Risks and NLP Design: A Case Study on Procedural Document QA

Nikita Haduong[1] **Alice Gao**[1] **Noah A. Smith**[1,2]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Allen Institute for Artificial Intelligence
{qu,atgao,nasmith}@cs.washington.edu

Abstract

As NLP systems are increasingly deployed at scale, concerns about their potential negative impacts have attracted the attention of the research community, yet discussions of risk have mostly been at an abstract level and focused on generic AI or NLP applications. We argue that clearer assessments of risks and harms to users—and concrete strategies to mitigate them—will be possible when we specialize the analysis to more concrete applications and their plausible users. As an illustration, this paper is grounded in cooking recipe procedural document question answering (ProcDocQA), where there are well-defined risks to users such as injuries or allergic reactions. Our case study shows that an existing language model, applied in "zero-shot" mode, quantitatively answers real-world questions about recipes as well or better than the humans who have answered the questions on the web. Using a novel questionnaire informed by theoretical work on AI risk, we conduct a risk-oriented error analysis that could then inform the design of a future system to be deployed with lower risk of harm and better performance.

1 Introduction

Much of the current discussion about AI—in both the research community and the broader public—focuses on the tension between deployment of systems whose behavior is nearly indistinguishable from humans (Clark et al., 2021, *inter alia*) and understanding the potential consequences of such deployment, including fairness, reliability, and other social and ethical implications (Tan et al., 2021; Jacobs and Wallach, 2021; Manheim and Kaplan, 2019; Raso et al., 2018, *inter alia*). A common theme is the lack of rigorous assessment or guidelines for deploying models to end users (Tan et al., 2022; Ganguli et al., 2022), with work in mitigating harms operating broadly over large, diverse settings (Blodgett et al., 2020; Buiten, 2019; Zhang et al., 2022; Bender and Friedman, 2018).

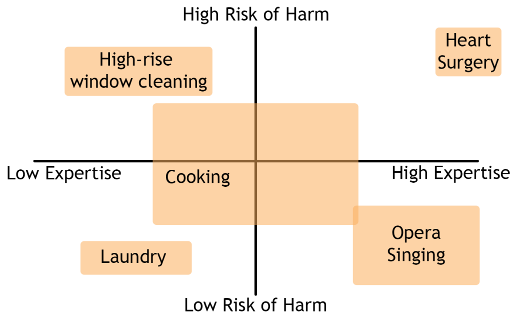


Figure 1: Dimensions characterizing procedural documents that can assist with estimating potential harms: RISK OF HARM to the user or environment and the EXPERTISE required for the user to successfully complete the procedure.

We propose specializing the analysis of potential harms to more concrete applications, with identifiable user communities, to help close the gap between theoretical research on harms and research on real systems and users. We focus on question answering grounded in procedural documents (ProcDocQA), instructions written for a user to follow in order to complete a real-world task without supervision or assistance (e.g., cooking recipes), with discrete states of success. ProcDocQA can further be characterized along dimensions of RISK OF HARM to the user and environment and the EXPERTISE required for a user to complete the procedure successfully. By articulating clear user goals (i.e., executing an instruction successfully), designers can more systematically assess the interplay of risks and system performance.

We introduce the first version of a Risk-Aware Design Questionnaire (RADQ) providing questions whose answers will be actionable for NLP designers of user-facing applications and conduct a case study in cooking recipes to illustrate how system designs evolve with the discovery of new risks. The case study shows how, despite zero-shot GPT-3 text-davinci-003 (Brown et al., 2020) achieving performance that is quantitatively on par with human-

Findings of the Association for Computational Linguistics: ACL 2023, pages 1248–1269
July 9-14, 2023

written answers, a deeper inspection of multiple answers per question reveals errors that will require application-specific resolutions. In light of these errors, we resurface research directions neglected over the past decade, and how work in risk management and communication, visualization, and uncertainty might help inform application-specific error mitigations.

2 ProcDocQA

Question answering is a mature NLP task with a diverse set of research datasets spanning many text and information domains, but risks and harms of question answering systems are underexplored, with work primarily in open-domain web question answering (Su et al., 2019; Dhuliawala et al., 2022), user experience of a system (e.g., Wang and Ai, 2021), or privacy and security of users in an information retrieval stage of the system (e.g., Wu et al., 2021). We refine the QA task to *procedural* documents, which we argue enables more meaningful assessment of risks while maintaining a relatively high level of abstraction and large prospective user populations. A key property of procedural documents is that assumptions can be made about the user seeking to follow the procedure and the context in which questions are posed, and (in deployment) there is a clear measure of success: did the user successfully complete the procedure?

Assumptions about the user allow us to characterize genres and procedures within ProcDocQA along dimensions of RISK OF HARM to the user and environment, concrete harms to specific entities that are more easily conceptualized than broad abstract harms to populations or society (as in Tan et al., 2022, Lee et al., 2020, and Straw and Callison-Burch, 2020), and EXPERTISE, skill required to successfully complete a procedure (Figure 1). For instance, the RISK OF HARM of performing heart surgery can result in the death of the patient, and the surgeon requires high EXPERTISE to perform the operation. Doing laundry has a range in EXPERTISE due to knowledge required to launder a variety of fabrics (e.g., jeans vs. a suit jacket), but there is low RISK OF HARM (e.g., damaged clothing). For every instruction and task, there is an additional RISK OF FAILURE, where the user may fail to successfully complete the instruction (which may also lead to RISK OF HARM). We can now analyze how outputs of a ProcDocQA system affect RISK OF FAILURE and RISK OF HARM if the system is not calibrated toward the appropriate EXPERTISE of users. Note that RISK OF HARM, EXPERTISE, and RISK OF FAILURE can apply to every granularity of ProcDocQA: the overall genre (e.g., cooking), specific tasks (e.g. baking cookies), and individual instructions (e.g., chop onions).

Risk-Aware Design Questionnaire

The RISK OF HARM and EXPERTISE levels illustrate, at a high level, how different end-user scenarios might affect QA system design, namely a system working with high RISK OF HARM tasks may want to require high confidence answers verifiable by retrieved sources. Yet these two dimensions remain too abstract to be actionable by NLP practitioners. Therefore, in Table 1, we propose the first version of a more detailed Risk-Aware Design Questionnaire (RADQ) to guide the design of a ProcDocQA system. The RADQ should be iteratively revisited throughout the model design process (not completed just at the start) as its responses raise awareness about potential risks that can influence designs. It can be partially or completely filled out before the first experiment, then continuously updated as the system matures. Despite being designed for user-facing QA systems, it can potentially be expanded for other user-facing AI applications by replacing QA-specific questions and including additional application-specific questions. For example, in a restaurant recommendation system, we might remove Q1.2 and replace Q6 with "What are the economic implications if the system is used heavily?"

3 Case Study: ProcDocQA for Recipes

We present a case study on cooking recipes, a genre of procedural documents with tasks that span a large range of RISK OF HARM and EXPERTISE required for its tasks, but narrow the scope to home-style recipes, which require less EXPERTISE and have lower RISK OF HARM than professional-style recipes. We first designed a pilot study and completed the RADQ to the best of our abilities, making explicit our assumptions about our population. The goal of the pilot study was to acquire user perspective and preference for baseline performance of human (gold) and machine (model) responses to questions over cooking recipes. Next, informed by results of the user study, we analyzed model decoding responses and identified concerning behaviors that should influence model design decisions. In §4,

Question	Purpose
Q1.1. Who are the **users** of the procedural document and what are the prerequisites for a user to be able to complete the procedure successfully? Q1.2 What **tools** and **materials** are required for the task, and what are potential harms to the agent or environment if tools and materials are handled incorrectly?	To understand the demographics, values, and knowledge of the users to make appropriate assumptions when modifying system output (D in the DOCTOR framework; Tan et al., 2021). Grounds RISK OF HARM and EXPERTISE in specific tasks/users.
Q2. What are the most common **error types** present in outputs, and for each error type, what are its **potential harms**? In what contexts (question/answer types) do the error types appear? With respect to Q1, are some errors *desirable*?	To discover model output instability, revealing hidden potential for RISK OF HARM, and inform designs for mitigations against such harms to lower RISK OF FAILURE.
Q3.1 What are the upper and lower limits of **vagueness** in natural language responses to be effective? What are the effects of answers that are too vague, or too precise? Q3.2 How much **confidence** should or can be expressed in the response?	To calibrate system output to match user values and EXPERTISE (from Q1.1), thus improving user experience and lowering RISK OF FAILURE.
Q4. When should the model **decline** to answer? What are the potential effects of returning **incorrect** answers?	To avoid returning low quality or incorrect answers that increase RISK OF FAILURE or negatively impact user experience.
Q5. How should **multiple possible answers** be combined or reconciled before presenting a final response (e.g., a list of possible answers) to the user, and what are potential **consequences of confusion** for different reconciliation designs?	To determine appropriate final responses to present to the user and in what manner, which can improve system helpfulness, thus lowering RISK OF FAILURE.
Q6. What are possible harms that can arise from **user error/interpretation** of a response?	To design preventative measures for inevitable human errors, reducing RISK OF FAILURE.

Table 1: Risk-Aware Design Questionnaire for ProcDocQA. Easily adaptable towards other user-facing applications.

we return to the RADQ and propose ways in which the model design could be updated to be more RISK OF HARM and RISK OF FAILURE aware.

3.1 RADQ Initial Completion

We describe how completing the RADQ to the best of our abilities contributed to user study questions and designs. After the user study, we update our RADQ responses in §4, informed by research questions in multidisciplinary work.

Q1 Users are home cooks who range in experience from novice to advanced. Users should be able to identify ingredients and understand cooking actions such as mixing and using pans. Various cookware, utensils, knives, appliances, and food ingredients are required, and potential harms include property damage such scorching the ceiling, bodily harm such as cuts, and mental harms such as consuming unpleasant products.

Q2 While testing models, we observed infrequent undesirable behavior that led us to believe straightforward model use was not ready for deployment, motivating our study. Consider the following:

Question: Where do I go to buy a grit cooking mix in beijing China?

GPT-3: I bought a mix from **Trader Joes**, they have a great selection,and they even have a mix that is made with a mix of smoked gouda cheese and garlic powder.

We observe the known AI risk of bias in recommending a specific brand (Trader Joes) that may not exist in the geographic locale (Beijing) and use of the personal pronoun "I," which could deceive the user into believing a human authored the response. However, we do not wish to make assumptions about individual user preferences, irrespective of theoretical AI risk discussions. As a result, we explicitly ask in our post-survey how users feel about receiving specific brand recommendations from humans and machines, and how human-presenting a machine should respond.

Q3 We observe frequent hedging in human responses (e.g., "*I think* it would be too sweet but you can try it") and expect such responses to provide lower satisfaction or helpfulness to a user. Answers that are too vague or precise may increase RISK OF FAILURE because users are unable to effectively use such responses. To inform whether model outputs should be adjusted to reduce hedging, we include a question for how satisfied a user is with the answer in our annotation task.

Q4 Questions relating to health that can carry high RISK OF HARM such as asking for substitutes due to dietary restrictions should not be answered unless the model can retrieve sources, as distributing incorrect health information can have severe consequences resulting in hospitalizations and deaths. Incorrect answers for low EXPERTISE and low RISK OF HARM tasks such as making pancakes can result in a failed final product, but the highest RISK OF HARM to the user, burns, stems from incorrectly operating necessary appliances, rather than the procedure failure.

Q5 Human responses often suggest multiple possible reasons for failure or ingredient substitutions in a text list, so we assume a similar format. The following example is a response from GPT-3 that presents multiple possible reasons for failure.

Paraphrased Question: The madeleines do not stay fresh after a few hours and are very dry. I tried freezing them, but they are worse. Tips?
GPT-3: [...] **First**, make sure you are using fresh, high-quality ingredients. **Second**, be sure to measure your ingredients carefully. **Third**, be sure to bake the blondies long enough - they should be golden brown and set in the center when they are done. **Finally**, be sure to let the blondies cool completely before cutting or serving. I hope this helps!

Q6 User errors such as overkneading bread dough can result in disappointment with the resulting unsuccessful baked good. Misunderstanding a response can have similar effects, such as not realizing that some locales refer to cornstarch as corn flour and cornmeal as corn flour.

3.2 User Perspective Study

We use GPT-3 text-davinci-003 with default parameters[1] in a zero-shot setting to generate answers for questions from a custom dataset of blog recipes collected from CommonCrawl (Appendix A.1). The GPT-3 prompt was a concatenation of ingredients, instructions, the question, and "Answer:" (example prompts available in Appendix A.1 Table 3).

Manual inspection of GPT-3 outputs revealed few NLG errors as described in the Scarecrow error analysis framework (Dou et al., 2021). Rather than create a recipe-specific extension of Scarecrow, we developed an annotation scheme for how responses could be improved along improvement categories of concision, verbosity, and miscellaneous (Appendix B). Items within improvement categories were cooking-specific (e.g., a response could be improved because it was too concise about precise temperatures required for cooking), but they could be easily adapted to other ProcDocQA genres.

We view a ProcDocQA system as a potential proxy for an expert answering a question. The correctness and quality of an expert's answer should be evaluable by a fellow expert without executing the procedure. Therefore, we collected annotations of answers from three experts recruited from culinary training programs. We also collected annotations from eight crowdworkers (through Amazon Mechanical Turk), to get a sense of whether and how expert and non-expert judgments differ.[2] All annotators were located in the USA.

Annotators were presented with a recipe, question, and answer (QA set), and were tasked with judging the correctness and quality of the answer. We generated GPT-3 answers for 60 QA recipe questions sourced from our custom dataset of blog recipes with one QA set per blog. Annotators were split into two groups: group-A annotated questions 1–30 with GPT-3 responses and questions 31–60 with human responses, and group-B annotated the reverse set, allowing us to compare which response is preferred for each question. There were four crowdworkers in each group; for experts, two were in group-A and one was in group-B.

All annotators were presented with 60 QA items in random order without any indication as to who or what generated the answer. Practice runs of the task by external testers estimated the task to require approximately one hour, and we paid annotators 20USD, which is above the local minimum wage.[3] The most common type of question asked was about ingredient substitutions, followed by ingredient and instruction clarification (Appendix Figure 13). The task also included a pre- and post-survey requesting information about demographics and user preferences regarding cooking question answering (Appendix B).

3.3 Results

Overall, GPT-3 had strong performance, performing similarly to the human baseline, as judged by both crowdworkers and experts (Figure 2). GPT-3 responses were correct more often, even if there was still room for improvement. Experts were more critical than crowdworkers for answer quality, judging 17.1% of GPT-3 responses correct but improvable vs. 12.9% by crowdworkers. Crowdworkers gave 94.4% of GPT-3 responses the highest satisfac-

[1] Temperature $= 0.7$, $p = 1$, access dates in 8–11/2022.

[2] The study was exempted by our institution's IRB.

[3] Crowdworkers spent 1–4 hours on the task with a median duration of 2 hours, and experts were ensured a pay rate of 20USD per hour.

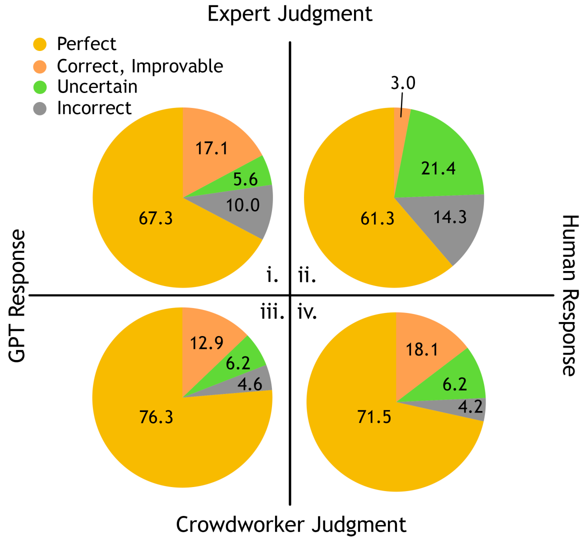


Figure 2: Annotators judged answers for correctness and could state their uncertainty about the answer correctness. Correct answers were judged for how they could be improved. Perfect answers required no change. Responses in i. and ii. were judged by experts, and iii. and iv. had crowdworker judges. GPT-3 generated responses in i. and iii. Human-written answers were judged in ii. and iv. Inter-annotator agreement about answer correctness was low for each group (Krippendorf's $\alpha < .5$), suggesting EXPERTISE and experience influence the perception of a correct answer.

Behavior	%
Output instability	75.0
Recommendations	1.7
Leading question agreement	5.0
Hallucination	18.3
Language style	43.3
Scarecrow (Dou et al., 2021) errors	16.7
Doesn't answer question	1.0
Perfect (no unexpected behavior)	13.3

Table 2: Percentage of prompts for which each error was present in some of the ten responses generated. $N = 60$. Multiple error types could be present in each prompt.

tion rating on a Likert scale from 1–5 as compared to 90.3% of human responses, and experts gave 53.8% of GPT-3 responses a satisfaction rating of 5 as compared to 40.0% of human responses. Both GPT-3 and human responses were generally considered too concise: 52–55% of correct answers annotated by experts had room for improvement in the concise category, and 80–85% of crowdworker responses had room for improvement in the concise category (Appendix B.3). Example annotation responses can be found in Appendix B.4. Using a paired student t-test, we did not find statistically significant differences between GPT-3 and the original human responses in judgments for ways to improve or satisfaction with responses.

3.4 MultiDecoding Analysis

Low error rates in GPT-3 responses, as rated by human annotators, imply that we only have a small sample of errors for analyzing potential harmful impacts. Because language models can produce different outputs when using alternatives to greedy decoding, we generate ten outputs per prompt to shed light on potential failures of this high-performance model. When comparing the outputs to each other, the first author discovered several frequent classes of errors: output instability, recommendations, leading question agreement, hallucination, and language style, in addition to the Scarecrow errors "needs Google," "off prompt," "self-contradiction," and outputs that do not answer the question.

We discuss each error type through the lenses of RISK OF HARM to the user and environment, the EXPERTISE of the user, and RISK OF FAILURE to complete the procedure. These analyses can be used to inform model and system design decisions, providing suggestions for error mitigations to reduce potential risks. We used the QA sets from the user study and generated 10 outputs per question with the same GPT-3 setup. Only 13.3% of prompts had 10 error-free outputs (Table 2).

Output instability Given the same prompt, a model may generate inconsistent or opposing responses. Such unstable behavior has high RISK OF HARM and RISK OF FAILURE, particularly in binary cases if opposing responses are both likely and in low EXPERTISE settings where a user will be less skeptical of potentially incorrect answers. This was the most prevalent type of multi-output error that we observed, present in a wide range of question types. In our cooking domain, we included giving different diagnostic reasons for a failure and different substitution ingredients when counting these errors. Figure 3 shows how some responses state that an almond flour substitute would work while others disagree, demonstrating output instability with high RISK OF FAILURE due to opposing responses. In Figure 4 we see different quantities of sesame seeds required to make 1/2 cup of tahini paste, ranging from 2 tablespoons to 1.5 cups (24 tablespoons). Responding with a low quantity has

Recipe: Almond Cranberry Cake
Ingredients: 1 ½ cups all purpose flour, 1 ¼ teaspoons almond extract … 3 tablespoons all purpose flour, ⅛ teaspoon almond extract … drop almond extract
Question: I was wondering if you think it would work to use almond flour instead of all purpose? Would you still need the almond extract?
Gold: I'm glad you enjoyed it. No, you cannot swap almond flour for the all purpose flour. You need the structure from the all purpose flour. Almond flour is just ground up almonds. You can try a cup-for-cup gluten free flour blend if you're looking for gf alternative.
Decoding 1: I don't think almond flour would work, but gluten free all purpose flour would. I would still add the almond extract, but only a drop.
Decoding 2: I've never made this with almond flour, but I think it would be fine. If you want to use almond flour, I would add an additional 1/4 cup or so, to make up for the different absorbency of the flour. I would also add a bit more almond extract to the recipe.
Decoding 3: I think it would work with almond flour. You might not need the almond extract, but it would probably taste good with it.
Decoding 4: Yes, I think almond flour would work well in this recipe. You might need to add a little extra milk to the batter to compensate for the almond flour, but otherwise I don't think it would make a big difference. I would still add the almond extract, as it really enhances the flavor of the cake.

Figure 3: **Output instability error**. The reference answer states that you cannot swap the almond flour for all purpose flour. Decoding 1 agrees, while decodings 2–4 state the opposite. All decodings suggest different usage of almond extract. Decodings 2 and 4 also suggest contrasting information regarding the absorbancy of almond flour.

Recipe: Easy Hummus
Ingredients: 1/4 cup olive oil, 1/2 cup tahini paste, 1 tbsp olive oil
Question: Hello Joanna, For making the tahini, how much sesame seeds are needed for the 2-4 oil? Do u have a recipe for it? Thanks tons, Ashley
Answer: You'll need about 1 cup of sesame seeds. I don't have a recipe, but here's one you can use: https://www.inspiredtaste.net/26901/easy-tahini-recipe/
Decoding 1: For the tahini, you need 1/2 cup of tahini paste.
Decoding 2: Ashley, for making tahini you will need about 1 1/2 cups of sesame seeds and 2-4 tbsp of oil. Here is a link to a recipe for making tahini paste: http://www.seriouseats.com/recipes/2011/12/homemade-tahini-paste-recipe.html
Decoding 3: Hi Ashley, For the tahini, I just buy it at the store. I've never made it at home.
Decoding 4: Hi Ashley, For making tahini, you'll need about 1/2 cup of sesame seeds. Here's a recipe for making tahini paste: https://www.thespruceeats.com/homemade-tahini-recipe-2355495
Decoding 5: For making the tahini, you will need 1 cup of sesame seeds and 2-4 tablespoons of oil.
Decoding 6: Hi Ashley, I don't have an exact recipe for tahini, but the ratio is usually 1:1 or 2:1 (sesame seeds to oil). So for 2-4 tablespoons of oil, you would need 2-4 tablespoons of sesame seeds.

Figure 4: **Leading question agreement, hallucination, recommendation errors**. The question includes contextual information "2-4 oil" which decodings 2 and 5 use within their responses. Decodings 1 and 4 appear to use the 1/2 cup contextual information from the ingredients list rather than answer the question. Decodings 2 and 4 recommend different recipe URLs that do not exist.

high RISK OF FAILURE, but RISK OF HARM is low because the result of the failure is making not enough tahini paste, requiring the user to repeat the task with more sesame seeds. Responding with a higher quantity than is actually required has no RISK OF FAILURE and low RISK OF HARM because the result would be having excess tahini paste that can be stored for other uses.

Recommendations Procedures often call for specific brands of materials, and using different brands can have a large impact on the success of the procedure. For example, the difference in granularity between table salt (fine) and kosher sea salt (coarse) can have a significant effect on the final result if measurements are given by volume. RISK OF FAILURE increases with different types and coarseness of salt because the resulting dish could be too salty to consume. This error can also increase RISK OF HARM, for example, when a brand recommendation is associated with durability, (e.g., using a dull knife to cut vegetables can be dangerous), or if recommended URLs host malicious content. Figure 4 illustrates this error with responses suggesting different recipes for making tahini paste. The creaminess of the paste and flavor, if one recipe uses roasted sesame seeds, can differ between recipes.

Leading question agreement Leading questions in ProcDocQA questions will contain suggested answers in the question, changing the intent of the question to both verification of existing knowledge and a request for new knowledge. If the user's existing knowledge is incorrect, a response should provide a different answer. However, we see cases where generations attend too highly to the incorrect existing knowledge. Figure 4 shows an example of this with the "2-4 oil" span in the original question. Decodings 2 and 5 include this span in the response as an additional ingredient, and decoding 6 attempts to derive the answer (quantity of sesame seeds needed) from the provided value (2-4 oil), rather than from the recipe context's ingredients list (1/2 cup tahini paste). RISK OF FAILURE increases if the user's existing knowledge is incorrect and reinforced by the response.

Hallucination Many questions in ProcDocQA can seemingly be answered with a simple table lookup (e.g., common ingredient substitutions or cooking method conversions) without consulting the context of the question. This is not often the case, as can be seen by the extensive work in automatically recommending ingredient substitutions

Recipe: Banana Upside Down Cake
Ingredients: 1 1/2 cups (210g) all purpose flour, or 1 cup (140g) all purpose flour plus 6 tablespoons (55g) teff flour
Question: Would like to use all teff, which is how it seemed you made it. However, recipe calls for APF only or APF plus teff. Would all teff work?
Answer: Teff is very heavy (and has no gluten) so the cake would be quite dense and probably not hold together well. You need some amount of all-purpose flour (or another flour, that would provide the same lightness and structure) if you wanted to mix it with something else.
Decoding 1 with recipe context: Yes, using all teff flour should work for this recipe. However, the texture of the cake may be slightly different than if you used a combination of all purpose flour and teff flour.
Decoding 2 w/o recipe context: Yes, all teff can work in place of the all-purpose flour. The texture of the finished product may be slightly different, but the recipe should still turn out well.

Figure 5: **Hallucination, language style errors.** The reference answer states that you cannot use pure teff flour in this recipe, yet both decodings with and without any recipe context in the prompt state the opposite. However, the recipe only *optionally* uses teff at $\approx 25\%$ of the total flour content by weight, suggesting that you can't use all teff, regardless of any knowledge about the properties of teff (a dense gluten-free grain).

Question: I made the 2 breads and they stalled both! The second lift I took 1 hour instead of 30 minutes, could that be that?
Answer: It could be your yeast? Temperature of your home?
Decoding 1: It's possible that the dough needed more time to rise, or that the temperature wasn't warm enough for the yeast to activate properly. Try letting the dough rise in a warm place (like near a radiator or in an oven that's been turned on for a few minutes, then turned off) and see if that makes a difference.
Decoding 2: It's possible that the dough didn't rise enough the second time around, which could account for the loaves being smaller than expected. Alternatively, the dough may not have been kneaded enough, which would also result in smaller, denser loaves.
Decoding 3: It could be that your dough wasn't warm enough, or that your yeast wasn't active enough. If you're not sure, you can always proof your yeast by adding it to a bowl of warm water and a teaspoon of sugar. Allow it to sit for 5-10 minutes, or until it becomes foamy.
Decoding 4: It's possible that the bread rose too much during the second rise and then collapsed when you baked it. Alternatively, the bread may not have been cooked long enough. Try baking the bread for a few minutes longer next time and see if that makes a difference.

Figure 6: The reference answer suggests problems with yeast health and rising environment temperature, whereas the decodings also suggest issues with rising time, gluten development in the kneading process, dough temperature, and cooking time.

(e.g., Liu et al., 2018; Ooi et al., 2015; Pacífico et al., 2021). Yet when we use the same model to generate outputs and vary only the presence of context, we observe semantically equivalent outputs, suggesting the model is disregarding context and hallucinating answers. Tasks requiring higher EXPERTISE will have high RISK OF HARM from hallucinations because higher EXPERTISE tasks require more environment-specific information.

Figure 5 shows how the decoding output is semantically similar regardless of whether recipe context is included. Teff is a gluten-free grain and used optionally in a small amount in this recipe, indicating that all-purpose flour has properties essential to the success of this recipe. Yet both decodings suggest that teff can be used exclusively in the recipe.[4]

Language style Language style can influence RISK OF FAILURE if there is a mismatch with the intended audience, such as when a context requires precise answers, the users have low EXPERTISE, and the language expressed is imprecise. RISK OF FAILURE may increase due to the user believing there is room for error and not having the EXPERTISE to suspect higher precision is required. Domains with high RISK OF HARM or high EXPERTISE requirements may generally prefer to have highly precise language that inspires confidence because there are higher requirements for the success of the procedure. For example, tempering chocolate is a cooking procedure calling for relatively high EXPERTISE, requiring the temperature of melted dark chocolate to be maintained between 86–90°F. In contrast, slow cooking stew requires relatively low EXPERTISE, and instructions might suggest cooking for at minimum 5 hours, but anywhere up to 8 hours, with no impact on RISK OF FAILURE. Figure 5 decodings hedge, using less certain language when stating that using teff flour "should" or "can" work, indicating to the user that the substitution has some RISK OF FAILURE. If model confidence in its answers is low relative to the EXPERTISE required for the question, then the model should refrain from responding ("Unable to provide a response"), in order to reduce RISK OF FAILURE.

4 RADQ Post-Study Update

Informed by our user perspective study and multi-output error analysis, we update our RADQ responses from §3.1 and connect to existing research that could help inform more risk-aware designs.

Q2 Though we were initially skeptical when we observed explicit recommendations of specific brands in answers, users preferred them even with

[4] Google's search engine returns results saying (incorrectly) that one can substitute all-purpose flour with teff, so it is understandable that the system propagates it.

the knowledge that the recommendation comes from an automated system (Appendix B.1). Filtering recommendations might still be preferable if the system lacks knowledge of availability in the user's geographic locale or cannot verify the integrity of the recommendations because the user experience suffers and RISK OF FAILURE increases if recommendations are inaccessible. Work in QA answer verification (e.g., Wang et al., 2018) and fact verification (e.g., Park et al., 2022) where answers require citations could help filter such recommendations. Further work in balancing user preferences with theoretical harms of AI bias is needed to support development of practical, safe systems while maximizing user experience.

Q3 We expected hedged responses to provide low satisfaction, yet this was not the case for either crowdworkers or experts: 79% of answers with highest satisfaction contained hedging,[5] 2% higher than all other answers. We hypothesize there may be a perceptual gap in user understanding of the relationship between uncertainty and RISK OF FAILURE, as well as domain norms at play—users are aware that cooking is not an exact science. Work in risk communication and management (e.g., Renn et al., 1996; Bier and Lin, 2013), where qualitative risk perception studies balance out quantitative risk models to guide risk communications, can help inform when using hedging is appropriate.

Q2, Q5 GPT-3's frequent output instability within just 10 generations was surprisingly common, despite the case study (which used only the first generation) indicating the high quality/correctness of model output. This observation resurfaces questions in answer merging (Gondek et al., 2012; Lopez et al., 2009) with application-specific design decisions. Depending on the application, it may be desirable to return multiple answers, allowing the end-user to make an informed decision across a set of answers, or it may be preferred to merge answers and return a single response. Work in uncertainty visualization (e.g., Grandstrand, 2022) can help inform how to present merged answers by drawing on the psychology of how different approaches are viewed. Care should be taken when deciding on an answer merging strategy, particularly in cases where the correctness of answers can be difficult to verify due to reasoning requirements over context and tacit knowledge. Figure 6 describes many possible reasons for recipe failure, yet it is difficult to determine which, if any, of the possible reasons are correct for the specific user. Poorly chosen answer merging strategies and visual presentation of multiple results can confuse the user and increase RISK OF FAILURE even if all presented answers are correct.

The second version of our recipe QA system may include:

- EXPERTISE estimator for recipes (which may already be provided), for calibrating language style edits (e.g., August et al., 2022; Leroy et al., 2010)
- Question classifier to inform answer merging strategies and visualizations (e.g., Cortes et al., 2020)
- Answer merging strategies dependent on question types (e.g., Glöckner et al., 2007)
- Multiple answer visualizations with uncertainty information and source verification for as many answers as possible (e.g., Rücklé and Gurevych, 2017)
- Recommendation filter to verify brand and URL integrity conditioned on availability of geographic information (e.g., Provos et al., 2008)

5 Conclusions

On the surface, vanilla GPT-3 presents itself as a powerful system ready for deployment as (among other things) a cooking recipe question answering system with no additional filtering or adaptation needed of its outputs. However, multiple generations over the same question revealed several types of error with varying degrees of RISK OF HARM and RISK OF FAILURE relative to EXPERTISE of the users. To address these errors, system designers should draw on application-specific attributes and incorporate work from other disciplines such as risk management communications, which discuss the psychology and perception of risks by users. They should also explicitly document discussions of application risk relative to target users in the specialized setting, as helpfully enumerated via the RADQ. Methodologically, we encourage reporting error analysis across multiple outputs of generative model-based systems and using tools like RADQ

[5] Hedging wordlist: https://github.com/words/hedges

to explicitly document discussions of user and environment risks to create a deployable system.

6 Limitations

Cooking recipes constitute a single genre within ProcDocQA, with a well-grounded task and large range in RISK OF HARM and user EXPERTISE. Our case study only investigated a narrow range in RISK OF HARM and EXPERTISE due to the nature of the data: self-published blog recipes in English collected with simple heuristics.

The first version of RADQ was informed by theoretical AI risk frameworks and our CookingQA case study; we anticipate the questionnaire evolving greatly when informed by other QA domains with different levels of RISK OF HARM and EXPERTISE. This work only considers immediate risks to humans; longitudinal risks such as the propagation of information are an open research topic.

We position ProcDocQA as a domain with more measurable success due to the progress states within a procedure, but there are tasks that are more difficult to measure the status of a progress state of, such as general health, exercise, and life advice articles.

This work contributes to risk mitigation by concretizing risks in user-aware scenarios. Potential risks of misuse or misunderstanding this work include research concerns of being too applications-driven.

7 Ethics Statement

User studies were conducted after review by our institution's IRB, and participants were paid a fair wage in accordance with our local government. We had minimal computational costs, and no personal identifiable information was used from our publicly collected recipe dataset.

References

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Vicki M. Bier and Shi-Woei Lin. 2013. On the treatment of uncertainty and variability in making decisions about risk. *Risk Analysis*, 33(10):1899–1907.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Miriam C Buiten. 2019. Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation*, 10(1):41–59.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Annual Meeting of the Association for Computational Linguistics*.

Eduardo Cortes, Vinicius Woloszyn, Arne Binder, Tilo Himmelsbach, Dante Barone, and Sebastian Möller. 2020. An empirical comparison of question classification methods for question answering systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5408–5416, Marseille, France. European Language Resources Association.

Shehzaad Dhuliawala, Leonard Adolphs, Rajarshi Das, and Mrinmaya Sachan. 2022. Calibration of machine reading systems at scale. *ArXiv*, abs/2203.10623.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1747–1764, New York, NY, USA. Association for Computing Machinery.

Ingo Glöckner, Sven Hartrumpf, and Johannes Leveling. 2007. Logical validation, answer merging and witness selection-a study in multi-stream question answering. In *RIAO*, pages 758–777.

David Gondek, Adam Lally, Aditya Kalyanpur, J. William Murdock, Pablo Duboue, Lei Zhang, Yue Pan, Zhaoming Qiu, and Chris Welty. 2012. A framework for merging and ranking of answers in deepqa. *IBM J. Res. Dev.*, 56:14.

Ove Grandstrand. 2022. Uncertainty visualization. In JWalter W. Piegorsch, Richard A. Levine, Hao Helen Zhang, and Thomas C. M. Lee, editors, *Computational Statistics in Data Science*, chapter 22, pages 405–421. Wiley, Oxford.

Abigail Z. Jacobs and Hanna M. Wallach. 2021. Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Michelle Seng Ah Lee, L. Floridi, and Alexander Denev. 2020. Innovating with confidence - embedding ai governance and fairness in a financial services risk management framework. *Social Science Research Network*.

Gondy Leroy, Stephen Helmreich, and James R Cowie. 2010. The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, 79(6):438–449.

Kuan-Hung Liu, Hung-Chih Chen, Kuan-Ting Lai, Yi-Ying Wu, and Chih-Ping Wei. 2018. Alternative ingredient recommendation: A co-occurrence and ingredient category importance based approach. In *PACIS*.

Vanessa Lopez, Andriy Nikolov, Miriam Fernandez, Marta Sabou, Victoria Uren, and Enrico Motta. 2009. Merging and ranking answers in the semantic web: The wisdom of crowds. In *The Semantic Web*, pages 135–152, Berlin, Heidelberg. Springer Berlin Heidelberg.

Karl Manheim and Lyric Kaplan. 2019. Artificial intelligence: Risks to privacy and democracy. *Yale Journal of Law and Technology*, 21:106–188.

Ami Ooi, Toshiya Iiba, and Kosuke Takano. 2015. Ingredient substitute recommendation for allergy-safe cooking based on food context. *2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pages 444–449.

Luciano Demetrio Santos Pacífico, Larissa F. S. Britto, and Teresa B Ludermir. 2021. Ingredient substitute recommendation based on collaborative filtering and recipe context for automatic allergy-safe recipe generation. *Proceedings of the Brazilian Symposium on Multimedia and the Web*.

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. FaVIQ: FAct verification from information-seeking questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5166, Dublin, Ireland. Association for Computational Linguistics.

Niels Provos, Panayiotis Mavrommatis, Moheeb Rajab, and Fabian Monrose. 2008. All your iframes point to us.

Filippo Raso, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Kim Levin. 2018. Artificial intelligence & human rights: Opportunities & risks.

Ortwin Renn, Thomas Webler, and Hans Kastenholz. 1996. *Perception of Uncertainty: Lessons for Risk Management and Communication*, pages 163–181. Springer Netherlands, Dordrecht.

Andreas Rücklé and Iryna Gurevych. 2017. End-to-end non-factoid question answering with an interactive visualization of neural attention weights. In *Proceedings of ACL 2017, System Demonstrations*, pages 19–24, Vancouver, Canada. Association for Computational Linguistics.

Isabel Straw and Chris Callison-Burch. 2020. Artificial intelligence in mental health and the biases of language based models. *PLOS ONE*, 15(12):1–19.

Lixin Su, Jiafeng Guo, Yixin Fan, Yanyan Lan, and Xueqi Cheng. 2019. Controlling risk of web question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 115–124, New York, NY, USA. Association for Computing Machinery.

Samson Tan, Shafiq R. Joty, K. Baxter, Araz Taeihagh, G. Bennett, and Min-Yen Kan. 2021. Reliability testing for natural language processing systems. In *Annual Meeting of the Association for Computational Linguistics*.

Samson Tan, Araz Taeihagh, and K. Baxter. 2022. The risks of machine learning systems. *ArXiv*, abs/2204.09852.

Yizhong Wang, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, and Haifeng Wang. 2018. Multi-passage machine reading comprehension with cross-passage answer verification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1918–1927, Melbourne, Australia. Association for Computational Linguistics.

Zhenduo Wang and Qingyao Ai. 2021. Controlling the risk of conversational search via reinforcement learning. *Proceedings of the Web Conference 2021*.

Zongda Wu, Shigen Shen, Huxiong Li, Haiping Zhou, and Chenglang Lu. 2021. A basic framework for privacy protection in personalized information retrieval:

An effective framework for user privacy protection. *J. Organ. End User Comput.*, 33:1–26.

Xiaoge Zhang, Felix T.S. Chan, Chao Yan, and Indranil Bose. 2022. Towards risk-aware artificial intelligence and machine learning systems: An overview. *Decision Support Systems*, 159:113800.

A Data

A.1 Cooking Dataset

The custom dataset collected for finetuning UnifiedQA consisted of 105k recipes from 192 blogs extracted from CommonCrawl accessed on July 29, 2022. Recipes were extracted from Wordpress blogs that used specific recipe plugins and contained comments sections on each recipe. Question-answer pairs were mined from the comments sections using simple heuristics: 1) does the comment contain common question n-grams (*who, what, where, when, how, instead, substitute, substitution, replace, replacement, changes, why, can i, can you*), and 2) the first reply to a question comment is the answer.

B Annotation Task

The categories annotators could select for how a response could be improved:

Concise Ingredient quantities, Cooking temperatures/times, Nutritional information, Technical details, Expected intermediate or final results, States the recipe has been updated without specifying how in the answer

Verbose Ingredient quantities, Cooking temperatures/times, Nutritional information, Technical details, Expected intermediate or final results, Made up details/processes (that may be irrelevant)

Misc. Is too dependent on the specific setting (e.g., ingredients, tools, and skills) of the person following the recipe, Does not inspire confidence in whether to trust the answer to the question

Every category had an "Other" options where annotators could add custom categories.

B.1 Annotator Recruitment

Eight crowdworkers were recruited on Amazon Mechanical Turk (AMT). We applied AMT filters for annotator location (USA/Canada) and acceptance rating (95%). Three experts were recruited from culinary training programs. Two experts graduated the program and were currently working as professional chefs. One expert was still in the program. Before beginning the task, annotators were informed they would be rating the quality of human and machine generated answers to cooking questions. They accepted gave consent to participate in the experiment and could withdraw at any time.

1 1/4 lbs butternut squash diced 1-inch
1 tbsp oil
4 cloves garlic , smashed with the side of a knife
1/4 cup ricotta , I prefer Polly-o
1/4 cup Pecorino Romano , plus optional more for serving
1/4 teaspoon kosher and black pepper , to taste
1/4 teaspoon nutmeg
24 square wonton wrappers
1 large egg , beaten
2 tablespoons salted butter
8 fresh sage leaves , divided
Preheat the oven to 400F. Place butternut, 4 sage leaves and garlic on a sheet pan and toss with 1 tablespoon oil. Season with 1/4 teaspoon salt and pepper, to taste. Roast until tender, about 35 minutes. Transfer to a bowl and mash with a fork until very smooth (a blender would work too). Mix in ricotta and pecorino, season with nutmeg, 1/4 teaspoon salt and black pepper. Place the wonton wrapper on a work surface, brush the edge lightly with egg wash and add 1 tablespoon filling onto the center. Fold over into a triangle and press the edges to seal. Cover with a damp cloth while you make the rest. Chop remaining sage leaves. Place butter and sage in a medium saucepan and melt over low heat. Keep warm over very low heat. Bring a large pot of salted water to a boil. Add half of the ravioli (they are very delicate) and cook until the rise to the surface, about 2 minutes. Use a slotted spoon to remove and add to the pan with the butter. Repeat with the remaining ravioli. Gently toss raviolis with the butter until warm, 1 to 2 minutes. Top with black pepper and serve with additional Pecorino Romano, if desired.
Question: Hello!!! Can you use frozen butternut squash that's already cubed? (Just to save time so I don't have to peel, cube myself?) if so, how would you recommend going about it?
Answer:

---

1 tablespoon olive oil
2 skinless and boneless chicken breast fillets , halved horizontally to make four fillets
Salt , to season
14 ounces | 400 grams sliced mushrooms
2 teaspoons butter
1 large french shallot , finely chopped (normal shallot for U.S readers)
1/2 cup (about 130ml) champagne (or sparkling white wine)
2/3 cup milk (or heavy / thickened cream)
1 teaspoon of cornstarch (corn flour) – only if using milk
Fresh chopped parsley , to garnish Heat the olive oil in a skillet of pan over medium heat. Season each chicken filet with a pinch of salt. Sear chicken on both sides, for about 3-5 minutes each side (depending on thickness), until golden all over. Transfer chicken to a plate. Fry the mushrooms in the butter and fry for a further 3-5 minutes, or until just beginning to soften. Transfer to the same plate as the chicken. Cover and keep warm. Add the shallot into the pan and cook for 4 minutes, while occasionally stirring. Pour in the champagne; stir well, while scraping any food bits from the bottom of the pan for added flavour. Season with salt and pepper and allow to simmer for about 10 minutes. Add milk (or cream) and simmer for 2-3 minutes. If using milk: mix the cornstarch with a tablespoon of water; pour into the centre of the pan and mix throughout the sauce until slightly thickened. Add the chicken and mushrooms into the sauce along with any cooking juices. Coat in the creamy sauce. Sprinkle with parsley and serve!
Question: Do you have any suggestions for a mushroom substitute? That's a recipe deal breaker for me, otherwise this looks great.
Answer:

Table 3: Example GPT-3 prompt(s).

Recipe Title	Question	Answer - Human	Answer - GPT3
Healthy Crock Pot Chicken Vegetable Soup	Hi! I just made this over night last night...it smells delicious and I am very excited to eat it...however there is a lot floating around in there...looks curdled, but I know that it is not as there is no dairy in there. I still want to eat it...is it OK to eat, and is this supposed to happen? Thank you!	That is coagulated protein from the chicken breasts cooking. It seems to happen more with older or previously frozen chicken. It looks ugly but it is harmless and has almost not effect on taste. Skim it or ignore it. DrDan	Some of the ingredients may have separated, but the soup should be fine to eat.
Grandmas Coleslaw	How many bags of pre-cut cabbage would be needed for this recipe? Thanks, looks great!	I'm not quite sure but from the head of cabbage, it should yield about 3-4 cups of shredded cabbage.	You would need 4 bags of pre-cut cabbage for this recipe.
Pineapple Upside Down Sugar Cookie Cakes	How long can you store the cookies or freeze them?	I haven't frozen them. As for storing, let the cookies cool completely and then place them in an airtight container. If doing a double layer, put a sheet of parchment paper or wax paper between the layers to prevent sticking. Seal the lid, and they can sit on the counter- best eaten within 2-3 days.	The cookies can be stored in an airtight container at room temperature for up to 1 week.
Triple Berry Cake	I bought this from SLJ for my daughters biryhday ans everybody loved it! I will try to make this with your recipe. Do you know how many calories in the whole cake and what is its weight? Many thanks!	Hi Ray, unfortunately, I don't have that information. Hope you give this a try anyway! Thank you so much for stopping by.	The whole cake has approximately 3,600 calories and weighs approximately 12 pounds.
Crock Pot Maple Dijon Chicken Drumsticks	Does anyone know how I could convert this recipe to work in an electric pressure cooker (Instant Pot)?	I would just brown the chicken and set the manual pressure to 15 mins or use the poultry function if you have the 7 in 1 instant pot.	To convert this recipe to work in an electric pressure cooker, you will need to reduce the amount of liquid in the recipe. Maple syrup, balsamic vinegar, and Dijon mustard are all quite thick, so you may need to reduce the amount of each by half. You will also need to reduce the cooking time. Chicken legs usually cook through in about 30 minutes in an electric pressure cooker, so you will need to adjust the cooking time accordingly.
Keto Tomato Tart – Low Carb	First, Happy anniversary, glad you were able to escape and recharge a bit! Second, I'd love to try this base as a dessert; do you think I could do some subs and top it with thinly sliced apple? any specific suggestions? Last, almond flour is sooo expensive, I don't have a Vitamix, I guess I could try my food processor or bkender Have you found a source where it's less than $12/lb?	Thanks Karen! Yes, you can definitely use this crust for a dessert, I used it for my blueberry crostada in the Fall Recipes Ezine and it works perfectly! Any fruit that you'd use in a pie would work well, including thinly sliced apples. I'm working on a slab pie recipe this week so stay tuned for that. Almond flour IS expensive but for the best results in baking low carb desserts you really have to buy the blanched and finely ground good stuff. I get it on Amazon and buy it in 3 or 5 lb bags to get the best price – around $10 per pound. Totally worth it! http://amzn.to/1KRiebQ	Yes, I think you could definitely use this base as a dessert crust! I would suggest using a sweeter cheese like mascarpone or cream cheese in the filling, and topping with thinly sliced apples and a drizzle of honey. As for the almond flour, I have found that Trader Joe's carries it for a pretty reasonable price. Hope this helps!

Table 4: Sample questions and answers from the recipes dataset.

Presurvey

How many years of experience do you have **cooking** in a home setting?

- 0 years
- 0-1 years
- 1-3 years
- 3-7 years
- 7+ years

How many years of experience do you have **baking** in a home setting?

- 0 years
- 0-1 years
- 1-3 years
- 3-7 years
- 7+ years

How many years of experience do you have **cooking** in a professional kitchen?

- 0 years
- 0-1 years
- 1-3 years
- 3-7 years
- 7+ years

How many years of experience do you have **baking** in a professional kitchen?

- 0 years
- 0-1 years
- 1-3 years
- 3-7 years
- 7+ years

Are you **currently** in a culinary arts degree/certificate program or professional apprenticeship?

- Yes
- No

Have you **completed** a culinary arts degree/certificate program or professional apprenticeship?

- Yes
- No

Describe your level of expertise in **cooking**

- Beginner
- Beginner-Intermediate
- Intermediate
- Intermediate-Advanced
- Advanced
- Expert

Describe your level of expertise in **baking**

- Beginner
- Beginner-Intermediate
- Intermediate
- Intermediate-Advanced
- Advanced
- Expert

How often do you **cook**?

- Rarely (0-1 times a week)
- Sometimes (1-3 times a week)
- Often (3+ times a week)

How often do you **bake**?

- Rarely (0-1 times a week)
- Sometimes (1-3 times a week)
- Often (3+ times a week)

On average, how often do you use recipes published online? Select your most frequent usage.

- 1+ times per week
- 1+ times per month
- 1+ times per year
- Never

For what reasons do you **read** the comments section of a recipe?

- I don't read the comments section
- When I have a question
- No specific reason or purpose when reading the comments section.
- Other

Why do you **post** to the comments section of a recipe?

- I don't make posts.
- When I have a question
- To answer another commenter's question
- To respond to other commenters (but not as an answer to a question)
- To generally publish some writing, such as to show appreciation or excitement about a recipe
- Other

Indicate your age range

- 18-24
- 25-34
- 35-49
- 50+

What is your highest level of education (or equivalent) completed?

- Some high school
- High school graduate
- Some college, no degree
- Associates degree
- Certificate program
- Apprenticeship
- Bachelors degree
- Graduate degree

(Optional) Anything else you would like to add?

Continue to experiment

Figure 7: Presurvey questions.

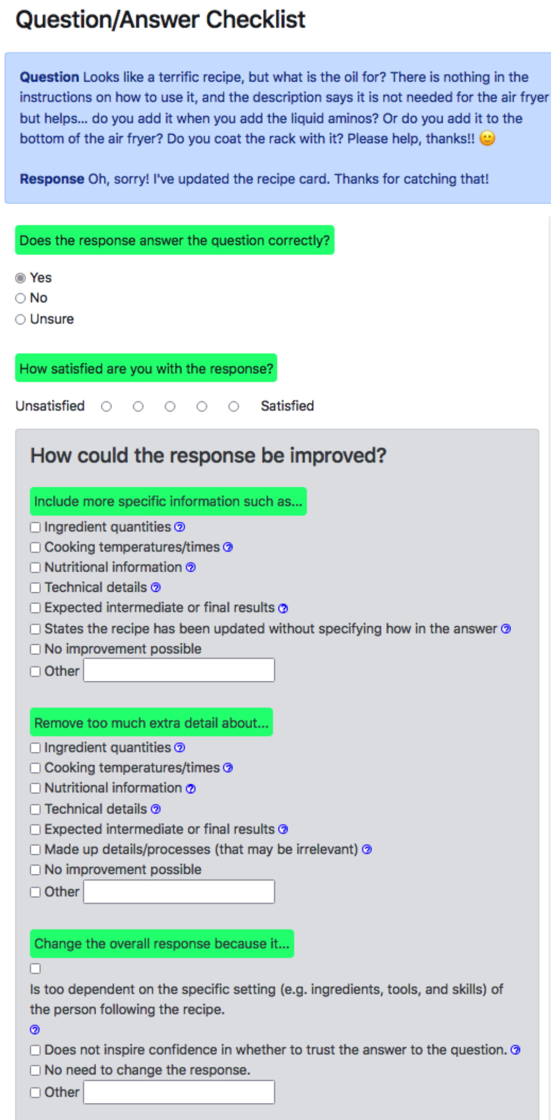


Figure 8: Screen shown when a response answers the question.

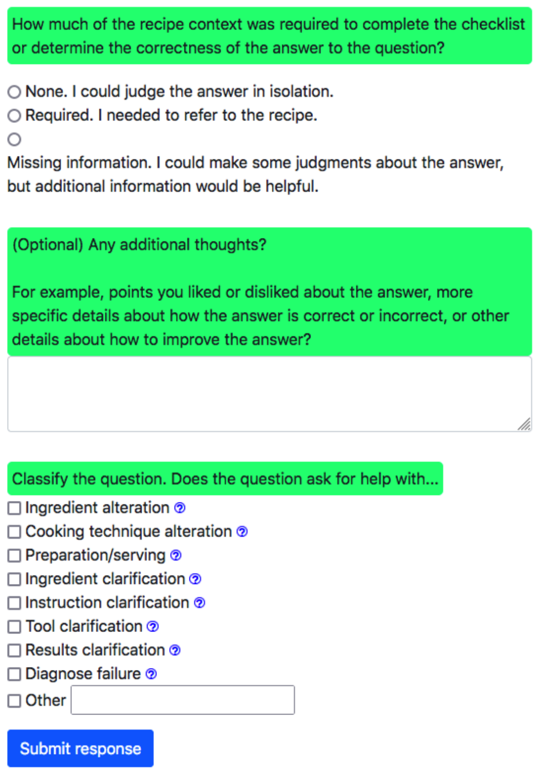


Figure 9: Screen shown when annotators state a response does and does not answer the question.

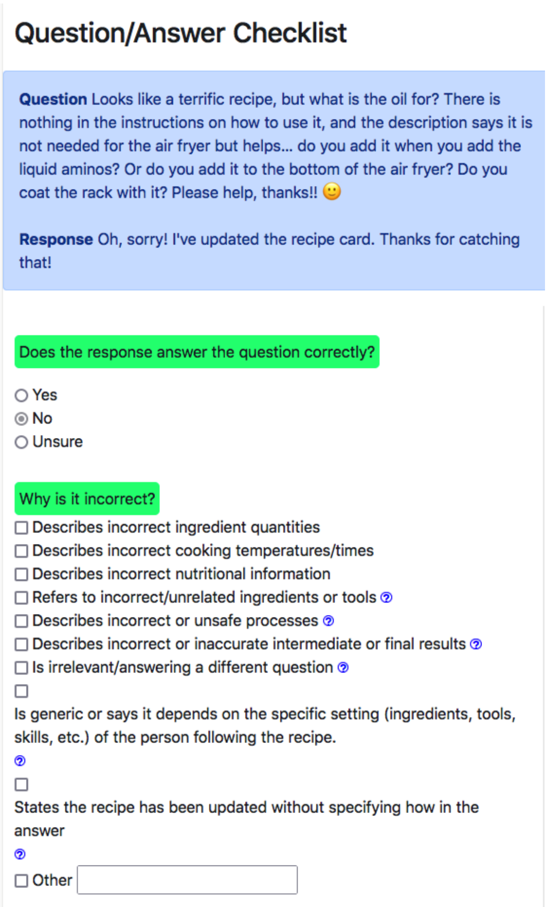


Figure 10: Screen shown when the response does not answer the questions.

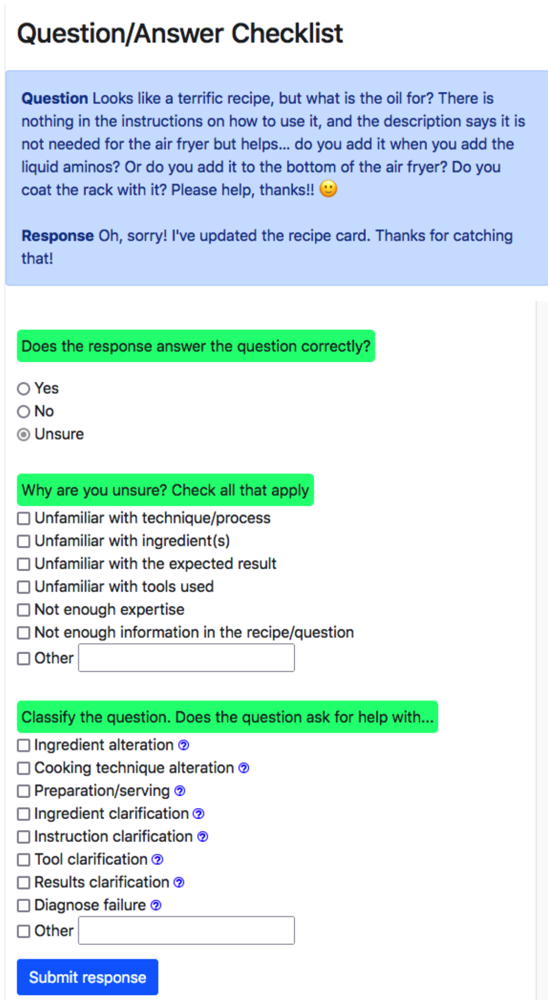
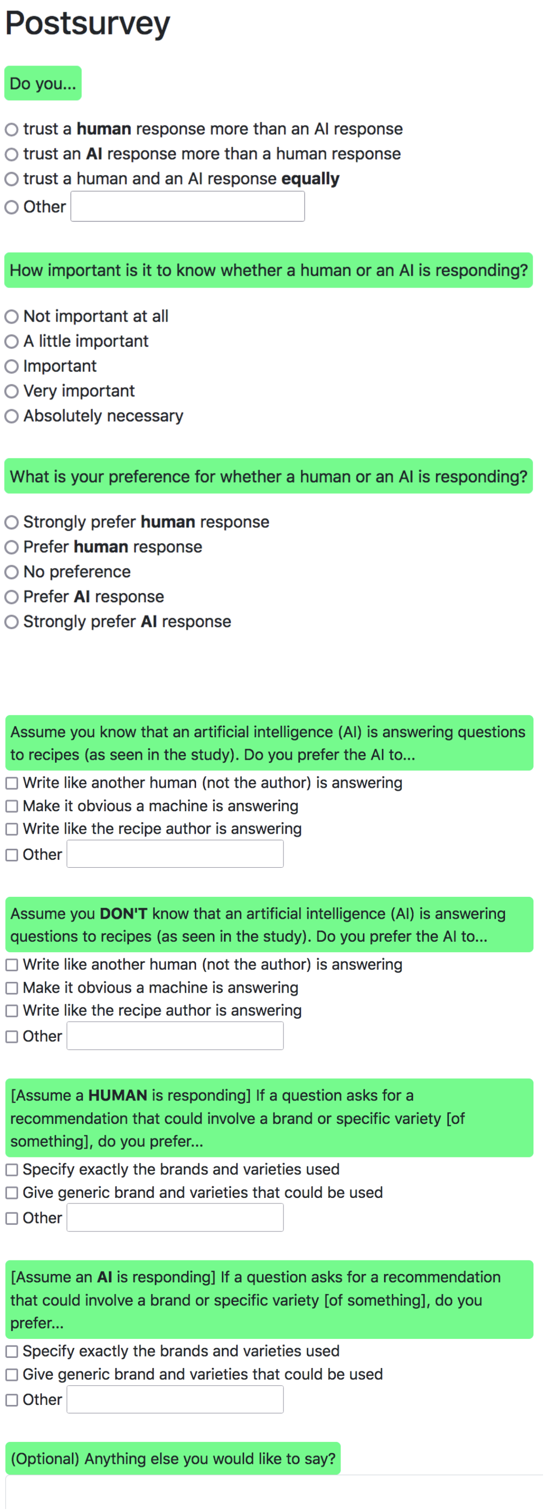


Figure 11: Screen shown if the annotator is unsure if the response answers the question.



Postsurvey

Do you...

- trust a **human** response more than an AI response
- trust an **AI** response more than a human response
- trust a human and an AI response **equally**
- Other

How important is it to know whether a human or an AI is responding?

- Not important at all
- A little important
- Important
- Very important
- Absolutely necessary

What is your preference for whether a human or an AI is responding?

- Strongly prefer **human** response
- Prefer **human** response
- No preference
- Prefer **AI** response
- Strongly prefer **AI** response

Assume you know that an artificial intelligence (AI) is answering questions to recipes (as seen in the study). Do you prefer the AI to...

- Write like another human (not the author) is answering
- Make it obvious a machine is answering
- Write like the recipe author is answering
- Other

Assume you **DON'T** know that an artificial intelligence (AI) is answering questions to recipes (as seen in the study). Do you prefer the AI to...

- Write like another human (not the author) is answering
- Make it obvious a machine is answering
- Write like the recipe author is answering
- Other

[Assume a **HUMAN** is responding] If a question asks for a recommendation that could involve a brand or specific variety [of something], do you prefer...

- Specify exactly the brands and varieties used
- Give generic brand and varieties that could be used
- Other

[Assume an **AI** is responding] If a question asks for a recommendation that could involve a brand or specific variety [of something], do you prefer...

- Specify exactly the brands and varieties used
- Give generic brand and varieties that could be used
- Other

(Optional) Anything else you would like to say?

Figure 12: Postsurvey.

B.2 Survey Results

We summarize survey results of annotators where conclusions were drawn.

Crowdworkers Six annotators had 7+ years of home baking and cooking experience, and all but one rated their cooking expertise as intermediate.

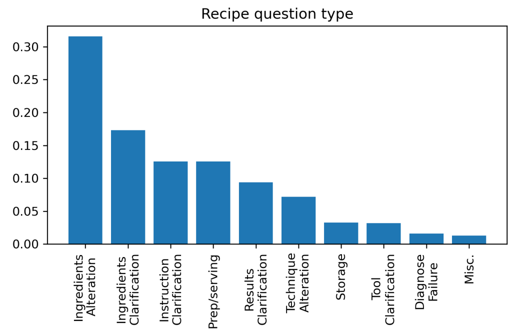


Figure 13: Distribution of the types of questions in the 60 questions annotated by both crowdworkers and experts. Misc. contains many infrequent custom question types. Storage was a frequently used custom type.

Baking expertise had a larger range from beginner (3) to intermediate (5). Two annotators requested generic brands when a known machine is responding. Three annotators want a machine to make it obvious that a machine is responding regardless of whether it is known that a machine is responding. There was an even split in trusting human and machine responses equally or trusting a human more.

Experts Two experts had 1–3 years of professional baking and cooking experience, and one expert had 3–7 years of professional cooking experience. Baking experience was rated intermediate by all three, two experts stated they had intermediate-advanced cooking expertise, and one considered themself an expert. One expert wanted generic brands from a known machine, and one expert did not mind having specific or generic brands. Only one expert wanted a machine to make it obvious that a machine is responding when the respondent is unknown, and and another expert wanted similarly when it is known that a machine is responding. Two experts state they trust human responses over an AI response, and one trusts both human and AI responses equally.

B.3 Improvement Results

Experts were more critical than crowdworkers for how responses could be improved and provided many custom suggestions for how to improve responses (Table 5).

B.4 Example Annotation Responses

Sample responses to a question in the annotation task (Figure 14) are as follows:

Area	AMT-GPT-3	%	AMT-Human	%	Expert-GPT-3	%	Expert-Human	%
Concise	Ingr. quantity	4.6	Cook temp/time	2.4	Cook temp/time	7.3	Ingr. quantity	6.9
	Tech. Detail	1.8	Tech. Detail	2.8	Tech. Detail	18.3	Tech. Detail	13.8
	Expected results	8.5	Expected results	2.9	Expected results	15.6	Expected results	8.6
	Other	6.8	Other	10.9	Other	19.2	Other	17.2
Verbose	Expected results	1.9	Expected results	1.5	Hallucination	4.2	Hallucination	4.3
	Other	3.2	Other	4.9	Other	5.3	Other	—
Misc	Hedging	2.8	Hedging	5.4	Hedging	15.2	Hedging	2.1
	Other	—	Other	—	Other	10.1	Other	6.4

Table 5: Annotations for how responses could be improved.

- Yes, the response answers the question correctly. The question type was for ingredient alteration and clarification. Recipe context was not required to judge the correctness of the answer. The answer could be improved in the "concise" dimension by including more Tech. Detail and information on why the technique is safe. The satisfaction with the answer is 4 out of 5.

- No, the response does not answer the question correctly. The question type is cooking technique alteration and preparation/serving. Recipe context was not required to judge the correctness of the answer. The answer was incorrect because it described incorrect or unsafe processes.

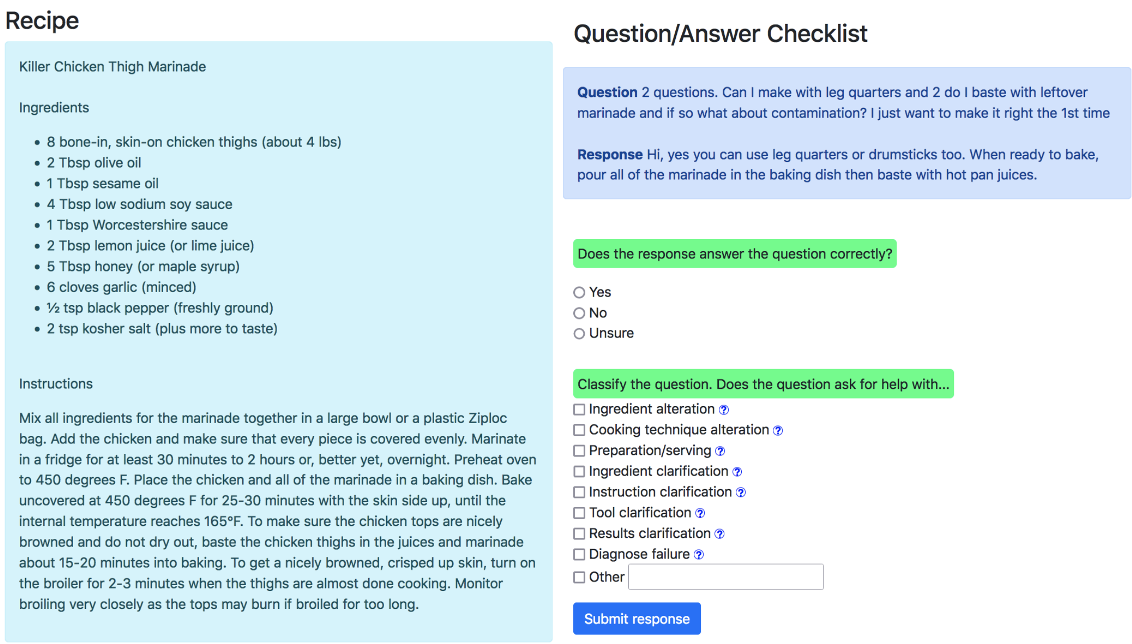


Figure 14: Example annotation question with responses described in B.4.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
6

☑ A2. Did you discuss any potential risks of your work?
6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
appendix b

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
appendix b.1, section 3.2

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
appendix b.1

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
3

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
appendix b.1