept a correct answer, we provide a 15 cent bonus, but when they accept an incorrect answer they lose the same amount. When they reject an answer, however, they do not receive any bonus.[5] This aims to simulate the real-world cost and utility of users choosing to believe answers of an ODQA model. The maximum cumulative reward is $ 2.70. These values were chosen to ensure workers earned minimum a $15 hourly wage.

Post-task survey. After the main task, we asked participants to (1) rate the length of responses, (2) rate their helpfulness and (3) give us general feedback on what worked and how explanations could be made better. For the *spoken modality*, we also asked participants to rate the clarity of the voice to understand if issues in text-to-speech confused them. Appendix E contains the complete survey.

Metrics for error-detectability. We quantify user performance at error-detectability using the following three metrics:

1. **Accuracy**: Percentage of times a user accepts correct and rejects incorrect answers. A high accuracy indicates high error-detectability.
2. **% Accepts | correct**: Indicates the true positive rate, *i.e.*, percentage of times the user accepts *correct* answers.

[5] if bonus is negative, no deductions re made from base pay. Bonus is instead set to zero

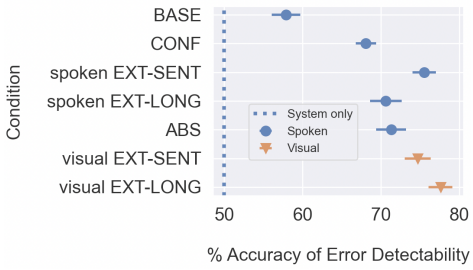

Figure 3: Accuracy of users. In the *spoken modality*, EXT-SENT explanations yield the best results and is significantly better than CONF. In the *visual modality*, EXT-LONG perform best. We observe a statistically significant ($p < 0.01$) difference between EXT-LONG in visual versus spoken, perhaps because of differences in user's cognitive limitations across modalities.

3. **% Accepts | incorrect**: Indicates the false positive rate, *i.e.*, percentage of times the user accepts *incorrect* answers. If a setting yields a high number, this would indicate that this setting misleads users more often.

We do not present true and false negative rates because conclusions are similar. We additionally measure time spent on each question and cummulative reward. These metrics are explained in Appendix F. When computing all metrics, we removed the first 4 questions for each worker to account for workers getting used to the interface. We preregistered this procedure prior to our final studies.

4 Results

To validate our hypothesis (Section 3.2) we compare explanation methods on the quantitative metrics (Section 4.1). To further understand participant behavior we analyze responses to the post-task survey (Section 4.2), and analyze common cases where explanations misled the users (Section 4.3).

4.1 Quantitative Results

Figure 3 displays average accuracy with 75 workers per condition. Similar to Lamm et al. (2020), to validate hypotheses and compute statistical significance, we fit a generalized linear mixed effects model using the `lme4` library in R and the formula `a~ c +(1|w) + (1|q)`, where `a` is accuracy, `c` is the condition, `w` is the worker id and `q` is the question id. We run pairwise comparisons of these effects using Holm-Bonferroni to correct for multiple hypothesis testing. For both the spoken and visual modalities, all conditions lead to significantly higher accuracies than BASE ($p < 0.01$).

Model confidence improved accuracy of error-detectability. In Figure 3, CONF achieves higher accuracy than BASE– 68.1% vs. 57.2%. This difference was statistically significant ($p < 0.01$), **validating H1**. While previous guidelines recommend displaying confidence to users (Amershi et al., 2019; Bansal et al., 2021), our observations provide the first empirical evidence that confidence is a simple yet stronger baseline against which explanations for ODQA should be compared.

Explaining via an evidence sentence further improved performance. The more interesting comparisons are between explanation types and CONF. In both modalities, EXT-SENT performed better than CONF. For example, in the *spoken modality*, EXT-SENT improved accuracy over CONF from 68.1% to 75.6% ($p < 0.01$); thus **validating H2**. Contrary to recent prior works that observed no benefit from explaining predictions, this result confirms a concrete application of explanations where they help users in an end-to-end task.

Longer explanations improved performance over concise explanations in the visual modality, but worsened performance in the spoken modality. Figure 3 shows that for the visual modality, EXT-LONG outperforms EXT-SENT explanations – 77.6% vs. 74.7% ($p < 0.4$). Conversely, for spoken, EXT-SENT is better than EXT-LONG– 75.6% vs. 70.4% ($p < 0.01$); thus **validating H3**. The decrease was severe enough that we no longer observed a statistically significant difference between EXT-LONG and CONF ($p = 0.9$), reemphasizing the importance of comparing against the latter. Although communicating the same content, *visual* EXT-LONG led to significantly better accuracy than their spoken version— 77.6% vs. 70.4% ($p < 0.01$); thus **validating H5**. These results indicate large differences, across modalities, in user ability to process and utilize explanations, and how these differences need to be accounted for while evaluating and developing explanations.

Despite improving conciseness, abstractive explanations did not help improve performance in the spoken modality. Figure 3 shows that ABS performs significantly worse than EXT-SENT in the spoken modality– 71.3% vs. 75.6% ($p < 0.01$) and thus we could **not validate H4**. This result indicates that the length of the explanation (*e.g.*, number of tokens) is not the only factor that affects user performance, instead, the density of information also increases cognitive load on users. This finding is in line with the Time Based Resource Sharing (TBRS) model (Barrouillet et al., 2007), a theory of working memory establishing that time as well as the complexity of what is being communicated, both play a role in cognitive demand. We also observe a similar effect in users' subjective rating of length of explanation (Section 4.2).

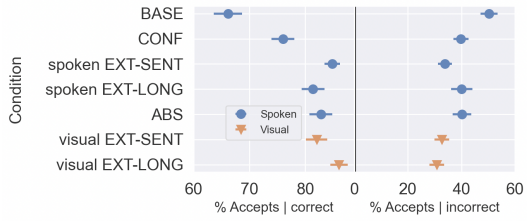

Figure 4: (Left) Explanations significantly increased participant ability to detect *correct* answers compared to CONF. (Right) However, only EXT-SENT in the spoken modality and both explanations in the visual modality decreased the rate at which users are misled.

All explanations significantly increased participants' ability to detect *correct* answers, but only some explanations improved their ability to detect *incorrect* answers. Instead of aggregate accuracy, Figure 4 splits and visualizes how often users accept correct and incorrect answers. For accepting *correct* model predictions, all *visual* and *spoken* explanation conditions significantly helped compared to CONF (at least $p < 0.05$).

For accepting incorrect predictions, in the *spoken modality*, only EXT-SENT is significantly better (*i.e.*, lower) than CONF—34% vs. 40% ($p < 0.05$). Whereas in the *visual modality*, both EXT-LONG and EXT-SENT lead to improvements over CONF—30% ($p < 0.01$) and 32% ($p < 0.05$), respectively. This shows that although explanations decrease the chance of being misled by the system, the least misleading explanations change with modality.

4.2 Qualitative Results

We analyzed user responses to the post-task survey to understand their experience, what helped them and how the system could serve them better. We discuss the main findings here and reserve additional results to the Appendix.

Length preference. We asked participants to rate the length of the explanation as *too short, short, right, long,* or *too-long*. Figure 5 shows the results. For EXT-LONG, over 85% of the workers perceived that in the *visual modality*, responses

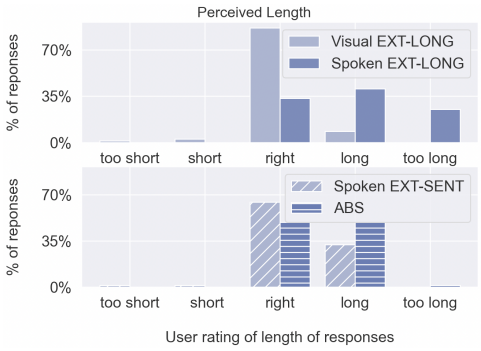

Figure 5: Users rated how they perceived the length of each explanation strategy. **Top:** Spoken explanations were perceived to be longer. **Bottom:** While EXT-SENT and ABS were the same length, the latter was rated as longer more often perhaps because of its complexity.

were the right length. On the other hand, in the *spoken modality*, only 30% of participants agreed the length was right. Thus, user's subjective ratings for the same explanation type were dramatically different across modalities. Indicating, in addition to affecting error-detectability, the modality also changes users' subjective preferences.

While ABS and EXT-SENT were the same duration, users rated ABS as longer than EXT-SENT. As mentioned before, this relates to the TBRS model of working memory (Barrouillet et al., 2007). We hypothesize that *our* ABS explanations, which integrate more information than EXT-SENT in the same amount of time, were more taxing for user's working memory, thereby reducing error-detectability and increasing perception of length.

User feedback. To understand how we can develop better explanations, we asked participants: *Do you have any additional feedback on what the system can improve?* To analyze responses, two annotators (authors) coded 400 responses. After removing responses that were not descriptive (*e.g.*, "can't think of anything to improve"), 175 responses remained for the final analysis. We computed the inter-annotator agreement using Cohen's k (=0.74). Here we describe the most interesting findings, but Appendix F shows additional results and details.

In BASE, where the answer was provided with no additional information, about 50% of **participants mentioned that they would have liked it if the voice changed with system certainty**. In CONF, around 30% of participants give this feedback.

For EXT-SENT in both modalities, EXT-LONG in the visual modality, and ABS, 10-35 % of **participants would like the level of detail to adapt to the model certainty**. Users would like to have more detail *only* when the model is not confident.

For EXT-LONG in the *spoken modality*, the feedback centered around length. 78% of participants mentioned that responses should be shorter, which aligns with the higher perceived length of the explanations in Figure 5. For the *visual modality*, 40 % of participants mention that highlighting some key items would have made it even easier and faster. Introducing highlights would improve the visual interface and would likely increase the differences in modality already observed.

Finally, for all explanation conditions, 20-45% of **participants said they would like to see explanations from multiple sources for an answer**, *e.g.*, from non-Wikipedia sources to help them better decide whether to trust the answer.

4.3 What Misleads Users?

To understand how explanations can mislead users, we analyzed questions where users frequently accepted incorrect predictions (false positives). A single annotator then followed a similar coding procedure to detect categories of such questions. We found that users were frequently misled on the same 30% of our study questions. Below we describe the two main categories:

Plausible explanations. We find 60-65% cases where an explanation does not confirm the predicted answer but makes it seem plausible, misleading users into accepting incorrect responses. This phenomenon is similar to prior work in psychology that has shown that people often fail to evaluate the accuracy of information when they have little prior knowledge and information seems plausible (Hinze et al., 2014).

> **Question:** *Who is the patron saint of adoptive parents?*
> **Response:** I am 37 percent confident that the answer is, **Saint Anthony of Padua**. I found the following evidence in a wikipedia passage titled, Anthony of Padua: Saint Anthony of Padua, born Fernando Martins de Bulhoes, also known as Anthony of Lisbon, was a portuguese catholic priest and friar of the Franciscan order.

In the example above, the model is incorrect (true answer is Saint William of Perth), but users were often misled to accept this answer because the evidence makes the prediction sound plausible.

Lexical overlap. The second most common mistake (from 30 to 35% of errors) that both *the model* and *the users* make is related to the lexical overlap between the question and the evidence. For

instance, in the example below, the evidence contains the correct answer (15 teams) but many users are misled by the phrase *"A total of 30 teams play in the National League."*

> **Question:** *How many teams are in the MLB national League?*
> **Response:** I am 60 percent confident that the answer is, **30**. I found the following evidence in a wikipedia passage titled, Major League Baseball: *A total of 30 teams play in the National League(NL)* and American League (AL) , with 15 teams in each league .

5 Discussion

5.1 Why Explanations Worked for ODQA

Unlike previous studies (Bansal et al., 2021; Chu et al., 2020; Hase and Bansal, 2020), we observed significant improvements from explanations over only communicating confidence. One reason for our positive results could be owing to the nature of ODQA i.e. unlike tasks such as sentiment classification, where humans may be able to solve the task without relying on explanations, ODQA requires satisfying a user's information need, which may take considerably longer without explanations; users *require* additional help to navigate through vast amounts of information. Another potential reason is, in ODQA, presenting a single good explanation can allow users to verify whether the prediction is correct. In contrast, in sentiment analysis, even if the explanation points to evidence for a positive sentiment ("the smell was delicious"), there is always a chance that another phrase ("but the taste made me puke") renders the net correct label as negative. It is worth noting that like previous works, not all of our explanation methods provide significant value (Figure 3); thus the success from showing explanations still cannot be taken for granted but should instead be measured using well-designed user studies.

5.2 Implications and Recommendations

Another interesting question is how can our findings inform future research in explainable NLP.

Develop modality-specific explanations. Our results showed that the best explanation varied across modalities, indicating that evaluating explanations on one modality (*e.g.*, visual UI) and deploying them on another (*e.g.*, voice assistant) can lead to sub-optimal deployment decisions. As a result, explanations should be optimized for and evaluated in the task and settings in which they will be deployed in-the-wild.

Further study abstractive explanations. Longer explanations helped improve error-detectability in the visual modality, but they hurt in the spoken case, perhaps because of the increased cognitive load. This may indicate a trade-off between *information content* of explanation and its *cognitive load*. While we hoped abstractive explanations would achieve an optimal balance, results showed that they did not improve end-performance. Perhaps because even though they were more concise, they still had high information density. Though, abstractive explanations showed some promise. For example, compared to longer explanations, they improved speed of error-detectability by 2.2 sec (discussed in Appendix, Table 2) and their length was rated as more satisfactory (Figure 5). Thus future work should explore whether benefits of abstractive increase when explaining multiple sources (*e.g.*, in (Yang et al., 2018)) or candidate answers.

Enable interactive explanations. To manage a balance between information content and cognitive load one may also use interactive explanations (Weld and Bansal, 2018), where the system presents a concise explanation and lets users request more details, *e.g.*additional evidence, sources, or candidate answers (Section 4.2). Another option is *adaptive* explanations, where the model switches explanation strategies based on its confidence (Bansal et al., 2021).

6 Conclusion

We conducted user studies to understand whether explanations from a state-of-the-art open-domain QA system help improve error-detectability for end-users. Our study showed that for ODQA, simple explanations based on evidence snippets can significantly improve error-detectability and beat strong baselines such as communicating model's confidence. We observed this for multiple modalities of interaction: spoken and visual modalities. However, results also indicated that not every explanation type is guaranteed to improve performance over confidence and the best explanation strategy may change with the modality, *e.g.*, due to differences in users' cognitive abilities across modalities. Thus, developers and researchers of explainable ODQA systems should not take the effectiveness of explanations for granted and should evaluate and tune them on the tasks and modalities where these models will be eventually deployed.

7 Ethical Impact Statement

Recent work has shown that explanations may increase blind trust in systems (Bansal et al., 2021). Deploying such explanations in the wild is ethically fraught, hence we should better evaluate explanations using human evaluation before deployment. Our study expands the knowledge in this direction and show that current explanation strategies can work, but they can still considerably mislead users to accept incorrect model predictions. We hope that our findings and recommendations will have a positive impact on how explainable NLP is developed and evaluated in future work; namely, through carefully designed user studies which inform us of the real-world utility of explanations. In terms of data collection, the study was approved by IRB and no sensitive or personally identifiable data was collected, and users were informed that their efforts would end in a research publication.

References

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-ai interaction. In *CHI*.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. pages 1–16.

Pierre Barrouillet, Sophie Bernardin, Sophie Portrat, Evie Vergauwe, and Valérie Camos. 2007. Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3):570.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464.

Gary Buck. 1991. The testing of listening comprehension: an introspective study1. *Language testing*, 8(1):67–91.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2018. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *ACL*.

Shi Feng and Jordan Boyd Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *IUI*, pages 229–239.

John Flowerdew, Michael H Long, Jack C Richards, et al. 1994. *Academic listening: Research perspectives*. Cambridge University Press.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. pages 1321–1330.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? pages 5540–5552.

Scott R Hinze, Daniel G Slaten, William S Horton, Ryan Jenkins, and David N Rapp. 2014. Pilgrims sailing the titanic: Plausibility effects on memory for misinformation. *Memory & Cognition*, 42(2):305–324.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. pages 6769–6781.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones,

Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering. *arXiv preprint arXiv:2009.06354*.

Wayne Leahy and John Sweller. 2016. Cognitive load theory and the effects of transient information on the modality effect. *Instructional Science*, 44(1):107–123.

John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019a. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019b. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

Randall J Lund. 1991. A comparison of second language listening and reading comprehension. *The modern language journal*, 75(2):196–204.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. pages 5783–5797.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Nobuko Osada. 2004. Listening comprehension research: A brief review of the past thirty years. *Dialogue*, 3(1):53–66.

Google PAIR. 2019. The People + AI Guidebook. https://pair.withgoogle.com/guidebook/. Accessed: 2021-1-28.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. pages 1938–1952.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. pages 5609–5626.

John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier.

Irene Thompson and Joan Rubin. 1996. Can strategy instruction improve listening comprehension? *Foreign Language Annals*, 29(3):331–342.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82. Citeseer.

Daniel S Weld and Gagan Bansal. 2018. Intelligible artificial intelligence. *ArXiv e-prints, March 2018*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. pages 2369–2380.

A Explanation Examples

In Table 1, we show an example of how the responses and explanations looked for each of the conditions. We also indicate in which modalities each explanation is shown in our experiments.

B Temperature Scaling

Temperature scaling (Guo et al., 2017), a multi-class extension of Platt Scaling (Platt et al., 1999), is a post-processing method applied on the logits of a neural network, before the softmax layer. It consists of learning a scalar parameter t, which decreases or increases confidence. t is used to rescale the logit vector z, which is input to softmax σ, so that the predicted probabilities are obtained by $\sigma(z/t)$, instead of $\sigma(z)$.

In our experiments, the model is set to pick from the top 100 solutions, however, in many cases the correct answer occurs within the top 10 items. For our purposes we calibrate the confidence scores of the top 10 outputs. We use the publicly available scripts provided by Guo et al. (2017).[6]

The model confidence before and after calibration can be seen in Figure 6.

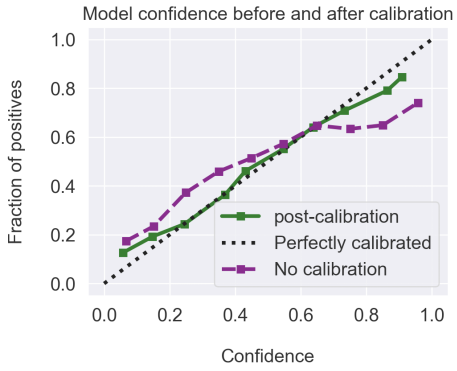

Figure 6: Confidence before and after calibration.

C Additional Preprocessing

Additional preprocessing to ascertain the quality of stimuli in each modality was required. Before sampling questions for the task, to ensure a high-quality and non-ambiguous experience for MTurk workers, we manually filter out several "problematic" questions:

- **Ambiguity in the question**: For various questions in NQ, multiple answers can exist. For example, the question: *when was King Kong released?*, does not specify which of the many King Kong movies or video games it refers to. These cases have been known to appear often in NQ (Min et al., 2020). We remove such questions from our subset.
- **The gold answer was incorrect**: Many examples in NQ are incorrectly annotated. As it is too expensive to re annotate these cases, we remove them.
- **Answer marked incorrect is actually correct** : We present both correct and incorrect questions to users. There are cases where the predicted answer is marked incorrect (not exact match) but is actually correct (a paraphrase). We manually verify that correct answers are paired with contexts which support the answer.
- **Correct answer but incorrect evidence**: The model sometimes, though not as often, chooses the correct answer but in the incorrect context. We discarded examples where the explanation was irrelevant to the question e.g. who plays Oscar in the office? *Oscar Nuñez, is a Cuban-American actor and comedian.*. In order to be able to make more general conclusions about whether explanations help in error-detectability, we restrict our questions to ones containing correct answers in the correct context.
- **Question and prediction do not match type**. We removed cases where the question asked for a certain type e.g. a date, and the prediction type did not match e.g. a location.

In the visual modality, to ensure readability, we fixed capitalizations. For the spoken modality, to ensure fluency and clarity, we manually (1) inserted punctuation to ensure more natural sounding pauses, and (2) changed abbreviations and symbols to a written out form e.g. *$ 3.5 billion* to *3.5 billion dollars*.

D Task Setup: Additional details

Platform and participant details. We conduct our experiments using Amazon Mechanical Turk[7]. We recruited 525 participants in total, with approval ratings greater than 95 % and had a maximum of 8

[6] https://github.com/gpleiss/temperature_scaling

[7] https://www.mturk.com/

EXPLANATION TYPE	RESPONSE+EXPLANATION	MODALITY
BASE	The answer is, **two**.	Spoken
CONF	I am 41 percent confident that the answer is, **two**.	Spoken
ABS	I am 41 percent confident that the answer is, **two**. I summarized evidence from a wikipedia passage titled, Marco Polo (TV series). Netflix cancelled the show after two seasons, as it had resulted in a 200 million dollar loss.	Spoken
EXT-SENT	I am 41 percent confident that the answer is, **two**. I found the following evidence in a wikipedia passage titled, Marco Polo (TV series). On December 12, 2016, Netflix announced they had canceled "Marco Polo" after two seasons.	Spoken/Visual
EXT-LONG	I am 41 percent confident that the answer is, **two**. I found the following evidence in a wikipedia passage titled, Marco Polo (TV series). On December 12, 2016, Netflix announced they had canceled "Marco Polo " after two seasons. Sources told "The Hollywood Reporter" that the series' two seasons resulted in a 200 million dollar loss for Netflix , and the decision to cancel the series was jointly taken by Netflix and the Weinstein Company. Luthi portrays Ling Ling in season 1, Chew in season 2. The series was originally developed at starz, which had picked up the series in January 2012.	Spoken/Visual

Table 1: **Explanation examples:** Example of how system responses looked for each explanation type and baseline, for the question *How many seasons of Marco Polo are there?*

days for approval of responses in order to minimize the amount of spamming.

We use a random sample of 120 questions from our dataset which remains the same across all conditions. In order to keep each session per participant at a reasonable time and ensure the quality of the data wouldn't be affected by workers becoming exhausted, we opted for three fixed batches of 40 questions, all split as 50 % correct and 50 % incorrect. Workers could only participate once (only one batch in one condition). Participants took around from 35-45 minutes to complete the HITs, but were given up to 70 minutes to complete.

We monitored if their screen went out of focus, to ensure that participants did not cheat. We ensured that we had 25 user annotations per question. When analyzing the data, we remove the first 4 questions of each batch, as it may take participants a few tries before getting used to the interface. In the end, we collect about 21,000 test instances.

Task Instructions. Imagine asking Norby a question and Norby responds with an answer. Norby's answer can be correct or wrong. If you believe Norby's answer is correct, you can accept the answer. If you believe it is wrong, you can reject it. If the answer is actually correct and you accept it, you will earn a bonus of $0.15. But, if the answer is wrong, and you accept it, you will lose $0.15 from your bonus. If you reject the answer, your bonus is not affected. (Don't worry, the bonus is extra! Even if it shows negative during the experiment, in the end the minimum bonus is 0). In total you will see 40 questions in this HIT (you will only be allowed to participate once) and the task will take about 40 to 45 minutes. You can be compensated a maximum of $13.50 for about 40-45 minutes of work. Some things to note:

1. You must listen to the audio before the options become available.
2. If you make it to the end there is a submit button there, however, in case of an emergency you can hit the quit early button above and you will get rewarded for the answers you provided.
3. You can play the audio as many times as you need but as soon as you click a choice you will be directed to the next item.
4. **IMPORTANT!!** Please do not look up questions in any search engine. We will monitor when the screen goes out of focus, so please keep the screen on focus or you might risk being rejected.
5. Finally, please do not discuss answers in forums; that will invalidate our results.

E Post-task Survey

1. I found the CLARITY of Norby's voice to be:

   (a) Excellent **(b)** Good **(c)** Fair **(d)** Poor **(e)** Very Poor

2. I found Norby's responses to be HELPFUL when deciding to Accept or Reject:

(a) Strongly Agree **(b)** Agree **(c)** Undecided **(d)** Disagree **(e)** Strongly Disagree

Can you give a few more details about your answer?

3. I found the LENGTH of Norby's responses to be:

   (a) Too Long **(b)** Long **(c)** Just right **(d)** Short **(e)** Too short

4. No AI is perfect and Norby is no exception. We are interested in helping Norby provide responses that can help users to determine whether to trust it or not (to accept or reject, just as you have done in this experiment). From your interaction with Norby, **do you have any additional feedback on what it can improve?**

F Results

Reward. Cummulative reward is the total dollar reward in bonuses earned by a worker based on the payoff described earlier. Note that, unlike accuracy, the payoff matrix is not symmetric wrt. user decision and correctness of predictions. We compute the differences in overall reward for each condition and observe the same trends as we discussed for accuracy. More specifically, all explanation conditions improve the final user reward, with EXT-SENT performing best in the spoken modality and EXT-LONG performing best overall. These differences are shown in Figure 7.

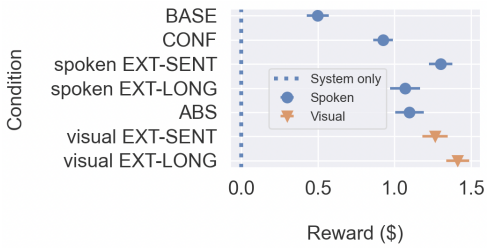

Figure 7: **Reward**: The scores presented here are out of $ 2.70. Although all explanations are better than CONF, the explanations leading to the highest rewards change across modalities.

Time differences. We measured the time (in seconds) that it took participants to complete each question. In Table 2 we present the median times averaged over all workers per condition. We also include an adjusted time, subtracting the length of the audio, in order to measure decision time.

CONDITION	SEC/QUESTION	ADJUSTED
SPOKEN MODALITY		
BASE	10.2 ± 1.6	8.3 ± 1.6
CONF	9.4 ± 1.5	6.0 ± 1.5
ABS	24.4 ± 1.5	7.0 ± 1.4
EXT-LONG	44.9 ± 1.6	9.2 ± 1.6
EXT-SENT	24.3 ± 1.7	7.6 ± 1.7
VISUAL MODALITY		
EXT-LONG	16.1 ± 1.7	-
EXT-SENT	10.4 ± 1.1	-

Table 2: Time differences across modalities. Time differences in the right column have been adjusted by removing the duration of the audio files. We observe that with additional information, users can make faster decisions than the BASELINE condition.

Voice quality. To verify that the quality of the text-to-speech tool that we employed did not negatively affect our experiments, we asked users to rate the clarity of the assistant's voice as *very poor, poor, fair, good*, or *excellent*. Around 90 % rated the voice as good or excellent. These results are shown in Figure 8.

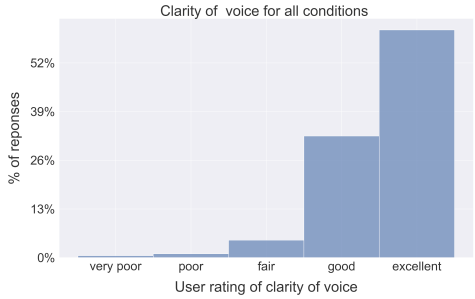

Figure 8: **Voice clarity:** Most participants found the voice of the assistant to be good or excellent.

Helpfulness. Participants were asked whether the responses helped them in their decision making. Their responses showed that CONF and all explanation conditions were perceived as helpful by at least 80% of participants, with no real differences among them except for EXT-LONG in the visual modality (which is perceived helpful by close to 90% of users). Interestingly, 50% of participants indicated BASE to be helpful. In contrast, our results in Figure 3 show that different explanations actually differ in their eventual helpfulness. These results suggest that subjective measures can sometimes correlate with actual performance when the differences are large, but for the most-part and smaller differences, the result from subjective rating can be unreliable. These findings align with prior observa-

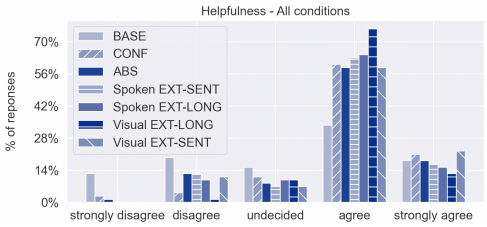

Figure 9: **Helpfulness:** Participants indicated how helpful responses were. These results reflect the large differences we see in performance (BASE vs the rest of the settings), but are not able to capture the more subtle differences among explanation strategies and CONF.

tion made (Buçinca et al., 2020) that showed that evaluating explanations on proxy metrics can lead to incorrect conclusions. These findings are shown in Figure 9.

User feedback. Users provided free-form written feedback on possible ways to improve the system. The prompt they saw was: *do you have any additional feedback on what the system can improve?* After converging on a final set of codes, two annotators coded up about 400 responses across all conditions. The codes and their descriptions can be found in Table 3. The codes are *not* mutually exclusive.

CODE	DESCRIPTION	CATEGORY
len-conciseness	users wish explanation was shorter	improvement on length
len-expand	users wish explanation was shorter	
adapt-detail	users wish details adapted with confidence	adaptability feature
adapt-voice	users wish voice adapted to confidence	
pres-change-confidence	users wish confidence would be communicated differently e.g. the answer is probably....	improve presentation
pres-highlighting	users wish important facts would be highlighted	
need-more-sources	users wish more source were provided	need additional info
need-confidence	users wish confidence was provided	
need-source	users wished a source was provided	
need-explanation	users wish an explanation would be provided	
need-link	users wish a link was provided	
need-multiple-answers	users wish more than 1 answer was provided	

Table 3: The codes used to uncover areas of improvement from the post-experimental user feedback.

We found that many users across most conditions, would like **adaptability features** added. Additionally, we found that participants would like to be provided with multiple sources which converge on the answer. We also observe that for spoken conditions, **improvements on length** are mentioned more often. The full distribution of codes across conditions is shown in Table 4.

CONDITION	CODE	% PARTICIPANTS
BASE	adapt-voice	50
	need-confidence	36
	need-explanation	25
	need-source	17
CONF	need-explanation	38
	adapt-voice	29
	pres-change-confidence	14
	adapt-detail	10
	need-multiple-answers	10
	need-link	5
Spoken EXT-SENT	need-more-sources	44
	adapt-detail	28
	len-conciseness	22
	need-multiple-answers	17
	need-link	11
	len-expand	11
	pres-change-confidence	6
Spoken EXT-LONG	len-conciseness	78
	need-more-sources	15
	pres-change-confidence	4
ABS	len-conciseness	52
	need-more-sources	22
	adapt-detail	22
	pres-change-confidence	13
	need-multiple-answers	4
Visual EXT-SENT	need-more-sources	33
	adapt-detail	33
	len-expand	27
	need-multiple-answers	7
Visual EXT-LONG	pres-highlighting	40
	need-more-sources	33
	adapt-detail	10
	need-link	10
	pres-change-confidence	7

Table 4: Distribution of codes across all conditions. Codes are **not** mutually exclusive.