Characterizing LLM Abstention Behavior in Science QA with Context Perturbations

Bingbing Wen[1] **Bill Howe**[1] **Lucy Lu Wang**[1,2]
[1]University of Washington, [2]Allen Institute for AI
{bingbw, billhowe, lucylw}@uw.edu

Abstract

The correct model response in the face of uncertainty is to abstain from answering a question so as not to mislead the user. In this work, we study the ability of LLMs to abstain from answering context-dependent science questions when provided insufficient or incorrect context. We probe model sensitivity in several settings: removing gold context, replacing gold context with irrelevant context, and providing additional context beyond what is given. In experiments on four QA datasets with four LLMs, we show that performance varies greatly across models, across the type of context provided, and also by question type; in particular, many LLMs seem unable to abstain from answering boolean questions using standard QA prompts. Our analysis also highlights the unexpected impact of abstention performance on QA task accuracy. Counter-intuitively, in some settings, replacing gold context with irrelevant context or adding irrelevant context to gold context can improve abstention performance in a way that results in improvements in task performance. Our results imply that changes are needed in QA dataset design and evaluation to more effectively assess the correctness and downstream impacts of model abstention.[1]

1 Introduction

Question-answering (QA) in scientific settings is typically defined as a context-dependent task, where models answer questions based on provided context or relevant context it identifies. When the provided or retrieved context is itself unreliable or inconsistent, however, the correct model response should be to abstain from answering. Prior work has studied the abstention capabilities of LLMs (Yin et al., 2023; Amayuelas et al., 2023) and proposed approaches to improve abstention when presented with insufficient context (Zhou et al., 2023; Slobodkin et al., 2023), but these approaches focus on general domain question-answering (e.g. SQuAD2) against a narrow range of models (e.g. ChatGPT). As we will show in our experiments (§5.2), general domain settings produce highly divergent results from scientific QA settings, and different model architectures also exhibit differing abilities to abstain. Further, as new LLMs continue to be developed and released, we need an extensible way to measure their ability to abstain from answering questions when provided context is irrelevant or uncertain.

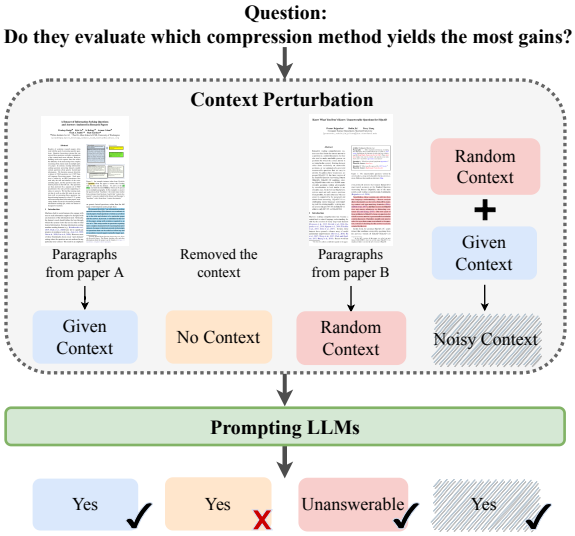

Figure 1: Our framework to probe the context sensitivity of LLMs for science QA. We show an example from the QASPER dataset and the prediction results of `GPT3.5` under different context perturbation settings. While the model fails to abstain when context is removed, it abstains appropriately when random context is provided.

To address these challenges, we introduce a framework to assess LLM abstention in science question-answering by removing, replacing, and augmenting provided contexts to control the answerability of questions. Using this framework, we probe abstention performance of four LLMs (`LLama2`, `Vicuna`, `Flan-T5`, and `GPT3.5`) on

[1]Code will be publicly available at https://github.com/bbwen/llm_scienceqa.

one general domain QA dataset (SQuAD2) and three scientific QA datasets (PubmedQA, BioASQ, QASPER). Several of these datasets include unanswerable questions by design, allowing us to analyze interactions between question answerability and noisy context. Our work aims to answer the following: (i) How well do LLMs abstain from answering questions when the correct context is *not* provided? (ii) How do context perturbations impact task performance and abstention performance? And (iii) How does question type impact task performance and abstention performance? We summarize our contributions below:

- We introduce a framework to study LLM abstention for science QA. Specifically, we probe models' ability to abstain from answering questions when the correct context is not provided, and how abstention is impacted by context perturbations and question type.
- Using this framework, we investigate the task and abstention performance of four LLMs on four QA datasets, ranging from general-domain factoid QA to context-sensitive, document-based science QA (§5). Our results show that no model consistently abstains in all settings where abstention is expected (unanswerable questions and no context/random context settings), though some models demonstrate a stronger ability to abstain for more context-dependent QA tasks, and instruction-tuned models are generally better at following abstention instructions.
- We investigate the impacts of context perturbations (§5.1) and question type (§5.2) on task accuracy and abstention ability. Substituting or augmenting context with random irrelevant context consistently facilitates better abstention performance across models and datasets, which can reflect as a counter-intuitive *improvement* in task performance. We also find that yes-no questions tend to interfere with all models' ability to abstain relative to other question types.

2 Related Work

Prompting LLMs for Science QA The few-shot capability of LLMs have been applied successfully to knowledge-intensive tasks like question answering (Wei et al., 2021; Chowdhery et al., 2022; Nori et al., 2023). Prompting strategies such as chain-of-thought (Wei et al., 2022), least-to-most (Zhou et al., 2022), and others (Kojima et al., 2022; Wang et al., 2022) significantly improve LLMs' zero or few-shot abilities on diverse QA benchmarks in the general (Rajpurkar et al., 2018; Yang et al., 2015) and scientific (Taylor et al., 2022; Pereira et al., 2023) domains. Pereira et al. (2023) adopt retrieve-than-read on the QASPER dataset (Dasigi et al., 2021), showing that current retrievers are the main bottleneck and readers (LLMs) are already performing at the human level. Inspired by this finding, we study how irrelevant or incorrect contexts (mimicking retrieval errors) can impact LLM performance on context-intensive QA tasks.

LLMs and Abstention in QA Several works have shown that LLMs become easier to steer with natural language prompts either as they become larger (Mishra et al., 2022a) or as they are exposed to more instruction tuning (Mishra et al., 2022b; Chung et al., 2022). Liao et al. (2022) introduced a prompt tuning strategy to improve performance on unanswerable questions, by mapping questions onto their specific templates. Other work tried to direct models toward abstention through data augmentation (Zhu et al., 2019). Asai and Choi (2021) provided an in-depth analysis of LMs' abstention ability, identifying paragraph selection and answerability prediction as two areas for improvement. Recent work introduced new datasets to study whether LLMs know what they do not know (Yin et al., 2023; Amayuelas et al., 2023), while Slobodkin et al. (2023) demonstrated that differences in LLM hidden states identify the boundary between known and unknown. In this work, we systematically characterize LLM abstention capabilities on context-dependent QA tasks.

Context Perturbation Prior work studying input perturbations for NLP tasks include approaches such as model-agnostic input transformations (Liang et al., 2022; Ravichander et al., 2022; Giorgi et al., 2022) and adversarial example generation (Jia and Liang, 2017; Wang et al., 2021). Liang et al. (2022) use semantics-preserving and semantics-altering perturbations in their robustness evaluation of LLMs. Pretrained language models can be negatively impacted by irrelevant context (Chowdhery et al., 2022; Liang et al., 2022), e.g., Shi et al. (2023) injected irrelevant numerical context for the MathQA dataset, after which ChatGPT performance dramatically decreased. However, Liang et al. (2022) evaluated T5 (Raffel et al., 2020) and PaLM (Chowdhery et al., 2022), demon-

Dataset	Context length (words)	Unans. proportion	Answer types	Test set size
SQuAD2	128	0.5	Ext	11873
PubmedQA	204	0.1	Bool	500
BioASQ	221	0.0	Bool	140
QASPER	149	0.1	Ext/Abs/Bool	1451

Table 1: QA Dataset statistics.

strating that finetuning these models with counterfactual and irrelevant contexts can improve model robustness to noisy context. In our framework, we leverage context perturbations to investigate the LLM abstention behavior for science QA, finding occasional counter-intuitive interactions between abstention and task performance.

3 Datasets

We conduct experiments on four QA datasets: SQuAD2 (Rajpurkar et al., 2018), PubmedQA (Jin et al., 2019), BioASQ (Nentidis et al., 2021), and QASPER (Dasigi et al., 2021). Dataset statistics are provided in Table 1. These datasets span general and science domains, and include extractive, abstractive, and boolean questions.

- SQuAD2 is a general-domain reading comprehension QA dataset. Answer contexts are extracted from Wikipedia.
- PubmedQA is a biomedical QA dataset. Questions are automatically derived from PubMed paper titles and are answered from the conclusion sections of the corresponding abstracts. All questions can be answered Yes/No/Maybe.
- BioASQ includes Yes/No questions that are formulated by biomedical experts, reflecting real-life information needs encountered during their work. Answers are provided by medical experts from paper abstracts.
- QASPER is a full document science QA dataset. Questions are written by domain experts and answers are annotated from the full text of associated computer science papers. Questions can be boolean, extractive, or abstractive, and multiple answers may be provided for each question.

Unanswerable questions Three of these datasets contain unanswerable questions (proportions in Table 1, examples in Table 2). SQuAD2 introduced unanswerable questions in machine reading comprehension; these unanswerable questions were curated by altering questions through negation, antonym swaps, entity swaps, mutual exclusion, impossible conditions, and other ways which make it such that the context paragraph no longer implies any answer. In SQuAD2, "unanswerable" questions imply *irrelevant* context passages. For PubmedQA, questions can be answered "yes", "no", or "maybe", and we interpret "maybe" as "unanswerable"; therefore, "unanswerable" in PubmedQA can be interpreted as answers with *high uncertainty* based on the given context. For QASPER, "unanswerable" questions are expert-labeled, and mean that *no answer is available* in the given document.

4 Framework

We describe our framework in terms of prompting strategies, context perturbation methods, the choice of models and handling of model output, and evaluation metrics.

4.1 Prompting strategies

We adopt prompting templates that achieved the best performance based on recent work (Pereira et al., 2023). We refer to these templates as *constrained prompts* since answer constraints (e.g., "Answer 'Yes' or 'No' for boolean questions") are added to achieve better task performance. For datasets that do not have boolean questions (e.g., SQuAD2), we do not include boolean answer constraints. The example prompt used for the QASPER dataset is the following:

> Create an Answer to the Question using following documents. **Pay attention to only answer "Yes" or "No" for boolean questions. Answer "Unanswerable" when you are not sure about the answer.**
> Context: {c}
> Question: {q}
> Answer:

Prompts for other datasets are in Appendix A.

We conduct zero-shot experiments. Given an input pair (c, q) where c is context and q is question, we prepend the constrained prompt instructions along with an explicit directive for handling unanswerable questions. Given complex interactions between model architecture, pretraining, instruction-tuning, dataset, question type, question answerability, context perturbations, and abstention, these settings confer significant complexity for our analysis. We therefore reserve the analysis of few-shot experiments and in-context learning to future work. We also conduct ablations with different prompting templates and abstention representations (results for these experiments in Appendix C).

Datasets	Unanswerable Examples
SQuAD2	Q: Who moved to Hollywood in 2004? C: ".... Following the move to **Holyrood in 2004** this building was demolished. The former Midlothian County Buildings facing Parliament Square..."
PubmedQA	Q: Does rugby headgear prevent concussion? C: "...In addition, coaches from all four levels were **questioned** about team policies and their personal opinions about the use of headgear to prevent concussion. Although the players tended to believe that the headgear could prevent concussion (62%), the coaches were **less convinced (33%)**..."
QASPER	Q: How many Universal Dependency features are considered? C: Empty.

Table 2: Example unanswerable questions from datasets (Q: question, C: context).

4.2 Context Perturbation

We conduct experiments to assess model sensitivity to context perturbations. We either provide the given context, or perturb the context by removing, replacing, or adding context passages (Figure 1).

Given context: We use the original/gold context provided by the dataset. For unanswerable questions in SQuAD2, given context is unchanged but the designers manually modified the question to render the context ineffective for inferring an answer. For PubmedQA, we label a question unanswerable if the given answer is "maybe" and do not change the context. For unanswerable questions in QASPER, given context is empty.

No context: We remove the given context.

Random irrelevant context: We replace the given context with the context from a random question in the train split.

Noisy context: We *append* context from a random question to the given context.

4.3 Models

We conduct experiments using LLamaV2-13b-chat (`LLama2`), Vicuna1.5-13b-chat (`Vicuna`), Flan-T5-XL (`Flan-T5`), and `gpt-3.5-turbo-0613` (`GPT3.5`). We select these models to have reasonable representation across the following attributes:

- Closed API (`GPT3.5`) vs open weights (`LLama2`, `Vicuna`, and `Flan-T5`)
- Encoder-decoder (`Flan-T5`) vs decoder-only (`LLama2`, `Vicuna`, and `GPT3.5`) architecture
- Models having less (`LLama2`) or more (`Vicuna`) instruction tuning

For all models, we use the same hyperparameters at inference time. We set temperature to 0 and top-p sampling to 1 to reduce the variability of model output (details in Appendix B).

Post-processing model output LLMs usually produce long outputs, so we post-process their outputs to obtain structured predicted answers. For boolean questions, we extract the first words from the model output. If these words contain "yes," "no," or "unanswerable," we map them to the corresponding classes.

4.4 Evaluation metrics

We report both task performance (F1 or Accuracy) and abstention performance (rate of abstention) across different experimental settings.

Task performance For SQuAD2 and QASPER, we evaluate *task performance* using n-gram F1 as reported in Rajpurkar et al. (2018). For PubmedQA and BioASQ (boolean questions only), we report Accuracy based on the original papers. Task performance is therefore comparable under different context settings, but is *not* comparable across datasets.

Abstention performance We measure model *abstention performance* by calculating the proportion of questions that is answered "unanswerable." Ideally, models that are perfect at abstaining would be expected to have an abstention proportion of 1.0 when no or irrelevant context is provided for all questions. However, in reality this is not the case, since many questions may be context-independent or could be answered using model parametric knowledge. Abstention is therefore a quality dependent on (i) model ability to follow instructions ("Answer 'Unanswerable' when you are not sure"), (ii) question context dependency (whether a question is answerable without context), and (iii) question difficulty (whether the question is answerable using model parametric knowledge). For clarity, we show task performance as plain numbers and abstention rates and deltas using (parentheses).

5 Results

We report baseline task performance (Table 3) and abstention performance (Table 4) for all models on all datasets. Results for context perturbations

	SQuAD2	PubmedQA	BioASQ	QASPER
SOTA	90.5*	77.6**	94.3**	61.4†
(a) Given context				
LLama2	51.7	52.6	**98.6**	16.8
Vicuna	61.0	36.4	93.6	30.5
Flan-T5	**87.4**	**73.2**	97.8	**60.1**
GPT3.5	60.4	61.2	97.9	57.8
(b) No context				
LLama2	-11.7	-17.8	-43.6	-2.7
Vicuna	-24.0	-19.4	-32.9	-10.4
Flan-T5	-38.2	-16.4	-30.8	-38.0
GPT3.5	-22.9	-39.0	-46.4	-37.7
(c) Random context				
LLama2	-1.9	-40.8	-63.6	-0.8
Vicuna	-11.8	-25.2	-41.5	-11.5
Flan-T5	-37.3	-59.4	-63.5	-37.2
GPT3.5	-10.4	-50.0	-64.9	-37.6
(d) Noisy context				
LLama2	-4.3	-13.2	-10.7	+1.9
Vicuna	-1.4	+4.8	-0.7	+4.7
Flan-T5	+0.0	-0.2	-0.7	+0.0
GPT3.5	-2.4	-2.8	-2.2	-2.4

Table 3: Model zero-shot task performance using constrained prompts. SOTA indicates the best zero-shot performance of LLMs reported in previous papers (*=Flan-UL2 (Slobodkin et al., 2023), **=Galactica (Taylor et al., 2022)), or best performance from a pretrained LM (†=UnifiedQA-large (Dasigi et al., 2021)). Best performances **bolded**, second best underlined among baseline performance (a). Colors indicate positive or negative delta from baseline task performance with different context perturbations: (b) no context, (c) random context, and (d) noisy context. While task performance generally degrades with context perturbations, this is not the case for QASPER due to interactions between abstention and task performance (see Section 5.1).

are shown as changes to task performance and abstention performance from baseline. Additional analysis by question type is shown in Figure 2.

Baseline task performance in the zero-shot prompting setting is presented in Table 3(a), alongside previously reported SOTA zero-shot performance of LLMs on each dataset. Surprisingly, Flan-T5 achieves the best performance on SQuAD2, PubmedQA and QASPER, LLama2 performs best on BioASQ. GPT3.5 achieves the second best performance (very close to the best one) on PubmedQA, BioASQ and QASPER. The task performance of tested models are close to and sometimes exceed the reported SOTA on each dataset. Though in this work, we focus on assessing the change to task performance under different context settings rather than maximizing model performance.

5.1 Impact of context perturbations

Random context facilitates abstention In Table 4(c), we observe that all models are much more likely to abstain from answering when given random context except on the BioASQ dataset. For SQuAD2 and PubmedQA, random context improves most models' abstention performance to close to 1. Specifically, random context improves Flan-T5's ability to abstain from answering on PubmedQA while the no context setting fails to do so. For QASPER, the abstention performance on answerable questions in the random context setting is much higher compared to no context, but is still far from 1. Oddly, for boolean questions in BioASQ, random context biases the model towards answering "no" rather than "unanswerable", and we observe no changes in abstention rate (further analysis in Section 6).

Adding noisy context can counter-intuitively improve task performance on some datasets While noisy context has mixed effects on abstention (Table 4(d)), this perturbation does not always translate to negative impacts on task performance. In Table 3(d), task performance increases on QASPER for LLama2, Vicuna, and Flan-T5. Since QASPER's unanswerable questions have empty context, the baseline task performance for unanswerable questions is actually computed with no context; perturbing no context by adding noisy context therefore leads to a trade-off in task performance and abstention performance (some models abstain much more for unanswerable questions and task performance consequently improves).

5.2 Impact of question type

Models' abstention capabilities vary by question type For extractive questions, Figure 2(a) and (d) show that model abstention varies significantly between given context and no context settings; under no context settings, models can achieve abstention performance close to 1. Abstractive answers show similar patterns. Notably, we find *all* models are reluctant to abstain from boolean questions for all datasets; this overconfidence is apparent even for QASPER, which is a highly document-specific QA dataset. Surprisingly, Flan-T5 and LLama2 demonstrate near zero abstention performance on boolean questions on any dataset, regardless of

Baseline	SQUAD2		Pubmed		BioASQ	Qasper	
Model	Ans.	Unans.	Ans.	Unans.	Ans.	Ans.	Unans.
LLama2	(44.7)	(82.0)	(37.3)	(34.5)	(0.0)	(3.1)	(30.9)
Vicuna	(21.1)	(64.9)	(61.8)	(60.0)	(0.0)	(30.0)	(57.0)
Flan-T5	(4.4)	(85.0)	(0.0)	(0.0)	(0.0)	(13.0)	(87.0)
GPT3.5	(3.7)	(53.3)	(20.4)	(21.8)	(0.0)	(15.0)	(89.2)
(b) No context							
LLama2	(+14.8)	(-11.7)	(-7.0)	(-3.6)	(+2.8)	(+17.7)	(+3.7)
Vicuna	(+28.3)	(-4.7)	(+26.7)	(+27.3)	(+10.7)	(+38.2)	(+20.7)
Flan-T5	(+91.2)	(+11.3)	(+0.0)	(+0.0)	(+0.0)	(+65.0)	(-2.8)
GPT3.5	(+37.1)	(+2.5)	(+57.6)	(+56.4)	(+36.4)	(+73.8)	(+8.4)
(c) Random context							
LLama2	(+54.8)	(+17.2)	(+58.7)	(+60.0)	(+0.0)	(+16.9)	(+17.8)
Vicuna	(+73.8)	(+31.7)	(+37.5)	(+40.0)	(+0.0)	(+44.0)	(+34.5)
Flan-T5	(+95.4)	(+15.0)	(+91.5)	(+92.7)	(+0.0)	(+54.6)	(-0.9)
GPT3.5	(+95.5)	(+45.9)	(+79.6)	(+78.2)	(+0.0)	(+66.9)	(+5.5)
(d) Noisy context							
LLama2	(+22.8)	(+3.2)	(+18.7)	(+18.2)	(+0.0)	(+2.2)	(+17.4)
Vicuna	(+8.2)	(+2.5)	(-11.0)	(-12.7)	(+0.0)	(+13.2)	(+34.7)
Flan-T5	(-0.9)	(-1.3)	(+0.0)	(+0.0)	(+0.0)	(-1.4)	(-2.7)
GPT3.5	(+1.5)	(-2.8)	(+3.2)	(+0.0)	(+0.0)	(+4.5)	(+4.6)

Table 4: Abstention performance across different models broken down by **answerable** and **unanswerable** questions across various datasets. Baseline abstention rates are shown at the top with a gray background. Abstention rates under (b) No context, (c) Random context, and (d) Noisy context settings are shown below as deltas from the base rate. Colors indicate a positive or negative delta. All context perturbations improve model abstention performance in some settings, though this is not uniform over datasets, models, question answerability, or perturbation setting.

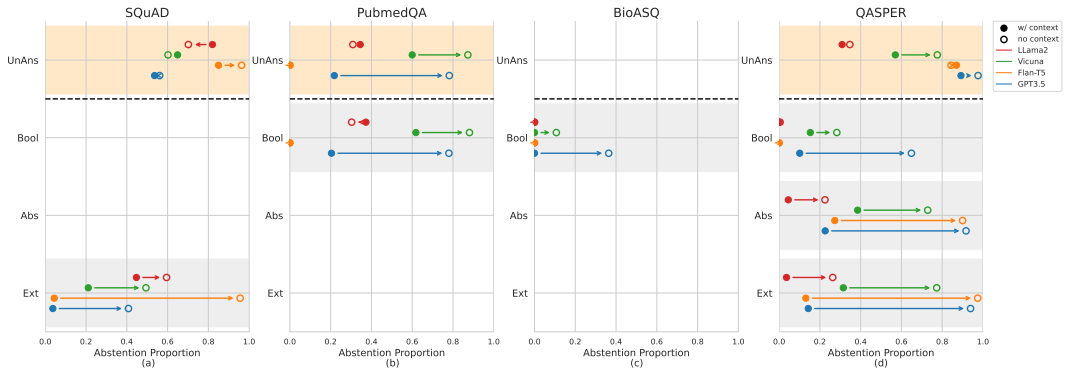

Figure 2: Abstention performance changes from "with context" to "no context" settings across different question types. Each row represents one question type, from top to bottom: "Unanswerable", "Boolean", "Abstractive" and "Extractive". White background indicates a dataset does not have questions of that type.

whether or not context is provided.[2] `Vicuna` and `GPT3.5` perform only slightly better on this front.

Abstention performance on unanswerable questions varies by dataset In Figure 2, we examine the abstention performance on unanswerable questions (yellow background) and find that performance varies significantly across datasets and models. For SQuAD2, `GPT3.5` performs consistently regardless of whether context is provided, indicating that `GPT3.5` may be able to recognize the irrelevance of the context. Surprisingly, `LLama2` and `Vicuna` abstain less when context is removed, which contradicts intuition. In SQuAD2, unanswerable questions are still related to the provided context; which may explain why when the confusing context is removed, these two models be-

[2]We conduct prompt ablations in Appendix C, showing the lack of abstention on boolean questions is highly sensitive to prompt wording. Without a boolean instruction in the prompt, both models abstain well for boolean questions.

Model	SQuAD2	PubmedQA	BioASQ	QASPER
LLama2	29.8	16.6	2.8	20.6
Vicuna	33.6	29.0	10.7	44.9
Flan-T5	**91.2**	0.0	0.0	66.0
GPT3.5	38.8	**59.3**	**36.4**	**75.7**

Table 5: % of answerable questions for which the model changes from answering to abstaining when context is removed. Best performances **bolded**, second best underlined.

come more likely to answer. For PubmedQA, unanswerable questions have high-uncertainty context; Flan-T5 and LLama2 consistently refuse to abstain from answering these questions, which are boolean, while Vicuna follows the "unanswerable" instruction quite well. For QASPER, abstention performance is generally higher across all models compared to SQuAD2 and PubmedQA since the questions are more document-grounded.

5.3 Abstention performance of LLMs

Instruction-tuned models more readily abstain Table 5 shows the proportion of answers for which the model changed from answering to abstaining when context is removed. The expectation for context-dependent questions would be that a model would abstain from answering if no context is provided. GPT3.5 achieves the largest abstention changes on PubmedQA, BioASQ, and QASPER, indicating that it is responsive to context loss. Vicuna, as an instruction-tuned model, also demonstrates relatively strong abstention behavior in these cases, though to a lesser degree than GPT3.5 for all datasets. This effect is contrary to LLama2, which abstains at a similar rate regardless of whether context is provided.

Flan-T5 readily abstains for non-boolean questions Flan-T5 abstains well when context is removed (Table 5). We observe large changes for SQuAD2 and QASPER (comparable or better than GPT3.5), although this behavior does not generalize *at all* to boolean questions—the change is 0.0 for PubmedQA and BioASQ, which consist only of boolean questions. Different prompting strategies may be necessary to enjoin certain models to abstain when answering boolean questions.

6 Error Analysis

We perform error analysis on the most consistently high performing model in our analysis, GPT3.5, and summarize main reasons why the model may fail to abstain. We sample 20 failure cases each from among answerable questions and unanswerable questions, for both gold context and no context settings. We present our findings by dataset.

SQuAD2 Answerable questions are either general (e.g., "What German poet was descended from Huguenots?") or could be answered without context (e.g., "What types of pumps are typically used in industrial boilers?"); the model tends to answer these regardless of whether context is provided. For unanswerable questions, half of failure cases are very open-ended like "What effect do technologies and resources generate?" and removing context will cause GPT3.5 to answer rather than abstain. Another 20% of questions contain popular entities but have no correct answer, e.g., "What lake contains part of the Rhine Falls?" misleads the model to generate "Lake Constance" as an answer since "The Rhine emerges from Lake Constance."

PubmedQA GPT3.5 is strongly inclined to answer "yes" for all failure cases among both answerable and unanswerable questions under the no context setting, perhaps due to the phrasing of automatically constructed questions.

BioASQ Around 90% of failure cases we sampled were answered correctly by GPT3.5 without context, including both "yes" and "no" answers. Another 10% of cases are affected by the model's tendency to hallucinate "yes", which is consistent with cases in PubmedQA. We manually substitute antonyms for the word in brackets on failure cases such as "Is Apelin usually [decreased] in diabetes?", "Does vesatolimod [inhibit] TLR7?" etc., and GPT3.5 still always answers "yes". Interestingly, for these factoid questions, substituting gold with random context skews the model strongly towards answering "no". These two types of hallucination behavior require further investigation. Our observations also raise concerns for assessing LLM performance using QA benchmarks containing boolean questions.

QASPER For answerable questions, 50% of failures seem to be caused by GPT3.5's ignorance of the ambiguity resulting from anaphora; for instance, GPT3.5 should not be able to resolve terms in questions such as "they","this study", and "the models" without context, yet the model answers these questions anyway under the no context setting. Around 30% of questions are very general, such as "What

	SQuAD2	QASPER
Context	Computational complexity theory is a branch of the theory of computation in theoretical computer science that focuses on classifying computational problems according to their inherent difficulty, and relating those classes to each other...	Table TABREF35 show the comparisons between tree and sequential based methods. We can see that, if we don't deploy CNN, simple Tree LSTM yields better result than traditional LSTM, but worse than Bidirectional LSTM...
Prompt	Q: What is a manual application of mathematical steps?	Q: Do they separately evaluate performance of their learned representations (before forwarding them to the CNN layer)?
Given context	Unanswerable	No ✗
No context	Calculation ✗	Yes ✗
Random context	Unanswerable	Unanswerable
Noisy context	Unanswerable	Unanswerable ✗
Ground truth	Unanswerable	Yes

Table 6: Examples of `GPT3.5` predictions under different context perturbations. For SQuAD2, removing context results in the model no longer abstaining, and responding inaccurately. For QASPER, the model answers the question incorrectly in the gold context setting; removing context results in an accurate but incorrect answer (the model should abstain in this case instead).

is a soft label?" and "Why is NER for tweets more challenging as the number of entities increases?"—these questions lead the model to answer rather than abstain. For unanswerable questions, the distribution of failure reasons is similar.

Table 6 shows two context perturbation examples with answers generated by `GPT3.5`. More examples are given in Appendix D.

7 Discussion & Conclusion

Our study investigates the impacts of context removal and perturbation on LLM performance on scientific QA. While lack of correct context should result in a model abstaining from answering a question, our results highlight that there are inconsistent patterns of model behavior based on model pretraining paradigms, question types, and the context-dependence of various QA datasets. For example, perturbing given context by substituting with random irrelevant context or adding noisy context would be expected to reduce task performance, but in some cases, the improvements in abstention that result can negate any reductions in task performance and potentially lead to gains (e.g., `LLama2`, `Vicuna`, and `Flan-T5` on QASPER). Additionally, we find that abstention varies greatly by question type, with all models in our experiments struggling to abstain on boolean questions.

Future work for enhancing models' abstention ability could investigate the impact of (i) **Different prompting strategies**: since model abstention ability is sensitive to constrained prompts (such as boolean instructions) as shown in Section 5.2 and Appendix C, how to select good prompting strategies to produce the best trade-off between task performance and abstention performance remains an interesting problem; (ii) **Alternate model architectures**: smaller models such as Flan-T5-XL with encoder-decoder architectures performed comparably to larger decoder-only models such as ChatGPT. Further exploration of encoder-decoder architectures or introducing an auxiliary module to foster understanding of context may be helpful; (iii) **Other context perturbations**: we show initial results that providing noisy context can counter-intuitively improve task performance in some datasets due to interactions with unanswerable questions. How this interacts with retrieval errors that occur in open-domain QA is a direction that could be explored in future work.

Our results have direct implications for dataset curators and model developers. Benchmark QA datasets with mixtures of unanswerable and answerable questions were designed to facilitate assessment of abstention ability, yet our experiments show conflation and a lack of clear assessment of either abstention ability or performance accuracy. In other words, while unanswerable questions were motivated by the need to measure abstention, when aggregated with other questions during task performance evaluation, this distinction is lost. Additionally, unanswerable questions in different datasets measure different phenomena, e.g., in SQuAD2, they measure model sensitivity to irrelevant context passages, in QASPER, they indicate that a question is unanswerable based on the given document, neither of which clearly map to the notion of abstaining under insufficient information. Given the importance of assessing model abstention capabilities, separating task and abstention assessment, coupled with changes in dataset construction, is

needed to better align model performance on these tasks with human expectations.

Beyond dataset curation, other variables—such as the task defined by the dataset, the types of questions posed, the architecture of models, instruction tuning techniques, in-context learning, and domain-specific pretraining—can affect a model's ability to effectively abstain and provide accurate answers. In this work, we attempt to disentagle some of these factors, though not all. Future studies could explore the extent to which in-domain pretraining and abstention-specific instruction tuning techniques impact model abstention performance.

On the other hand, our results also hold implications for downstream builders/users of interactive systems who rely on LLMs as question-answering tools. We show that in some cases (e.g., boolean questions), LLMs exhibit little to no ability to abstain, even for highly context-dependent questions such as those in QASPER. System designers should be cautious of these limitations and when/how to infer model uncertainty and communicate this uncertainty to users. For communicating abstention capabilities to users, confidence scores or similar indicators that reflect model certainty, accompanying explanations, or guidance on how to interpret model answers can empower users to make more informed decisions.

Limitations

This work evaluates LLM abstention in situations with incorrect or noisy context, which mimics retrieval errors in retrieval-augmented systems. The problem of abstaining is also crucial in other settings, such as from the fairness, privacy/copyright, and safety perspectives, which we do not address in this work. In the future, we plan to propose a unified framework for abstention evaluation that considers these settings as well.

The interplay between abstention and task performance can only be studied in datasets with unanswerable questions, which is scarce in the landscape of QA datasets. We only conduct experiments on three such datasets, so our results may have difficulty generalizing. Additionally, we note the different ways that unanswerable questions were constructed among these datasets, which imply that our experiments may be measuring different notions of "unanswerability." In addition, although question type diversity is common in real application, there are insufficients numbers of QA datasets with question type diversity. Therefore, our results may not generalize, as several question types are only represented in one or two of our datasets. We also note the sensitivity of LLMs to prompt phrasing, as with constrained and free-form prompts for boolean questions, and emphasize that it may be difficult to acquire consistent results without significant efforts made in prompt engineering.

Due to the budget constraints, we could not conduct experiments with every single model, dataset, or context perturbation method from existing work, though our framework can be applied to other datasets and models. We will make our framework publicly available such that new models and datasets, as well as context perturbations could be added. Other factors affecting abstention such as in-domain pretraining and in-context learning were also not studied in this work and could be incorporated in the future.

References

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*.

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1492–1504, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022. Open domain multi-document summarization: A comprehensive study of model brittleness under retrieval. In *EMNLP Findings*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Jinzhi Liao, Xiang Zhao, Jianming Zheng, Xinyi Li, Fei Cai, and Jiuyang Tang. 2022. Ptau: Prompt tuning for attributing unanswerable questions. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1219–1229.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022a. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022b. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Luis Gasco, Martin Krallinger, and Georgios Paliouras. 2021. Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12*, pages 239–263. Springer.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Jayr Pereira, Robson Fidalgo, Roberto Lotufo, and Rodrigo Nogueira. 2023. Visconde: Multi-document qa with gpt-3 and neural reranking. In *European Conference on Information Retrieval*, pages 534–543. Springer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

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.

Abhilasha Ravichander, Matt Gardner, and Ana Marasović. 2022. Condaqa: A contrastive reading comprehension dataset for reasoning about negation. *arXiv preprint arXiv:2211.00295*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093*.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625, Singapore. Association for Computational Linguistics.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4238–4248, Florence, Italy. Association for Computational Linguistics.

A Prompting templates

Prompts with context Prompts for other datasets are provided below. For SQuAD2, we do not include a boolean directive as the dataset contains no boolean questions.

Example prompt for SQuAD2:

> Create a shortest Answer to the Question using the following documents. **Answer Unanswerable when you are not sure about the answer**. Please only output the exact answer and keep the answer concise.
> Context: {c}
> Question: {q}
> Answer:

Example prompt for PubmedQA:

> Create an Answer to the Question using following documents. **Pay attention to answer only "yes","no" or "Unanswerable". Answer "Unanswerable" when you are not sure about the answer.**
> Context: {c}
> Question: {q}
> Answer:

Example prompt for BioASQ:

> Create an Answer to the Question using following documents. **Pay attention to answer only "yes" or "no". Answer "Unanswerable" when you are not sure about the answer.**
> Context: {c}
> Question: {q}
> Answer:

Prompts without context

The prompting template for no context is different from the original. We remove the instruction expression of "use the documents" since no documents are provided. This is the example prompt used for QASPER when context is removed:

> Create an Answer to the Question. **Answer "Yes" or "No" for boolean questions. Answer "Unanswerable" when you are not sure about the answer.**
> Context: {c}
> Question: {q}
> Answer:

Correspondingly, we remove this instruction phrase from the prompts for SQuAD2, PubmedQA, and BioASQ.

B Experiment Details

We run all experiments on an A100 GPU with 40GB of memory for `LLama2`, `Vicuna` and `Flan-T5`. We set max context length to 3096 and max new tokens to 256. For all models, evaluation batch size is 2. We randomly sample 10% of SQuAD2 test dataset in this work.

C Ablation Study

We conduct experiments removing constraints from boolean question prompts. We refer to the resulting prompts as "free-form" prompts. The example used for QASPER is given below:

> Create an Answer to the Question using following documents. **Answer "Unanswerable" when you are not sure about the answer.**
> Context: {c}
> Question: {q}
> Answer:

	PubmedQA	BioASQ	QASPER
(a) Constrained prompt			
`LLama2`	(30.3)	(2.8)	(0.5)
`Vicuna`	(89.5)	(10.7)	(28.4)
`Flan-T5`	(0.0)	(0.0)	(0.0)
`GPT3.5`	(77.7)	(36.4)	(97.8)
(a) Free-form prompt			
`LLama2`	(50.3)	(55.7)	(21.6)
`Vicuna`	(83.3)	(54.2)	(8.9)
`Flan-T5`	(84.0)	(80.0)	(92.8)
`GPT3.5`	(48.1)	(29.3)	(43.1)

Table 7: Abstention performance comparison on boolean questions using constrained prompt and free-form prompt under no context setting. Free-form prompt enables `Flan-T5` to successfully abstain across three datasets.

	PubmedQA	BioASQ
(a) Unanswerable		
`LLama2`	(30.3)	(2.8)
`Vicuna`	(89.5)	(10.7)
`Flan-T5`	(0.0)	(0.0)
`GPT3.5`	(77.7)	(36.4)
(a) Maybe		
`LLama2`	(2.6)	(0.0)
`Vicuna`	(9.0)	(0.0)
`Flan-T5`	(0.0)	(0.0)
`GPT3.5`	(70.9)	(40.2)

Table 8: Abstention performance comparison on boolean questions using different abstention representation ("Unanswerable" vs "Maybe") under no context setting. "Unanswerable" enables models to abstain across three datasets.

Free-form vs. constrained prompts for boolean questions The free-form prompt consistently improves the abstention ability of smaller models on boolean questions across datasets as shown in Table 7. Compared with zero abstention performance (obtained under the contrained prompt), the free-form prompt enables `Flan-T5` to successfully abstain across three datasets; in fact, `Flan-T5` demonstrates the best abstention ability for boolean questions under the free-form prompt setting across

	BioASQ	PubmedQA
Context	BCL11B mutations in patients affected by a neurodevelopmental disorder with reduced type 2 innate lymphoid cells. Using massively parallel sequencing we identified 13 patients bearing heterozygous germline alterations in BCL11B. Notably, all of them are affected by global developmental delay with speech impairment and intellectual disability; **however, none displayed overt clinical signs of immune deficiency.** Six frameshift mutations, two nonsense mutations, one missense mutation, and two chromosomal rearrangements resulting in diminished BCL11B expression, arose de novo...	...Postoperative recovery rates in the surgery group at 1 week and 4 weeks were -7.4% and -1.1%, respectively. Only 5 cases had showed clinical improvement, and the condition of these 5 patients had worsened again at averaged 7.4 weeks after surgery. Postoperative oral steroid therapy was initiated at an average of 6.4 weeks and the average initial dose was 54.0 mg in the surgery group, while 51.3 mg in the nonsurgery group. The recovery rate of the Japanese Orthopedic Association score, which increased after steroid therapy, **was better in the nonsurgery group (62.5%) than in the surgery group (18.6%) with significant difference (P<0.01)**.
Prompt	Q: Is there a link between BCL11B haploinsufficiency and syndromic neurodevelopmental delay?	Q: Is decompressive surgery effective for spinal cord sarcoidosis accompanied with compressive cervical myelopathy?
Given context	Yes ✗	Unanswerable ✗
No context	Yes ✗	Yes ✗
Random context	No ✗	Unanswerable
Noisy context	Yes ✗	Yes ✗
Ground truth	No	No

Table 9: Examples of `GPT3.5` predictions under different context perturbations for BioASQ and PubmedQA. `GPT3.5` tends to hallucinate responses for boolean questions when no or incorrect context is provided. For BioASQ, the model answers the question incorrectly in the gold context setting; removing context results in the model skewing to "yes" and random context results in the model answering "no", both cases should be "unanswerable" ideally. For PubmedQA, the model answers the question incorrectly in the gold context setting; removing context results in hallucination to "yes" (the model should abstain in this case instead).

datasets. For `LLama2`, the free-form prompt also notably increases its abstention performance.

Different ways to represent abstention We further ablate the wording used to represent abstention in our prompt (results in Table 8). Instead of using "unanswerable", we also try the term "maybe" in the prompt. We find that using "maybe" makes models less likely to abstain. This may be due to "unanswerable" being used more frequently in training corpora to represent abstention.

D Further error analysis

Table 9 presents example failure cases from BioASQ and PubmedQA.