Decoding the Metrics Maze: Navigating the Landscape of Conversational Question Answering System Evaluation in Procedural Tasks

Alexander Frummet and David Elsweiler
University of Regensburg
{alexander.frummet, david.elsweiler}@ur.de

Abstract

Conversational systems are widely used for various tasks, from answering general questions to domain-specific procedural tasks, such as cooking. While the effectiveness of metrics for evaluating general question answering (QA) tasks has been extensively studied, the evaluation of procedural QA remains a challenge as we do not know what answer types users prefer in such tasks. Existing studies on metrics evaluation often focus on general QA tasks and typically limit assessments to one answer type, such as short, SQuAD-like responses or longer passages. This research aims to achieve two objectives. Firstly, it seeks to identify the desired traits of conversational QA systems in procedural tasks, particularly in the context of cooking (RQ1). Second, it assesses how commonly used conversational QA metrics align with these traits and perform across various categories of correct and incorrect answers (RQ2). Our findings reveal that users generally favour concise conversational responses, except in time-sensitive scenarios where brief, clear answers hold more value (e.g. when heating in oil). While metrics effectively identify inaccuracies in short responses, several commonly employed metrics tend to assign higher scores to incorrect conversational answers when compared to correct ones. We provide a selection of metrics that reliably detect correct and incorrect information in short and conversational answers.

Keywords: metrics, conversational search, question answering

1. Introduction

Conversational systems are frequently used for a variety of tasks, such as setting timers, getting the weather forecast for the day, or retrieving factual information from the web. For such general question answering (QA) tasks, users can ask questions such as "What is the capital of Germany?". Responses from conversational agents can vary, from short, concise answers, e.g., "Berlin", to more conversational responses, such as "The capital of Germany is Berlin." However, the accuracy of these different answer types cannot be guaranteed. Therefore, there is a crucial need for reliable metrics to evaluate the effectiveness of such conversational systems. For general QA tasks, many studies have explored the effectiveness of commonly used metrics, spanning *word overlap-based metrics*, e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), F1 Score, and Exact Match, measuring the degree of overlap between words in the ground truth answer and those generated by a QA model, to *embedding-based metrics*, such as Semantic Answer Similarity (SAS) (Risch et al., 2021) and BERTScore (Zhang et al., 2020), which measure semantically equivalent responses.

To determine the utility of these metrics, studies commonly examine their alignment with user preferences. For instance, numerous studies have examined the efficacy of word overlap-based metrics, and have concluded that METEOR exhibits the strongest correlation with human evaluations of QA model outputs, in comparison to metrics, such as ROUGE and BLEU (Blagec et al., 2020; Nema and Khapra, 2018; Chen et al., 2019). Others have investigated the performance of embedding-based metrics, including SAS and BERTScore. Chen et al. (2019) found that while BERTScore is superior at capturing semantic information, it does not correlate as strongly with human assessments of how closely a model prediction matches the ground truth. Additionally, SAS and BERTScore struggle with spatial awareness, numbers, and conversions (Mustafazade et al., 2022). In one particularly thorough analysis by Liu et al. (2021), conversational search evaluation metrics were evaluated from three different perspectives: reliability, fidelity, and intuitiveness. Based on their analysis, METEOR and BERTScore were determined to be the most reliable, whereas METEOR and BLEU were found to be the most intuitive.

These studies primarily assess metrics for general, domain-agnostic QA using open-domain datasets such as MSDialog or Wizard of Wikipedia (Liu et al., 2021). However, for domain-specific procedural QA tasks, such as cooking and DIY, where conversational agents are increasingly popular[1], such meta-evaluations fall short. While these domains share similarities with general QA, we lack

[1] see cooking/DIY focus in Alexa Task-Bot challenge: https://www.amazon.science/alexa-prize/taskbot-challenge

Proceedings of the HumEval @LREC-COLING 2024, pages 81–90
Turin, Italy. May 21st, 2024.

a deep understanding of the unique challenges in procedural QA conversations, making evaluation difficult. For example, consider the recipe step "Add apples, oranges, and sugar to a large pitcher and muddle with a muddler or wooden spoon for 45 seconds". When asked, "How much sugar do I need to add?" the answer can be as concise as "50g" or more conversational, e.g., "You need to add 50g of sugar". These responses differ in the level of context provided about the cooking process. "50g" lacks context, while the other response specifies that "50g" refers to sugar. In-situ studies with voice assistants have shown that users have a preference for agents that provide clarifications (Luger and Sellen, 2016) and in the cooking domain, an analysis of human-human dialogues revealed that people often seek reassurances regarding the answers they receive from an assistant (Frummet et al., 2022).

While these studies hint that answers clarifying context may be preferred, we do not yet know what constitutes a good answer in procedural tasks, such as cooking. Since we lack knowledge of user answer preferences, selecting a reliable metric to evaluate answer correctness in procedural QA tasks remains challenging.

This differs from "general" non-procedural QA tasks where studies evaluate metric effectiveness using responses of the same type, whether short, SQuAD-like answers (Mustafazade et al., 2022; Nema and Khapra, 2018; Bulian et al., 2022), conversational, sentence-length answers (Shi et al., 2023; Siblini et al., 2021) or long, paragraph-like answers (Xu et al., 2023). While these studies provide insights into reliable metrics for non-procedural scenarios, it is unclear if these findings extend to procedural tasks. It is uncertain if the metrics used there remain reliable, valid, and applicable when various answer formulations are possible, as is the case in procedural tasks.

To assess conversational agents effectively for procedural tasks, we need to understand 1) which answers are preferred by users and 2) whether the metrics traditionally employed in conversational QA yield reliable results for evaluating conversational systems in procedural tasks.

Having reliable evaluation results is crucial for the success of most procedural tasks, as having the correct information, such as quantities and next steps, is vital to successfully completing the associated task (Frummet et al., 2022). Just as with human-generated responses, multiple answer formulations are possible, and metrics must account for these variations. Additionally, they must be sensitive to incorrect aspects of answers, given that large language models sometimes "just make stuff up" (Shah and Bender, 2022). Users tend to trust these models because they mimic human language (Araujo, 2018; Dinan et al., 2021).

This study aims to achieve two objectives within the context of procedural assistance tasks. Firstly, it seeks to identify the desirable traits of QA systems for humans (RQ1). Secondly, it aims to analyse how commonly used metrics in conversational QA reflect these traits and vary for different categories of correct and incorrect answers (RQ2).

2. Methodology

This section outlines the methods and resources used to address the research questions, including the dataset used, the various answer types and metrics evaluated, and a user study to complement our system-based analyses.

2.1. Dataset

In this paper, we target cooking-related procedural assistance tasks. Existing datasets, such as CookDial (Jiang et al., 2022) and Wizard of Tasks (Choi et al., 2022), require post-processing to meet our needs. For example, Wizard of Tasks lacks grounding for conversational answers to specific parts of the recipe, making it challenging to evaluate different answer types within the recipe context. To address these limitations, we have created a new dataset tailored to our study's requirements. As a basis for our experiments, we use 298 randomly selected questions and answers [2] from a conversational cooking QA dataset (QookA) (Frummet and Elsweiler, 2024). The full dataset contains 1268 pairs of question-answer, where the questions are expressed in written natural language after being transcribed from spoken questions gathered from 95 participants ($M_{age} = 35$, 73% female, 26% male, 1% diverse) who followed a recipe from SeriousEats[3] in a simulated cooking scenario. Each question maps to an information need type from the taxonomy presented in Frummet et al. (2022) and has an associated answer derived from the appropriate recipe. Examples can be seen in the first two columns of Table 1.

2.2. Answer Formulations

Our study examines how metrics relate to various styles of answer. To this end, we devised three distinct answer types, each of which can be considered as "correct". First, *Classic QA* answers adhere to the extractive, span-based SQuAD format (Rajpurkar et al., 2018) and contain only the relevant information that is needed to answer a question.

[2] These 298 selected samples are available here: https://osf.io/6g5ud/?view_only=7f4e6040bff642cb8fc2638fe3a7295b

[3] https://www.seriouseats.com/

QookA Dataset		Classic QA	Conv. QA			Passage
Question	Answer		HA1	HA2	GPT-3	
What do I add after the mustard and cheese?	eggs.	eggs.	After mustard and cheese are combined, you need to add eggs one at a time.	After adding the mustard and cheese, you should add the eggs, one at a time.	After adding the mustard and cheese, beat the mixture with a wooden spoon until homogeneous. Then add eggs one at a time, ...	To Finish by Hand: Remove pot from heat. Add mustard and cheese and beat with wooden spoon until ...
How much fat should I reserve?	1 cup.	1 cup.	After skimming off the fat, 1 cup should be left behind.	According to the recipe you should reserve about 1 cup of the solid fat	Reserve 1 cup of fat when skimming the sauce.	Carefully skim off most of the fat, leaving behind about 1 cup total. (For more precise measurement ...
How long will it take for the coconut fat to separate?	about 4 minutes.	about 4 minutes.	It will take about 4 minutes for the coconut fat to separate.	Until coconut fat separates and solids start to sizzle, about 4 minutes.	It will take approximately 4 minutes for the coconut fat to separate when heated over medium-high heat and stirred frequently.	Open 1 can of coconut milk and carefully spoon off about 3 tablespoons of the solid fat from the surface and place it in a large Dutch oven...

Table 1: Example entries from the dataset used. The left-most two columns are from the QookA dataset. The remaining columns represent the answer types studied in this work. HA1 and HA2 denote the two human annotators.

Second, we investigate answers that follow the style of *conversational QA* datasets, such as QuAC (Choi et al., 2018), CoQA (Reddy et al., 2019) and Wizard of Tasks (Choi et al., 2022). To establish these answers, two human annotators were asked to formulate what they would consider to be "ideal" responses given a question and the classic QA answer, as well as the step text from the QookA dataset. They were free to formulate the answer in any way they wished, and did so independently. As an automated comparison, we provided the same instructions and information to GPT-3 DaVinci-3 as a prompt:

> Question: ‹*Question*›
> Answer: ‹*Answer*›
> Context: ‹*Recipe Step Text*›
> Rephrased Answer:

The authors checked these manually to establish that they were still "correct".

Last, we investigate passages that contain the answer and the surrounding context. These were attained by identifying the recipe step that included the pertinent information. A passage-based answer is appropriate since answers of this type are evaluated in conversational information retrieval (IR) assessment frameworks, such as CAsT (Dalton et al., 2020), and is a plausible information unit to present to users since past research has revealed that cooking assistant users value contextually embedded answers (Frummet et al., 2019, 2022). Examples answers of all three types can be found in Table 1.

2.3. Incorrect answers

To obtain incorrect versions of the same classes of answer we leveraged GPT-3 to generate responses that were factually inaccurate using the following prompt:

Question: <*Question*>
Context: <*Recipe Step Text*>
Correct Answer: <*Correct Answer (either ClassicQA/Conv. QA)*>
Factually wrong answer:

GPT-3 was provided with the question, the correct answer (either Classic QA or Conversational QA form) and the corresponding recipe step. In the case of Conversational QA, we randomly chose one of the two human annotators and used their answer as the ground truth. As an example, for the second row in Table 1, the incorrect Classic QA answer was "a tablespoon" and the incorrect Conversational QA answer was "You should reserve 2 cups of fat".

To derive incorrect passage formulations, we opted to randomly select another passage from our dataset to replicate the potential for an improperly retrieved passage that may arise during a TREC CAsT experiment.

2.4. Metrics studied

We study the most commonly applied QA metrics of the classes outlined in the introduction: Word-overlap based and embbeding-based metrics. Specifically, we evaluate ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), Exact Match and F1 for the word-overlap domain. From the embedding domain, we study Semantic Answer Similarity (Risch et al., 2021) and BERTScore F1 (Zhang et al., 2020).

To calculate the metric, we selected one of our answers as the prediction and compared it against all other answers. For example, if testing a classic QA answer (i.e., prediction) then the equivalent HA1, HA2, GPT-3, and Passage answers served as the references/ground truth to compute the metric values. We utilised Huggingface's evaluation library [4] to calculate the ROUGE, BERTScore, METEOR, BLEU, and Exact Match scores, while FARM's evaluation library[5] was used to compute F1 scores. Additionally, we customised the script from the SCAI-QReCC-22 shared task to determine the Semantic Answer Similarity (SAS).

2.5. User perception of correct answers

We conducted an online experiment with a within-groups design to determine how participants perceive answers of different types. In this experiment, each participant evaluated the appropriateness of five separate answers (one for each type, presented in a random order) for questions chosen from our dataset. As illustrated in Figure 1, ratings were provided on a Likert scale ranging from 1 to 5. To provide the necessary context, each answer was accompanied by the corresponding question and recipe step. To learn why our participants preferred certain answer types over others, we requested that they provide an explanation for their ratings.

The experiment was designed to resemble an interaction with a conversational assistant in a kitchen. To this end, we employed Google's Text-to-Speech API to convert the answers to audio files, which participants then listened to. Attention checks were used to confirm that participants had indeed listened to the answers. As an attention check, one of the answer audio files illustrated in Figure 1 contained the following instruction: "Please click the left circle and write the answer to six multiplied by four into the text field below."

Study participants were recruited via Prolific. According to the power analysis performed using G*Power (Faul et al., 2007), a total of 32 individuals were needed to achieve the required statistical power for conducting an ANOVA test with repeated measures[6]. All participants were recruited from the UK as we selected a British accent for the audio answer files. 53.13% of our participants were female, 46.88% male, with most being between 25 and 34 years old (37.5%). Since our experiment is in the cooking domain, we wanted to know how much people enjoy cooking on a scale from 1 to 5. The people in our study generally enjoy cooking ($M = 3.66, SD = 1.32$).

3. Results

This section presents our findings. Sections 3.1 and 3.2 provide an insight into the human perspective (RQ1) by examining the answers provided by human annotators and reporting the outcomes of our online study. Meanwhile, Sections 3.4 and 3.5 shed light on the metrics' ability to differentiate between correct and incorrect answers, and how they reflect the characteristics users desire (RQ2).

3.1. Evaluating Human Provided Answers

In a first step, we analyse the "ideal" answers provided by two human annotators to understand the conveyed information and the methods of communication. An initial observation suggests that these answers are considerably lengthier than Classic QA answers (which have a mean word count of $\bar{x} = 3.69$), but shorter than passage answers ($\bar{x} = 71.67$). On average, Human annotator 1 (HA1)

[4] https://huggingface.co/evaluate-metric

[5] https://farm.deepset.ai/_modules/farm/eval.html

[6] desired Power: 0.8; effect size $f = 0.25$; significance threshold $\alpha = 0.05$; num. of groups: 1; num. of measurements: 5; Corr. among rep. measures: 0.2

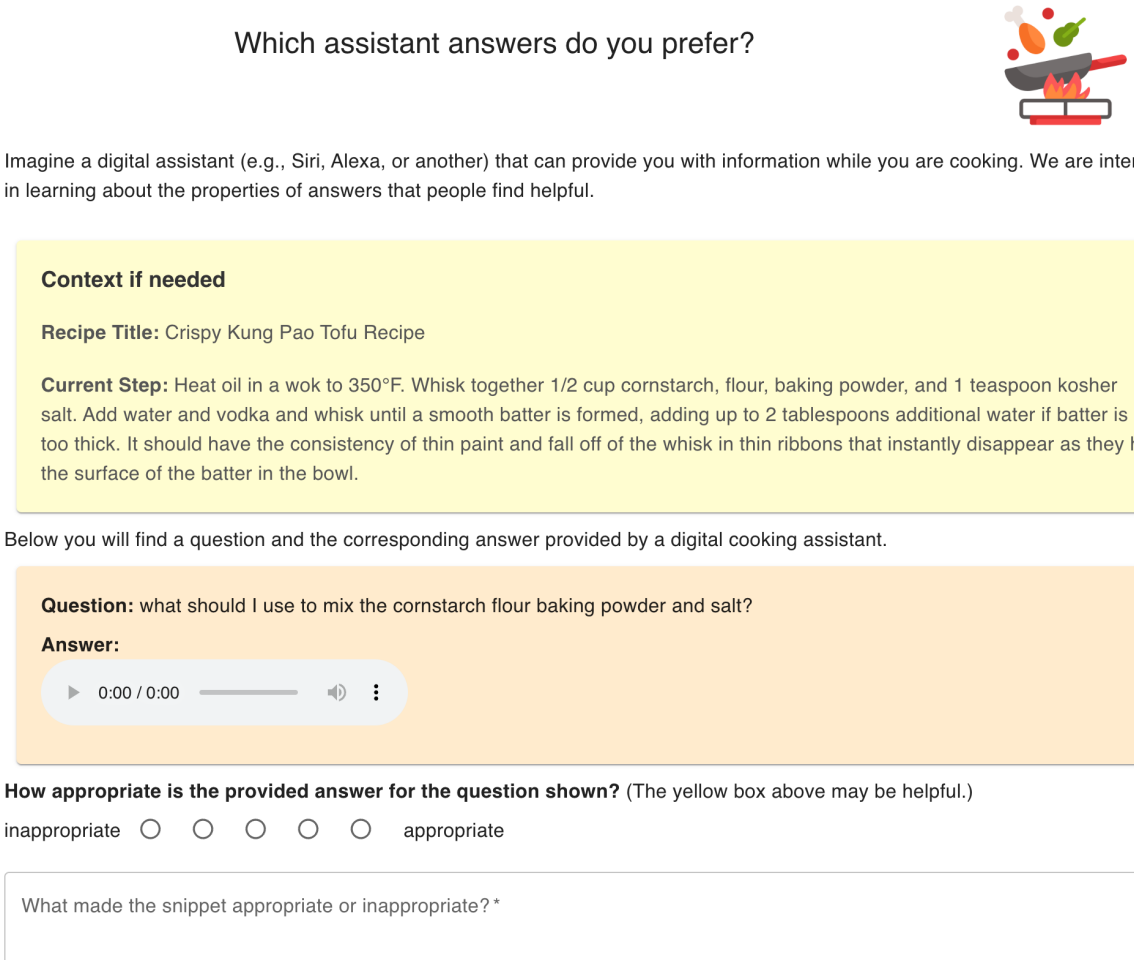

Figure 1: Screenshot of our online experiment tool.

provided responses containing $\bar{x} = 12.40$ words, while Human annotator 2 (HA2) had answers averaging $\bar{x} = 9.78$ words.

Examining the answers more closely reveals that both human annotators chose their formulations carefully. Many of the answers were phrased, such that the user would be reassured that the question was correctly understood. The first entry in Table 1 exemplifies that both HA1 and HA2 embed their answer in the conversational and step context. Both answers start with "After adding the mustard and cheese...". This linguistic pattern is used to reassure the user that their question has been understood correctly, a pattern observed in naturalistic cooking QA investigations (Frummet et al., 2022). Rather than simply providing times, e.g., "about 4 minutes", as the classic QA answer does, both HAs provide contextual information to provide cues to aid understanding e.g., "It will take about 4 minutes for the coconut fat to separate". Such techniques were common as is reflected in the jaccard similarity between HAs and questions (20% HA1, 11.58% HA2) and HAs and passage (13.15% HA1, 12.12% HA2).

3.2. Establishing User Answer Preferences

Participants rated conversational QA answers higher (GPT-3: $M = 4.53, SD = .67$, HA1: $M = 4.34, SD = 1.07$, HA2: $M = 4.22, SD = 1.10$) than classic QA answers ($M = 3.97, SD = 1.36$) and passage answers ($M = 3.00, SD = 1.34$). An ANOVA with repeated measures showed a significant difference ($F(4, 124) = 9.14, p < .001$). Bonferroni-adjusted post-hoc t-tests revealed that passage answers were rated significantly lower than GPT-3 answers ($p < .001$), HA1 answers

Answer Type	ROUGE1	ROUGE2	ROUGEL	ROUGELSum	BERTScore	SAS	BLEU	METEOR	EM	F1
Cor. Classic QA	44.03	30.64	38.67	38.67	90.36	68.53	21.89	58.63	0.34	43.29
Inc. Classic QA	13.50	1.88	13.35	13.36	92.59	42.83	1.80	26.25	0.00	12.62
Cor. Conv. QA	58.37	39.95	54.68	54.68	93.41	78.85	34.63	61.99	1.51	57.48
Inc. Conv. QA	59.29	43.87	58.39	58.39	94.42	56.67	38.86	56.52	0.34	56.66
Cor. Passage	45.24	25.49	45.02	45.02	90.54	67.45	15.32	32.87	2.01	49.16
Inc. Passage	22.01	4.66	15.56	15.56	84.67	44.43	2.63	21.07	1.34	19.76

Table 2: Metric Experiment results (in percent) grouped by Answer Type. The results for Conv. QA answer types are the mean scores for HA1, HA2 and GPT-3, which show little variation.

($p < .001$) and HA2 answers ($p < .01$).

The results indicate a preference for conversational QA answers compared to other types, which is supported by the justifications provided with the ratings.

3.3. Understanding User Answer Preferences

To gain insights into participants' preferences for specific answers, we conducted a qualitative analysis of the explanations they provided for their ratings.

3.3.1. Passage

In line with the ratings discussed in Section 3.2, the majority of feedback related to passage-style answers was negative (84.38%). Participants often criticised these answers for being overly lengthy, containing "too much information", and lacking a "direct" and "specific" response. Consequently, they found these responses "confusing" and "unhelpful". However, some participants did find them "totally appropriate" without providing further elaboration.

3.3.2. Classic QA

Conversely, explanations for classic QA responses were more balanced, with 56.25% of the feedback being positive and 43.75% negative. Participants appreciated the brevity and clarity of these responses, finding them "short", "to the point", "clear", and "concise". Some mentioned that they particularly liked these answers in situations requiring "quick" reactions, for example, when they are "in the midst of heating oil". However, in contexts not demanding immediate responses, participants critiqued the lack of detail (e.g., "relevant but too brief", "could be more exact and informative", "I think the AI should be clearer") and called for more contextual information to fully comprehend the answer within the cooking process. For instance, a participant felt that from the given answer it was "unclear when to add eggs". Another individual said: "It correctly said stand mixer but could have added that it needs a paddle attachment". Others suggested providing "extra information", such as instructions on how "to score with a knife".

3.3.3. Conversational QA

Conversational answers generated by HA1, HA2, and GPT-3 received the most positive feedback, with an average of 73.94% of answers rated positively and 21.88% rated negatively. Participants favored these responses for being straightforward, clear, and appropriately detailed. Unlike classic QA, conversational answers provided "a good amount of detail" and offered extra context that "helps to clarify what part of the recipe it is referring to at the same time as getting the answer". Participants noted that these answers helped them plan ahead in the cooking process, (e.g., "The answer told me to put the ingredients in a bowl, however it also went beyond and spoke about the type of mixer.", "useful to know when to sprinkle the parmesan", "vocalised in chronological order the steps necessary to complete this section of the recipe").

The few negative comments suggested some answers could be shorter for conciseness (e.g., "'In a large skillet' would've been enough and more concise").

3.3.4. Summary

In conclusion, our analysis of user preferences for different answer types reveals the following key insights:

- **Passage**: Users predominantly criticise lengthy, unclear passage-style answers, often finding them confusing and unhelpful.
- **Classic QA**: Classic QA answers are appreciated for their brevity, particularly in situations requiring quick responses, but some users call for more detail to fully understand the context.
- **Conversational QA**: Conversational answers generated the most positive feedback due to their clarity, detail, and suitability for planning the cooking process. However, a few users suggested that some responses could

be made more concise for improved user experience.

3.4. Understanding variance in metric scores across type of correct answer

The effectiveness of the metrics were assessed by grouping the results by type of answer (Classic QA, Conversational QA, Passage). As indicated in Table 2, the exact match metric yielded extremely poor results across all conditions (i.e., $< 2.5\%$). Both the conventional machine translation metrics, BLEU, ROUGE, and METEOR, and the commonly used F-Measure also achieved low scores for correct answers. Of these, however, METEOR provided the highest scores, which aligns with past research indicating that METEOR is the most robust among the common Machine Translation metrics (see, for example, Chen et al. (2019); Blagec et al. (2020); Nema and Khapra (2018)). The embedding-based metrics BERTScore and SAS, in contrast, yield much higher scores than all other metrics for correct answers with BERTScore providing the highest of all. Contrasting the metric scores across the three classes of answer reveals that all metrics provide higher scores for the user-preferred conversational answers.

3.5. Understanding variance in metric scores across type of incorrect answer

Our results show that all metrics provide lower scores when we provide factually incorrect classic QA and passage style answers. This is an expected and desired outcome. However, as illustrated in Table 2, most metrics yield higher scores when an incorrect answer in the conversational QA format is provided. BERTScore even yields the highest scores overall ($> 94\%$). This indicates that BERTScore is not a suitable metric for dealing with incorrect information. The only exceptions are the SAS and METEOR metrics which decrease for incorrect answers. The decreasing METEOR metric score, again, evidences its robustness as pointed out in Chen et al. (2019) and Liu et al. (2021).

We performed a Kruskal-Wallis test to determine if the differences in ConvQA answer type are statistically significant. The different metrics results served as dependent variables. Our independent variable is correct/incorrect answer. A Posthoc Dunn's test with bonferroni-adjusted p-values revealed that ROUGE2/L/LSum ($p < 0.05$) and BERTScore ($p < 0.001$) achieved significant higher scores for incorrect answers. METEOR ($p < 0.01$) and SAS ($p < 0.001$) yielded significantly higher results for correct answers.

4. Discussion

In this work, we have tried to "decode the metrics maze" by evaluating popular question answering metrics in the light of two research questions.

4.1. User Preferences for Answer Types

In RQ1, we studied the desired traits of procedural QA systems for users and whether all types of correct answers were equally preferred. Our findings from the qualitative analysis provide some hints that user preferences for answer types may vary depending on the specific task they perform. In situations requiring quick responses, such as boiling or heating something in oil, users tend to favour short and to the point answers. During most stages of the cooking process, however, users prefer concise, contextual conversational answers helping them to plan the cooking process. It's evident that not all correct answers are viewed equally by users, aligning with observations from previous naturalistic research on conversational cooking QA scenarios (Frummet et al., 2019, 2022, 2024). Consequently, for procedural conversational question answering, such as cooking, metrics should reliably distinguish between correct and incorrect answers for both short (=Classic QA) and conversational responses.

4.2. Metrics Performance in Procedural Tasks

In RQ2, we examined how commonly used metrics in conversational QA align with user preferences from RQ1 and their performance across various answer categories in a cooking QA scenario. Users generally favour conversational answers but appreciate short, concise responses in specific situations. To meet this need, metrics should effectively distinguish between correct and incorrect answers in both response styles.

Our findings, presented in sections 3.4 and 3.5, revealed that, except for SAS, METEOR and F1, all metrics exhibited higher scores for incorrect answers compared to correct ones in the case of conversational responses. This trend was consistent for metrics commonly employed in machine learning tasks, including ROUGE, BLEU, and BERTScore, with METEOR being the only exception. Hanna and Bojar (2021) provided an explanation for the generally high BERTScore values stating that "BERTScore fails to assign low scores when a bad candidate sentence has high lexical overlap with the reference in terms of content words" (Hanna and Bojar, 2021, p. 515). This phenomenon is attributed to the lexical similarity between correct and incorrect answers in our study, where only a few words differ. Both Blagec et al.

(2020) and Chen et al. (2019) have argued that machine translation metrics such as ROUGE and BLEU are unsuitable for question answering tasks as they struggle to identify incorrect information due to their n-gram-based approach (Blagec et al., 2020). However, Blagec et al. (2020) noted that METEOR is good at capturing semantic differences.

Our results suggest that commonly used conversational QA metrics may not accurately evaluate the correctness of information. Instead, a comprehensive suite of metrics is needed to assess answer accuracy. For short answers, F1, METEOR and SAS are dependable choices for distinguishing between correct and incorrect information. In the case of conversational responses, METEOR and SAS are better choices than F1 in capturing these distinctions more effectively. The choice of metrics should align with the answer type preferred by the user in the current cooking task (see Section 4.1). For short answers, the appropriate suite includes F1, METEOR, and SAS, while for conversational answers, it consists of METEOR and SAS.

5. Conclusion and Future Work

Currently, conversational assistant systems are being evaluated with diverse and, as we have shown here, potentially inappropriate metrics. We suggest using a suite of metrics to accurately assess the effectiveness of such systems. The choice of metrics should be tailored to the specific task and how users respond within that task.

Although we discovered hints that users tend to favour brief answers during time-sensitive situations (e.g., heating in oil) and conversational answers in most other cooking stages, the generalisability of these preferences remains uncertain. Future work should investigate these critical scenarios further. Moreover, forthcoming research should focus on developing robust evaluation methods for handling inaccuracies. While our study on a cooking QA dataset underscores this challenge, further confirmation using other datasets such as Wizard of Tasks (Choi et al., 2022) and CookDial (Jiang et al., 2022) is needed.

6. Bibliographical References

Theo Araujo. 2018. Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85:183–189.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald. 2020. A critical analysis of metrics used for measuring progress in artificial intelligence. *CoRR*, abs/2008.02577.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Jason Ingyu Choi, Saar Kuzi, Nikhita Vedula, Jie Zhao, Giuseppe Castellucci, Marcus Collins, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2022. Wizard of tasks: A novel conversational dataset for solving real-world tasks in conversational settings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3514–3529, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. Cast-19: A dataset for conversational information seeking. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Emily Dinan, Gavin Abercrombie, A Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021. Anticipating safety issues in e2e conversational ai: Framework and tooling. *arXiv preprint arXiv:2107.03451*.

Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*power 3: A flexible statistical power analysis program for the social,

behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191.

Alexander Frummet and David Elsweiler. 2024. Qooka: A cooking question answering dataset. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2024, Sheffield, United Kingdom, March 10-14, 2024*, pages 406–410. ACM.

Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2019. Detecting domain-specific information needs in conversational search dialogues. In *Proceedings of the 3rd Workshop on Natural Language for Artificial Intelligence co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AIIA 2019), Rende, Italy, November 19th-22nd, 2019*, volume 2521 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Alexander Frummet, David Elsweiler, and Bernd Ludwig. 2022. "what can I cook with these ingredients?" - understanding cooking-related information needs in conversational search. *ACM Trans. Inf. Syst.*, 40(4):81:1–81:32.

Alexander Frummet, Alessandro Speggiorin, David Elsweiler, Anton Leuski, and Jeff Dalton. 2024. Cooking with conversation: Enhancing user engagement and learning with a knowledge-enhancing assistant. *ACM Trans. Inf. Syst.* Just Accepted.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2022. Cookdial: A dataset for task-oriented dialogs grounded in procedural documents. *Applied Intelligence*, 53(4):4748–4766.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zeyang Liu, Ke Zhou, and Max L. Wilson. 2021. Meta-evaluation of conversational search evaluation metrics. *ACM Trans. Inf. Syst.*, 39(4).

Ewa Luger and Abigail Sellen. 2016. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5286–5297, New York, NY, USA. Association for Computing Machinery.

Farida Mustafazade, Peter Ebbinghaus, and Seth Darren. 2022. Evaluation of semantic answer similarity metrics. *International Journal on Natural Language Computing*, 11:15.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chirag Shah and Emily M Bender. 2022. Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 221–232.

Zhengliang Shi, Weiwei Sun, Shuo Zhang, Zhen Zhang, Pengjie Ren, and Zhaochun Ren. 2023. RADE: Reference-assisted dialogue evaluation for open-domain dialogue. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12856–12875, Toronto, Canada. Association for Computational Linguistics.

Wissam Siblini, Baris Sayil, and Yacine Kessaci. 2021. Towards a more robust evaluation for conversational question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing (Volume 2: Short Papers), pages 1028–1034, Online. Association for Computational Linguistics.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.