

Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering

Vaibhav Adlakha^{1,2}

Parishad BehnamGhader^{1,2,*}

Xing Han Lu^{1,2,*}

Nicholas Meade^{1,2,*}

Siva Reddy^{1,2,3}

¹Mila – Quebec AI Institute

²McGill University

³Facebook CIFAR AI Chair

{firstname.lastname}@mila.quebec

Abstract

Retriever-augmented instruction-following models are attractive alternatives to fine-tuned approaches for information-seeking tasks such as question answering (QA). By simply prepending retrieved documents in its input along with an instruction, these models can be adapted to various information domains and tasks without additional fine-tuning. While the model responses tend to be natural and fluent, the additional verbosity makes traditional QA evaluation metrics such as exact match (EM) and F1 unreliable for accurately quantifying model performance.

In this work, we investigate the performance of instruction-following models across three information-seeking QA tasks. We use both automatic and human evaluation to evaluate these models along two dimensions: 1) how well they satisfy the user’s information need (correctness), and 2) whether they produce a response based on the provided knowledge (faithfulness). Guided by human evaluation and analysis, we highlight the shortcomings of traditional metrics for both correctness and faithfulness. We then propose simple token-overlap based and model-based metrics that reflect the true performance of these models. Our analysis reveals that instruction-following models are competitive, and sometimes even outperform fine-tuned models for correctness. However, these models struggle to stick to the provided knowledge and often hallucinate in their responses. We hope our work encourages a more holistic evaluation of instruction-following models for QA. Our code and data is available at <https://github.com/McGill-NLP/instruct-qa>

1 Introduction

One of the goals of natural language processing (NLP) is to enable systems to perform tasks based on natural language instructions as this would empower users to interact in an intuitive and flexi-

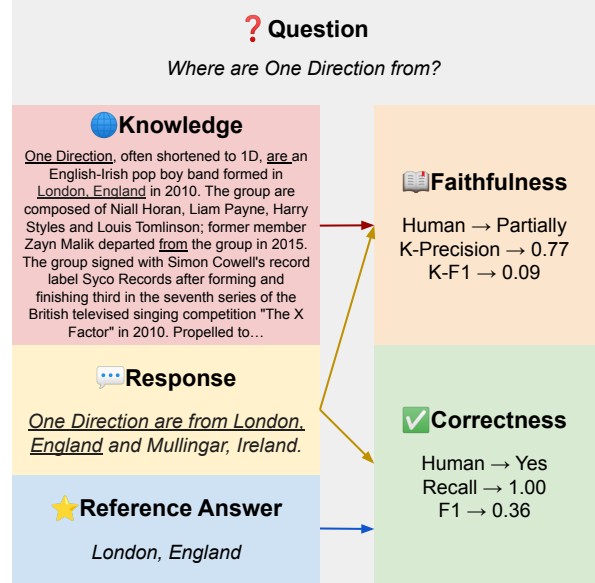


Figure 1: Sample response generated by GPT-3.5. The model response is correct w.r.t information need but only partially faithful w.r.t knowledge as only one of the two locations mentioned in the response can be found in the knowledge (truncated for readability). Recall (§4.2) and K-Precision (§5.1) are automatic metrics that approximate human judgment.

ble manner. Instruction-following models are a type of language models that aim to achieve this goal. Training these models usually involves exposing large language models (LLMs; Brown et al. 2020; Zhang et al. 2022; Thoppilan et al. 2022; Rae et al. 2022; Touvron et al. 2023a) to thousands of tasks formulated as natural language instructions through supervised examples (Sanh et al., 2022; Wei et al., 2022; Chung et al., 2022; Ouyang et al., 2022a; Iyer et al., 2023; Touvron et al., 2023b) or other forms of supervision (Ouyang et al., 2022b; Wang et al., 2022a; Taori et al., 2023; Peng et al., 2023). These are known to generalize to many tasks with little exposure to examples of those tasks (Mishra et al., 2022). In this paper, we evaluate instruction-following models for their ability to perform question-answering (QA) on a given set of text passages.

*Core contributor.

Instruction-following models can perform QA when provided with a prompt describing the task, the question, and relevant text passages to reason upon retrieved by a retriever (Chung et al., 2022). These model-generated answers are known to be natural, informative, and verbose, a useful trait that helps to build users’ trust and engagement but these models also generate hallucinate information that can mislead users (Dziri et al., 2022b; Chiesurin et al., 2023). Moreover, many QA datasets have short reference answers that render traditional evaluation metrics like exact match (EM) and F1 word overlap unreliable when evaluating these verbose answers (Kamalloo et al., 2023).

Consider, for instance, the scenario in Figure 1, where the user question is “Where are One Direction from?”. A comparison between the reference response “London, England” and the first part of the model’s response “One Direction are from London, England” yields an EM score of 0 and F1 score of only 0.5, despite both answers being effectively equivalent (The entire response gets 0.36 F1 score). Moreover, the second part of the response asserts that One Direction is from Mullingar, Ireland, a fact which despite being correct, is not entailed by the provided knowledge. As EM and F1 only compare against reference answers, they are unsuitable to estimate the alignment of the model’s response with the provided knowledge.

In this work, we advocate that the performance of instruction-following models for retrieval-augmented QA should be evaluated along two dimensions — 1) *correctness w.r.t information need*, which measures a model’s efficacy in satisfying a user’s information needs, and 2) *faithfulness w.r.t provided knowledge*, which measures a model’s capability to ground responses in provided knowledge. A model demonstrating robust performance across both these dimensions can potentially be considered useful and safe for the user in information-seeking scenarios.

Along these dimensions, we evaluate several recent instruction-following models such as Llama-2 (Touvron et al., 2023b), GPT-3.5 (sibling model of Ouyang et al. 2022a), Flan-T5 (Chung et al., 2022), and Alpaca (Taori et al., 2023) on three popular QA datasets that correspond to three diverse QA tasks — Natural Questions (NQ; Kwiatkowski et al. 2019) for open-domain QA, HotpotQA (Yang et al., 2018) for multi-hop QA, and TopiOCQA (Adlakha et al., 2022) for conversational QA. We conduct a human

analysis of 900 model responses and correlate them with several automatic metrics for correctness and faithfulness.

Our findings suggest that, for correctness, *recall* – the proportion of tokens in the reference answer also present in the model response – exhibits the highest correlation than lexical overlap metrics like EM or F1. For faithfulness, *K-Precision* – the proportion of model response tokens that appear in the knowledge snippet – correlates better with human judgments than any other token-overlap metric. Among model-based metrics, i.e., using a model to determine the correctness/faithfulness of an answer w.r.t. reference answer/knowledge, GPT-4 correlates the most but it is expensive and prone to systematic biases (Wang et al., 2023). However, we find that lexical overlap metrics are close to model-based metrics, allowing us to evaluate several instruction-following models at a large-scale.

A faithful model should not only answer a question when relevant knowledge is provided, but it should also abstain from answering when irrelevant knowledge is provided. Hence, we also measure the model’s ability to abstain from answering as an evaluation for faithfulness.

To summarize, our contributions are as follows:

- We evaluate the performance of four instruction-following models – Llama-2, GPT-3.5, Flan-T5, and Alpaca – in retrieval-augmented settings across three diverse QA tasks. We collect human annotations for both correctness and faithfulness.
- We analyze several metrics in relation to human judgments, finding that GPT-4-based evaluation as the most correlated for both correctness and faithfulness. Additionally, we analyze failures of traditional QA metrics and highlight that models are unfairly penalized for verbosity.
- We propose simple token-overlap based metrics for both correctness and faithfulness – *recall* for correctness and *K-Precision* for faithfulness – and demonstrate their strong correlation with human judgments.
- Our results indicate that instruction-following models can surpass the performance of fine-tuned models in terms of correctness. However, these models struggle to be faithful to provided knowledge, often demonstrating a tradeoff between the ability to remain faithful to relevant and irrelevant knowledge.

2 Related Work

Instruction-Following Models Fine-tuning pre-trained models on a collection of NLP tasks formatted as natural language instructions result in instruction-following models. These models can generalize to new unseen tasks based solely on instruction and optionally a few demonstrations, often outperforming LLMs in zero-shot and few-shot settings while being only a fraction of their size (Mishra et al., 2022). Depending on the nature of the datasets used for training, these models can be broadly classified into three categories.

The majority of instruction-following models in the research community are trained on publicly available NLP datasets verbalized by human annotators (Wei et al., 2022; Mishra et al., 2022; Wang et al., 2022b; Chung et al., 2022; Iyer et al., 2023). The number of tasks ranges from a few tens (e.g. 62 in Wei et al. 2022) to several hundred (e.g. 1800+ in Iyer et al. 2023).

Ouyang et al. (2022a) conjecture that public NLP datasets are limited in scope and lack sufficient diversity in user inputs. To address this, they train *InstructGPT* on a mix of human-written prompts submitted to the OpenAI API and prompts created by expert labelers. The model is further fine-tuned with human feedback to align it more closely with human preferences (RLHF; Christiano et al. 2017). Llama-2 (Touvron et al., 2023b) is another recent model in this category, trained on a mix of public NLP datasets and high-quality expert annotations of dialogue-style instructions, followed by RLHF.

Finally, *self-instruct* (Wang et al., 2022a) is an alternative paradigm to reduce reliance on human-generated task instructions. Starting from a small manually-annotated task pool, an LLM is prompted to generate instructions and demonstrations of new tasks. The resultant synthetic dataset is used to train a language model to follow instructions (Taori et al., 2023; Peng et al., 2023).

Datasets for instruction-tuning often contain several QA tasks. However, these tasks are either reading comprehension (i.e. answering a question about a provided passage) or closed-book QA (i.e., without using a large information source). In this work, we explore a more practical setting, where an instruction-following model is paired with a retriever, a paradigm known as retrieval-augmented generation (RAG; Lewis et al. 2020).

Retrieval-Augmented Generation RAG entails using a *retriever* to select relevant passages from an information source, which are subsequently passed to a *generator* to produce a response. This two-step retrieve-generate process has been shown to reduce hallucinations (Shuster et al., 2021), while lending interpretability and configurability to the model (Lewis et al., 2020).

RAG is a dominant paradigm for several information-seeking QA tasks such as open-domain QA (Chen et al. 2017; Lee et al. 2019; Sachan et al. 2021, *inter alia*), multi-hop QA (Asai et al. 2020; Qi et al. 2021; Izacard et al. 2022; *inter alia*), and conversational QA (Anantha et al. 2021; Adlakha et al. 2022; *inter alia*). Various works differ on how to train the generator to utilize information from the retrieved passages, for e.g. by extracting snippets (Chen et al., 2017; Clark and Gardner, 2018; Wang et al., 2019; Karpukhin et al., 2020) or by jointly attending encoded passages and previously generated tokens (Fusion-in-Decoder; Izacard and Grave 2021).

Recent works have also explored using off-the-shelf language models as generators in the RAG pipeline, alleviating the need to fine-tune or learn additional parameters. Lazaridou et al. (2022) demonstrated that few-shot prompting an LM conditioned on the web results outperforms a vanilla LM for several open-domain QA tasks. Shi et al. (2023) showcase that pairing LLMs like GPT-3 (Brown et al., 2020) with retrievers improves language modeling performance as well. Separate from these works, we evaluate retrieval-augmented instruction-following models based only on natural language instruction. In the absence of training instances or demonstrations, these models do not learn the distribution of reference answers of the target QA dataset, raising new challenges for evaluation.

Evaluation in QA Lexical matching between a set of reference answers and model response remains a dominant approach for evaluation across multiple NLP tasks. As QA tasks generally consist of short reference answers, previous works have primarily relied on Exact Match (EM) and F1 to evaluate and benchmark models (Rajpurkar et al., 2016; Reddy et al., 2019). For tasks that require generating longer sequences, such as summarization and translation, subsequence-based lexical matching is generally employed (Papineni et al. 2002; Banerjee and Lavie 2005; Lin 2004, *inter alia*).

A major shortcoming of lexical matching is that it depends on a set of reference answers which may be incomplete. To overcome this limitation, subsequent model-based metrics compute the semantic similarity between the reference answer and the model response using contextualized embeddings (Zhang et al., 2020) or train a specialized classifier (Bulian et al., 2022) to predict equivalence. More recently, several works resort to prompting LLMs like GPT-4 (OpenAI, 2023) to act as evaluators (Chiang et al., 2023; Peng et al., 2023; Chiang and Lee, 2023; Kamaloo et al., 2023; Liu et al., 2023c). In this work, we explore evaluating both correctness and faithfulness using GPT-4.

Concurrent to our work, Kamaloo et al. (2023) evaluate the correctness of InstructGPT in zero-shot and few-shot settings along with several fine-tuned models for open-domain QA. They highlight the shortcomings of traditional QA metrics and propose BEM (Bulian et al., 2022) and LLM-based evaluation as viable alternatives. However, they do not consider InstructGPT in *retrieval-augmented* settings. In contrast to their work, we investigate both correctness and faithfulness of multiple instruction-following models across three diverse QA tasks and propose simple token-overlap based metrics that correlate highly with human judgments.

Faithfulness and Groundedness Conversational models have been shown to produce factually incorrect or unsupported statements (Rashkin et al., 2021b; Dziri et al., 2022b), known as *hallucinations*. To alleviate those issues, various works attempt to reduce hallucinations via methods such as iterative refinement (Dziri et al., 2021), linguistic calibration (Mielke et al., 2022; Lin et al., 2022), or by editing instances of hallucinations (Dziri et al., 2022a), thus improving *faithfulness* of these models. Several metrics have also been developed to measure faithfulness. Honovich et al. (2021) proposed Q^2 , an automatic faithfulness evaluation metric that checks for factual consistency based on automatic question generation and question answering. FaithCritic (Dziri et al., 2022a) is another model-based metric that predicts the degree of hallucination in a model’s response.

For information-seeking, previous works have considered *groundedness* — the extent to which the generator relies on retrieved passages (Paranjape et al., 2022), quantified using Knowledge-F1 (K-F1; Shuster et al. 2021). In this work, we consider

a model response to be faithful, if it is grounded in the passage *relevant* to the user’s information need. Concurrent to our work, Chiesurin et al. (2023) investigated hallucination of retrieval-augmented GPT-3 in for conversational QA (Adlakha et al., 2022) task. They found that GPT-3 is likely to produce responses that appear trustworthy but are unfaithful.

3 Experimental Setup

3.1 Tasks

We evaluate our approach on the validation splits of three information-seeking QA tasks. The total number of questions and passages for each dataset are provided in Table 1. We describe the datasets used for each task below.

Open-domain QA Natural Questions (NQ; Kwiatkowski et al. 2019) includes questions sourced from Google queries, with reference answers written by human annotators. We use the open version of NQ (Lee et al., 2019) that consists of short answers based on 100-token passages from English Wikipedia (indexed in Dec. 2018).

Multi-hop QA We use HotpotQA (Yang et al., 2018) for this task, where each question requires reasoning across two Wikipedia passages. The passages are taken from the initial paragraphs from English Wikipedia articles (indexed in October 2017).

Conversational QA We use TopiOCQA (Adlakha et al., 2022) for this task, a dataset for open-domain information-seeking dialogue. At each turn of the conversation, an *agent* responds to a *user*’s questions based on knowledge from Wikipedia. Each turn has an associated 200-token gold passage from English Wikipedia (indexed in Oct. 2020).

3.2 Instruction-following Models

To evaluate retrieval-augmented instruction-following language models, we present the models with an instruction, followed by the retrieved passages and the query. The prompt template for open-domain QA and multi-hop QA tasks is given in Figure 2, whereas conversational QA differs slightly, replacing the question with conversation history (Figure 3). We consider four instruction-following models that primarily differ based on the type of training data used. We use the same generation parameters for all instruction-following models, described in Appendix A.1.

Dataset	# Questions	# Passages
Natural Questions	3,610	21,015,324
HotpotQA	7,405	5,233,329
TopiOCQA	2,514	25,700,593

Table 1: Statistics for datasets used in this work. We use the validation split from each dataset for our evaluation as the test sets are hidden.

```
Please answer the following question given
the following passages:
- title: {Passage title}
{Passage text}

- title: {Passage title}
{Passage text}
...
Question: {Question}
Answer:
```

Figure 2: The prompt template used for open-domain QA and multi-hop QA tasks.

```
Please answer the following question given
the following passages and the conversation
history:
- title: {Passage title}
{Passage text}

- title: {Passage title}
{Passage text}
...
User: {Question 1}
Agent: {Answer 1}
...
User: {Question k}
Agent:
```

Figure 3: Prompt template for conversational QA task.

Flan-T5 We use the 11B parameter version of T5 (Raffel et al., 2020), which has been trained by Chung et al. (2022) using the instruction fine-tuning methods proposed by Wei et al. (2022). Flan-T5 is trained on multiple publicly-available instruction-following datasets (Sanh et al., 2022; Wang et al., 2022b; Wei et al., 2022). Together, these datasets encompass more than 1800 tasks, of which over 200 are QA tasks. Out of the three datasets on which we evaluate, the training split of NQ and HotpotQA are included in Flan-T5’s training regime.

GPT-3.5 We use the *turbo* version of GPT-3.5¹ which is described² as a sibling to the InstructGPT model (Ouyang et al., 2022a). The model’s training

¹openai.com/blog/introducing-chatgpt-and-whisper-apis

²openai.com/blog/chatgpt

incorporates user data submitted to the OpenAI API as well as expert annotations, however, the exact distribution of training tasks and datasets is not publicly available.

Alpaca We use the 7B variant of Alpaca (Taori et al., 2023), a fine-tuned version of LLaMA (Touvron et al., 2023a) trained on demonstrations generated using GPT-3 (Brown et al., 2020). The demonstrations were collected using the *self-instruct* framework (Wang et al., 2022a).

Llama-2 We use the 7B chat version of Llama-2 (Touvron et al., 2023b). The model is initially bootstrapped on similar instruction-following dataset as Flan-T5, followed by fine-tuning for dialogue-style instructions.

Fine-tuned Generators To compare against instruction-following models, we select FiD (Izacard and Grave, 2021) as our fine-tuned baseline for all three tasks. This encoder-decoder model separately encodes each retrieved passage with the query, resulting in a set of vectors. The decoder then autoregressively generates the answer by attending to the input passages and the previously generated tokens. For NQ and TopiOCQA, we use the publicly available FiD checkpoints, while for HotpotQA, we fine-tune our own variant using the default hyperparameters.

4 Correctness w.r.t Information Need

In this section, we investigate if retrieval-augmented instruction-following models can produce responses that satisfy user information needs. We first describe our experimental setup by providing details of the retriever used in each task (§4.1) and the metrics used for evaluating model responses (§4.2). Next, we describe our human evaluation setup and present the results from our analysis (§4.3). Finally, equipped with a better understanding of evaluation metrics, we conduct large-scale evaluation of instruction-following models and present the results (§4.4).

4.1 Retrieval

For each task, we use a task-specific variant of DPR (Dense Passage Retrieval; Karpukhin et al. 2020) as the retriever. The general architecture of DPR consists of a question and a passage encoder. The dot product between the dense vector representations of the passage and the query is used as a ranking function.

For NQ, we adopt a pre-trained checkpoint from Karpukhin et al. (2020). This checkpoint was trained on four QA datasets — NQ, TriviaQA (Joshi et al., 2017), WebQuestions (Berant et al., 2013), and CuratedTREC (Baudis and Sedivý, 2015). For HotpotQA, we utilize a multi-hop variant of DPR proposed by Xiong et al. (2021). This version retrieves reasoning chains iteratively, selecting subsequent passages based on the query and previously retrieved passages. For TopiOCQA, we utilize the checkpoint provided by Adlakha et al. (2022). This variant of DPR is uniquely suited for conversational QA tasks as it encodes the conversation history in the question encoder.

In all of the tasks, the retriever selects passages from the associated Wikipedia dump, as detailed in Section 3.1. The number of retrieved passages provided to instruction-following models and fine-tuned models for each task are provided in Appendix A.2.

4.2 Evaluation Metrics

Evaluation in QA usually involves comparing model responses to human-annotated gold answers. The metrics used for this comparison can be divided into two categories:

Lexical Match These metrics score a model response based on its token overlap with the gold standard answer. While some metrics perform bag-of-words matching (e.g., Exact Match (EM), F1), others consider the order of the tokens by n -gram matching such as METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004).

In this work, we also consider *Recall* — the proportion of tokens in the reference answer that are present in the model response. Recall does not penalize verbose model response, as long as the response contains the reference answer tokens. Recent works that have evaluated the verbose responses generated by instruction-following models (Liu et al., 2023a; Mallen et al., 2022) have used a similar metric *accuracy*, whereby a model’s response is considered correct if any reference answer appears as a substring within the model’s response. This is a stricter version of recall that cannot handle small variations between reference answer and model response, such as if the reference answer is *John Kennedy* and the model response is *John F Kennedy*. To avoid any confusion, we refer to this metric as *Recall (S)*, indicating it as a stricter version of token-level recall.

Semantic Similarity Unlike the previous class of metrics that face *strictness* issues (Kamalloo et al., 2023), *semantic similarity-based* metrics typically leverage a trained model to predict if the model response is semantically equivalent to the gold answer. BERTScore (Zhang et al., 2020), which we refer to as *BertS*, is a commonly used metric for text generation that computes precision, recall, and F1 based on token similarity between model response and reference gold answer using contextual BERT embeddings. Furthermore, *BEM* (BERT matching, Bulian et al. 2022) employs a trained BERT model to evaluate question-answering models by predicting the semantic equivalence based on the question, reference gold answer, and model response. We extend BEM to conversational QA task by providing the question from the last turn of the conversation as input. Moreover, we also consider an evaluation metric based on prompting LLMs (referred to as *GPT3.5-Eval* and *GPT4-Eval*) to act as evaluation agents. In principle, the setup is similar to the one proposed by Kamalloo et al. (2023), however, with a different prompt, as described in Appendix B (Figure 7). Specifically, we prompt these models to act as evaluators by providing a natural language instruction along the question (or conversation history), reference gold answer, and model response.

4.3 Human Evaluation

We conduct a human evaluation on a subset of responses generated by three instruction-following models – GPT-3.5, Flan-T5, and Alpaca – to establish a basis for comparing evaluation metrics. Specifically, we focus on cases where retrieved passages provided to the model include the gold passage. Therefore, any inaccuracies in the response can be attributed to the model’s failures, rather than inaccurate retrieval. For every task, we collect annotations for 100 samples.

In our evaluation setup, the annotator is presented with the question or conversation history, the reference answer, and the anonymized model response. The annotator’s task is to assess if the model response is *correct*, i.e. it satisfies the information need underlying the question. For each of the 100 samples, we collect annotations for three instruction-following models, resulting in 900 labeling tasks. Each task is completed by two different annotators (authors of the paper). The inter-annotator agreement achieved was 92.7%. In instances where the annotators disagreed, a third an-



Figure 4: Failure cases of F1 metric. *More Elaborate Answers* is the most common failure sub-category, followed by *Open-ended Questions*.

notation is collected and a majority vote is taken.

The results of this human evaluation are presented in Table 8 (Appendix D), along with scores of automated metrics on this subset. Traditional QA evaluation metrics like EM and F1 tend to score model responses much lower than human assessments, highlighting the well-known problem of strictness in lexical matching (Min et al., 2021; Kamaloo et al., 2023).

Qualitative Analysis of Failure Cases For a more granular understanding of the shortcomings of traditional QA metrics, we analyze the models’ responses that have less than or equal to 0.3 F1 score, but were deemed correct according to the human evaluations. This resulted in 296 samples out of 900. Our classification of errors is adapted from Kamaloo et al. (2023) (which itself was based on Min et al. 2021), modified to focus on instruction-following models. Specifically, we exclude some error classes relevant to fine-tuned models and include some classes for instruction-following models. The resultant categories are:

- **Semantic Equivalence:** Here, the model response is semantically similar to the reference answer. Sub-categories include **Multinomial entities**, e.g., *John Kennedy* and *John F Kennedy*, **Synonymous Answers**, e.g., *from India* and *Indian nationality*, and **More Elaborate Answers**, e.g., *yes* and *yes, he is member of the band*.

- **Symbolic Equivalence:** This primarily refers to different possible representations of numeric quantities, e.g. *four seasons* and *4 seasons*, or *3000 BC* and *Early Dynastic Period*.

- **Intrinsic Ambiguity in Questions:** This refers to queries with multiple valid interpretations, leading to a range of correct answers, e.g. *Who is command sergeant major of the army?* could be seeking the person’s identity or a description of the position itself. This category also includes cases where the correct answer is dependent on the specific point in time being referenced, e.g. *Who won NFL football coach of the year?*.

- **Granularity Discrepancies:** The level of specificity in the model’s response may not align with that in the reference answer. This discrepancy in granularity can be **Temporal**, e.g., *August 25, 1939* and *1939*, or **Spatial**, e.g., for the question *Where’s the tv show The Crossing filmed?*, *Vancouver* and *British Columbia, Canada* are both correct answers.

- **Incomplete Reference Answers:** These cases occur when the reference answers, despite their number, fail to cover the entire spectrum of correct responses. We break this category into two types — **List of named entities** which includes questions like the cast of a movie or members of the band, and **Open-ended questions** which included questions that can be answered in multiple different ways, all of which are not captured by reference answers., e.g., *What was the Watergate scandal?*.

- **Enumeration of Reference Answers:** This error happens especially in NQ samples, where the question asks for a list of entities (e.g., all states of a country), but each reference answer includes only one entity (e.g., a single state). The instruction-following models often generate all the entities in their response, which results in low overlap with each individual reference answer.

- **Satisfactory Subset Responses:** This category pertains to instances where the model’s response, although containing less information than the reference answer, still provides an acceptable response to the user’s query. For instance, for the question *"name some of her*

Error Type: More Elaborate Answers
Dataset: HotpotQA

Question: At what theater is the composer and lyricist for the musical Big Fish a residential artist?
Ref. Answer: Ars Nova Theater
Response (Alpaca): The composer and lyricist for the musical Big Fish, Andrew Lipa, is a residential artist at the [Ars Nova Theater](#) in New York City.

Scores:
F1: 26.0, Recall: 100.0, METEOR: 54.5, Rouge-L: 22.2, BertS (F1): 85.4, BEM: 98.9, GPT4-Eval: 100.0

Error Type: Open-ended Questions
Dataset: TopiOCQA

Question: what was the watergate scandal?
Ref. Answer: It was an array of clandestine and often illegal activities undertaken by members of the Nixon administration.
Response (Flan-T5): The Watergate scandal was a political scandal in the United States involving [the administration of U.S. President Richard Nixon](#) from 1972 to 1974 that led to Nixon’s resignation.

Scores:
F1: 21.8, Recall: 26.7, METEOR: 15.6, Rouge-L: 15.4, BertS (F1): 86.1, BEM: 73.9, GPT4-Eval: 100.0

Error Type: Enumeration of reference answers
Dataset: NQ

Question: all of the states in the northeast region
Ref. Answers: 1. New Hampshire, 2. Maine, 3. Rhode Island, 4. Pennsylvania, 5. Vermont, 6. New York, 7. Connecticut, 8. New Jersey, 9. Massachusetts
Response (GPT-3.5): The states in the northeast region include [Maine](#), [New York](#), [New Jersey](#), [Vermont](#), [Massachusetts](#), [Rhode Island](#), [Connecticut](#), [New Hampshire](#), and [Pennsylvania](#).

Scores:
F1: 20.0, Recall: 100.0, METEOR: 39.0, Rouge-L: 17.4, BertS (F1): 82.7, BEM: 98.9, GPT4-Eval: 100.0

Figure 5: Qualitative examples cases where F1 fails, along with scores from other evaluation metrics.

songs”, the reference answer might list 5-6 song names, while the model response includes only 1-2. This situation is predominantly observed in the TopiOCQA dataset.

Figure 4 displays the distribution of error cases based on our classification. A significant portion of the errors (55.63%) fall under the *More Elaborate Answers* category. This suggests that traditional QA metrics often penalize models unjustly due to the verbose nature of their responses. The next most common sub-category, *Open-ended Questions* (13.99%), suggests that models are occasionally penalized for providing correct answers that were not included in the reference responses. The percentage share and exact count of all categories

Metric	Spearman ρ	Kendall τ
EM	26.466	26.168
F1	49.323	41.096
Precision	44.846	37.772
Recall	62.72	57.142
Recall (S)	54.072	53.464
METEOR	51.57	41.886
Rouge-L	46.993	39.001
BertS (F1)	36.862	29.691
BertS (Precision)	24.379	19.519
BertS (Recall)	42.886	34.58
BEM	53.649	43.727
GPT3.5-Eval	63.514	62.801
GPT4-Eval	70.152	69.363

Table 2: Correlation of several lexical matching and semantic similarity evaluation metrics with human judgments for correctness w.r.t information need. GPT4-Eval achieves the highest correlation overall. Recall is the highest correlated among all lexical overlap metrics.

are reported in Table 7 (Appendix C).

In Figure 5, we provide qualitative examples of common failure modes, along with their associated evaluation metrics scores. Recall appears to be an effective fix for sub-categories such as *More Elaborate Answers* and *Enumeration of Reference Answers*. However, both lexical match based and semantic similarity based metrics struggle with *Open-ended Questions*. Although GPT4-Eval appears to be relatively robust based on examples in Figure 5, this metric has some failures, with most common failure sub-category being *Open-ended Questions*. The complete distribution of failure cases according to sub-categories is reported in Figure 10, along with qualitative examples in Figure 11 (Appendix C).

Overall, the results of our human evaluation and analysis indicate that traditional metrics such as EM and F1, typically used in the literature for fine-tuned QA models, are not well-aligned with the verbose nature of instruction-following models. To determine more suitable metrics for these models, we analyze the correlation of each metric with human assessments.

Correlation Between Automatic Metrics and Human Judgement Table 2 presents the correlation between different metrics with human judgments. Apart from metrics detailed in Section 4.2, we include token-level precision, as well as precision and recall as computed using BERTScore. We report Spearman’s ρ and Kendall’s τ correlation.

Notably, GPT4-eval has the highest agreement with human judgments, with 70.15 Spearman cor-

Dataset	Model	EM	F1	Recall	METEOR	Rouge-L	BertS (F1)	BEM
NQ	FiD	46.57	53.93	54.45	42.94	54.33	92.57	58.81
	GPT-3.5	1.27	15.12	58.56	25.68	14.57	83.08	69.45
	Flan-T5	<u>41.16</u>	<u>50.62</u>	54.03	<u>40.80</u>	<u>51.02</u>	<u>91.77</u>	58.74
	Alpaca	8.78	20.3	46.23	23.17	20.61	84.67	55.97
	Llama-2	0.61	11.85	52.37	21.16	11.38	82.58	<u>62.30</u>
HotpotQA	FiD	<u>48.43</u>	60.16	60.55	46.03	60.18	93.02	67.94
	GPT-3.5	5.63	22.16	66.77	31.56	21.67	84.16	78.16
	Flan-T5	58.12	71.14	71.28	53.44	71.16	94.37	76.19
	Alpaca	16.25	33.54	56.76	33.23	33.5	86.88	67.74
	Llama-2	1.39	15.91	<u>67.55</u>	27.48	15.23	83.08	<u>78.05</u>
TopiOCQA	FiD	36.48	58.52	61.64	52.46	58.26	92.37	<u>66.55</u>
	GPT-3.5	2.63	36.07	66.72	47.35	33.81	88.14	69.34
	Flan-T5	<u>18.34</u>	<u>43.17</u>	52.54	42.42	<u>42.88</u>	<u>89.42</u>	56.57
	Alpaca	5.85	28.61	41.3	31.08	27.75	87.07	46.41
	Llama-2	0.32	25.16	55.3	35.16	23.42	86.06	56.33

Table 3: Performance of retrieval-augmented instruction-following models on three diverse information-seeking QA tasks. Among the metrics reported, *Recall* is most correlated with human judgements. Based on recall, instruction-following models outperform fine-tuned FiD on all three tasks.

relation and 69.36 Kendall correlation, closely followed by GPT3.5-Eval. We speculate that the language comprehension capabilities and inherent world knowledge embedded in LLMs like GPT-3.5 and GPT-4 help them overcome many of the challenges associated with evaluating responses of instruction-following models that we identified in our human evaluation study.

After GPT4-eval and GPT3.5-Eval, Recall achieves the highest correlation with human judgement. This simple token-overlap metric correlates better than other lexical matching-based metrics or more complex semantic similarity metrics like BERTScore and BEM, likely because it does not penalize verbosity in model responses.

Surprisingly, BERTScore fares worse than token-overlap F1, even when only considering the recall component of the metric. We hypothesize that the underlying issue is the poor quality of BERT token embeddings in short strings (Bommasani et al., 2020), a common characteristic of reference answers in QA datasets. For example, for the reference answer *yes, that is correct*, the model response *yes* receives the BERTScore of 0.806 and *no* receives a slightly higher score of 0.815. Although BEM performs better than F1, it still falls short of token-overlap recall. Given that BEM’s training data includes model responses of QA systems trained on SQuAD (Rajpurkar et al., 2016), it probably doesn’t generalize well to more verbose responses of instruction-following models.

Although LLM-based evaluation, such as GPT4-

eval and GPT3.5-eval, have the highest correlation with human judgements on the selected subset of responses, they also have certain limitations. Accessing these proprietary models incurs substantial API costs, which renders them impractical for automatic evaluation on large-scale datasets. Moreover, the reliability of LLMs as evaluators is still unclear, as recent studies have shown that they may exhibit systematic bias (Wang et al., 2023) and can be sensitive to input instructions (Bowman, 2023). Secondly, it is currently unclear how reliable LLMs are as evaluators, with some recent works demonstrating that they exhibit systematic bias (Wang et al., 2023) and are sensitive to input instructions (Bowman, 2023). Given these considerations, we rely on Recall to compare model performance.

4.4 Automatic Correctness Evaluation

The performance of both instruction-following and fine-tuned models in a retrieval-augmented generation setup across multiple datasets is reported in Table 3 using several lexical matching and semantic similarity metrics. Unsurprisingly, traditional QA metrics like EM and F1 assign much lower scores to instruction-following models, compared to fine-tuned FiD. The only exception is Flan-T5, that outperforms FiD with a 17.72% gap. However, it should be noted that Flan-T5 is trained on a wide range of QA tasks, including NQ and HotpotQA (Section 3.2).

Based on our finding in Section 4.3, we consider Recall to get true estimate of model perfor-

mance. Using recall, the performance gap between instruction-following and fine-tuned models narrows significantly, with some instruction-following models even outperforming FiD. Notably, GPT-3.5 outperforms the fine-tuned FiD across all three QA task – 7.55% gap in NQ, 10.27% in HotpotQA, and 8.24% in TopiOCQA. These results suggest that in retrieval-augmented settings, instruction-following models are equally, or even more capable than fine-tuned generators in generating correct responses w.r.t user information needs.

5 Faithfulness w.r.t Provided Knowledge

As previously noted, instruction-following models often produce verbose responses. Consequently, responses from these models often contain supplementary information which can be hallucinated (Rashkin et al., 2021a; Dziri et al., 2022b; Chiesurin et al., 2023). In this section, we conduct an analysis of the faithfulness of instruction-following models w.r.t knowledge provided as part of the input. We posit that an optimal generator’s response should rely *solely* on the knowledge relevant to the user information need. Based on this hypothesis, we split our analysis into two parts – 1) faithfulness w.r.t relevant knowledge, where we prompt the instruction-following model with the user question paired with the corresponding gold passage and evaluate the groundedness of the response in the provided knowledge, and 2) faithfulness w.r.t irrelevant knowledge, where we provide a related but irrelevant passage and measure how often the model refuses to answer.

In this section, we first describe the automatic faithfulness metrics (§5.1). Next, similar to correctness, we conduct a human evaluation and compute correlations for all metrics, followed by large-scale evaluation of faithfulness w.r.t relevant knowledge (§5.2). Finally, we analyze the capabilities of models to refrain from answering in the presence of irrelevant knowledge (§5.3).

5.1 Faithfulness Metrics

Here we describe the metrics that we use for automatic evaluation in Section 5.2. Given the user question or the conversation history (denoted by \mathcal{H}), the gold passage \mathcal{K} , and the model response u , the goal is to check if u is grounded in \mathcal{K} . We consider both faithfulness and groundedness metrics in the literature for this task.

K-F1 Knowledge-F1 (denoted K-F1) is a lexical overlap metric that checks for F1 overlap between the tokens of u and \mathcal{K} . Although it has been widely used for knowledge-grounded dialogue (Shuster et al., 2021; Dziri et al., 2022a), we argue it is unsuitable for assessing groundedness in information-seeking tasks. In information-seeking, model responses tend to be shorter than the knowledge snippet. Hence, even if the model selects precise information from the knowledge, it is penalized for not utilizing the entire knowledge snippet by K-F1.

K-Precision To counter the shortcomings of K-F1, we propose K-Precision – the proportion of tokens in the model response u that are present in \mathcal{K} . The intuition behind this is that in information-seeking, grounding u in \mathcal{K} is inherently an asymmetric task, i.e., u can be a subset of \mathcal{K} but \mathcal{K} cannot be a subset of u .

K-BertS Following Shuster et al. (2021) and Dziri et al. (2022a), we use of BERTScore to measure semantic similarity between \mathcal{K} and u based on contextual BERT token embeddings. We refer to this as **K-BertS** to differentiate it from BertS (Section 4).

FaithCritic We use the hallucination critic model by Dziri et al. (2023) to evaluate whether a response entails a given passage.³ It outputs a score between 0 and 1 indicating how likely a given response is hallucinated. Here, lower scores are indicative of lesser hallucination within a model’s responses, hence, more groundedness.

Q² Q^2 (Honovich et al., 2021) is an evaluation metric used to quantify factual consistency between responses and provided passages using automatic question generation, question answering, and natural language inference (NLI) models.

LLMCritic Similar to correctness, we investigate prompting LLMs to act as evaluator for groundedness. More specifically, we prompt GPT-3.5 and GPT-4 to annotate whether a given response uses *only* the knowledge present in the provided passage. The actual prompt is provided in Appendix B (Figure 8).

5.2 Faithfulness w.r.t Relevant Knowledge

In this section, we investigate the faithfulness of model responses when they are provided a passage relevant to the user query. We first conduct human

³RoBERTa-Large checkpoint: huggingface.co/McGill-NLP/roberta-large-faithcritic

Metric	Spearman	Kendall
K-F1	-2.67	-2.074
K-Precision	46.482	41.536
K-Recall	-4.258	-3.388
K-BertS (F1)	3.583	3.009
K-BertS (Precision)	19.721	16.07
K-BertS (Recall)	-10.3	-8.22
FaithCritic	11.741	9.528
Q^2 (F1)	27.883	23.932
Q^2 (NLI)	27.524	24.228
LLMCritic (GPT-3.5)	27.189	26.789
LLMCritic (GPT-4)	50.485	49.742

Table 4: Correlation of evaluation metrics of faithfulness with human judgments. LLMCritic (GPT-4) is most correlated with human judgements. K-Precision is a close second.

evaluation on a subset of samples, and use it to compare several evaluation metrics. Finally, we present the results of large-scale automatic evaluation of instruction-following models.

We conduct experiments on all three information-seeking tasks. For HotpotQA and TopiOCQA, the gold passage(s) for each query is provided as part of the dataset. For NQ, we follow Karpukhin et al. (2020) and provide each question and reference answer as a query to BM25 and take the first ranked passage as the gold passage. For all instruction-following models, we use the prompt provided in Section 3.

Human Evaluation For each example, we provide annotators with a question (or the conversation history), response, and retrieved passages and task them with determining whether the response is grounded in the provided passages. We allow annotators to provide two labels – 1) to determine if the provided passage is actually a relevant passage to the user’s query, and 2) to determine if the model response is “completely,” “partially,” or “not” found in the presented passages. The model response is given a score of 1.0 if the label is “completely,” 0.5 for “partially” and 0 for “not.” We collect two annotations for each example and resolve all conflicting annotations by collecting a third annotation and taking the majority vote.

We randomly sample 50 examples from Natural Questions, HotpotQA, and TopiOCQA for evaluation. We first filter out annotations for which the passage is not relevant to the query. This resulted in 39 samples for NQ, 47 for HotpotQA, and 49 for TopiOCQA. The high number of non-relevant for NQ is probably due to heuristic matching of gold

Dataset: HotpotQA

Question: Which 2008 American documentary film, Pond Hockey or I.O.U.S.A., delves into fiscal issues?

Knowledge #1: Pond Hockey (film) - [...] The film is an examination of the changing culture of pond hockey.

Knowledge #2: I.O.U.S.A. - I.O.U.S.A. is a 2008 American documentary film directed by Patrick Creadon. The film focuses on the shape and impact of the United States national debt [...] and was known as the "Fiscal Wake-Up Tour."

Ref. Answer: I.O.U.S.A.

Response (Alpaca): Pond Hockey delves into fiscal issues. I.O.U.S.A. focuses on the shape and impact of the United States national debt.

Scores: K-F1: 15.91, K-Precision: 82.35, K-BertS (F1): 84.64, Q^2 (F1): 20.0, Q^2 (NLI): 20.0, FaithCritic: 0.53, LLMCritic(GPT-3.5): 0.0, LLMCritic(GPT-3.5): 0.0

Dataset: NQ

Question: When did they replace lead with graphite in pencils?

Knowledge: Pencil - [...] many people have the misconception that the graphite in the pencil is lead, [...] even though it never contained the element lead. [...]

Ref. Answer: never contained the element lead

Response (Flan-T5): 1835

Scores: K-F1: 0.0, K-Precision: 0.0, K-BertS (F1): 76.22, Q^2 (F1): 0.0, Q^2 (NLI): 0.0, FaithCritic: 99.96, LLMCritic(GPT-3.5): 100.0, LLMCritic(GPT-3.5): 0.0

Dataset: TopiOCQA

Question: What are dragonfly flight speeds?

Knowledge: Dragonfly (Flight speed) - [...] a maximum speed of 36–54 km/h (22–34 mph) with average cruising speed of about 16 km/h (9.9 mph). Dragonflies can travel at 100 body-lengths per second in forward flight, and three lengths per second backwards

Ref. Answer: It can travel at 100 body-lengths per second in forward flight, and three lengths per second backwards.

Response (GPT-3.5): Dragonflies like the hawks have a maximum speed of around 30 miles per hour with an average cruising speed of about 10-15 miles per hour. They can also travel at 100 body-lengths per second in forward flight

Scores: K-F1: 54.54, K-Precision: 72.97, K-BertS (F1): 89.48, Q^2 (F1): 40.0, Q^2 (NLI): 40.0, FaithCritic: 99.55, LLMCritic(GPT-3.5): 100.0, LLMCritic(GPT-3.5): 100.0

Figure 6: Examples of non-faithful responses alongside relevant metric scores. Text in purple indicates hallucination, while teal responses are grounded to teal provided knowledge.

passage to the question. We consider three models – GPT-3.5, Flan-T5, and Alpaca, resulting in 405 samples. We compute scores from all evaluation metrics on this subset, including LLMCritic (for both GPT-3.5 and GPT-4). These are presented in Table 9 (Appendix D).

In Table 4, we present correlations between different automatic groundedness metrics and human evaluation. We find that LLMCritic based on GPT-4 correlates the most with human evaluation. K-

Dataset	Model	K-F1 \uparrow	K-Precision \uparrow	K-BertS (F1) \uparrow	Q^2 (F1) \uparrow	Q^2 (NLI) \uparrow	FaithCritic \downarrow
NQ	GPT-3.5	19.66	65.78	85.34	38.17	43.07	19.37
	Flan-T5	5.84	94.04	80.9	36.54	38.27	82.42
	Alpaca	13.29	70.44	83.40	30.18	33.46	69.92
	Llama-2	20.42	70.9	84.94	–	–	32.37
HotpotQA	GPT-3.5	16.61	81.19	84.18	49.32	56.07	38.95
	Flan-T5	3.26	92.12	78.57	36.03	37.97	64.31
	Alpaca	9.55	87.03	82.68	43.51	49.05	50.32
	Llama-2	17.7	76.9	83.65	–	–	38.53
TopiOCQA	GPT-3.5	26.82	71.96	87.01	54.74	60.44	30.71
	Flan-T5	23.74	86.37	86.42	61.30	64.75	44.89
	Alpaca	19.26	66.91	84.96	40.21	44.83	58.28
	Llama-2	24.75	64.64	86.19	45.00	50.72	42.55

Table 5: Results for faithfulness w.r.t relevant knowledge. We report both token-based and model-based metrics. For all metrics except FaithCritic, higher scores indicate greater response groundedness.

Precision, the token-overlap based metrics that is invariant to the length of the knowledge snippet in a close second, better than other model-based faithfulness metrics like K-BertS, FaithCritic, and Q^2 . This indicates that models trained to detect hallucinations in knowledge-grounded dialogues do not generalize well to information-seeking QA tasks. We present some examples of model hallucinations in Figure 6, along with associated scores of evaluation metrics.

Automatic Evaluation In Table 5, we present the results for faithfulness w.r.t relevant knowledge on NQ, HotpotQA, and TopiOCQA. Traditional faithfulness metrics such as K-F1, K-BertS, and FaithCritic, rank either Llama-2 or GPT-3.5 as the most faithful model for all the three tasks.

On the other hand, K-Precision, the metric most correlated with human judgments, denotes a completely different trend. GPT-3.5 is the *least* faithful for NQ, while Llama-2 is least faithful for HotpotQA and TopiOCQA. K-Precision ranks Flan-T5 as the most faithful instruction-following model for all three tasks. We hypothesize that K-F1 faces a similar issue as F1 in correctness evaluation – there is a length mismatch between the model response and the provided knowledge snippet. Our preliminary examination of model responses reveals that Flan-T5 responses are generally short, which is probably why K-F1 assigns it a low score.

These findings further highlight that verbose responses from instruction-following models are often not grounded in provided passages. For example, in Figure 6, GPT-3.5 hallucinates by outputting numbers that are completely different from what was provided, whereas Alpaca fails to reason properly based on provided passages.

5.3 Faithfulness w.r.t Irrelevant Knowledge

In the retrieval-augmented setting, an ideal model should comprehend passage contents and avoid answering if the passage lacks relevant information. To test this, we provide the models with an irrelevant passage by selecting the 1001 ranked passage from the list of retrieved passages.

Prompt Setup Our preliminary experiments demonstrated that without an explicit instruction, Flan-T5 and Alpaca did not refrain from answering at all. Hence, we modified the prompt to make this behavior more explicit and instructed the model to output *I don’t know* if the passage is deemed irrelevant, as demonstrated in Figure 9 (Appendix B). We report the proportion of model responses that contain *I don’t know* and other observed synonymous expressions.⁴ Note that for these experiments, we only investigate whether a model refused to answer. We do not verify the correctness of any generated responses. Moreover, to measure the impact of this new instruction, we also experiment with providing the gold passage and report the proportion of model responses that do *not* contain *I don’t know* and other synonymous expressions.

Results We present our results in Table 6. We find that when provided with an irrelevant passage, Llama-2 most often refuses to answer on open-domain and multi-hop QA datasets (more than 99% in NQ and HotpotQA). GPT-3.5 performs the best for TopiOCQA, refraining to answer on 88.15% turns. However, for both of these models, the inclination to not answer also extends to when the gold passage is actually present. In comparison, Flan-T5 is well balanced on datasets it was exposed to

⁴“UNANSWERABLE”, “..passages do not contain..”

Dataset	Model	Incorrect Psg. \uparrow	Gold Psg. \downarrow
NQ	GPT-3.5	98.5	48.01
	Flan-T5	91.99	24.76
	Alpaca	0.06	0.00
	Llama-2	99.34	75.84
HotpotQA	GPT-3.5	98.54	26.39
	Flan-T5	77.14	1.58
	Alpaca	0.09	0.11
	Llama-2	99.16	76.96
TopiOCQA	GPT-3.5	88.15	32.42
	Flan-T5	40.77	7.68
	Alpaca	1.27	0.80
	Llama-2	87.59	61.77

Table 6: Percentage of model responses that contain *I don’t know* and other synonymous expressions when provided with an incorrect passage (higher is better) or the gold passage (lower is better).

during training, however, it remains overconfident on TopiOCQA, which was not included in the training. Alpaca adheres the least to the instruction and answers even if the passage is not relevant to the information need of the user. [Appendix E](#) demonstrates some failure examples of these models in both scenarios. Further research is required to optimally design and prompt models to better identify when to answer and when not to answer.

6 Discussion and Limitations

Below, we highlight several key findings of this paper and discuss some of its limitations.

Which Evaluation Metrics are Best? Our analysis on correctness (§4) and faithfulness (§5) demonstrates that widely-used metrics are not suitable for evaluating the correctness (due to errors such as elaborate answers, open-ended questions, and list of named-entities) and faithfulness (due to partially grounded responses). Correlating the metrics with human judgements ([Table 2](#) and [Table 5](#)) reveals that **Recall** and **GPT4-Eval** are the best lexical and model-based metrics for correctness and **K-Precision** and **LLMCritic (GPT-4)** are the best lexical and model-based metrics for faithfulness, respectively. However, these model-based metrics, especially the ones based on LLMs, are usually slow to run, expensive, difficult to reproduce, and may exhibit systematic biases.

While we propose that Recall and K-Precision are the most widely-accessible and human-aligned metrics for correctness and faithfulness, respectively, we emphasize that these simple lexical-

based metrics are easy to hack. One model can copy all the retrieved knowledge as the output, leading to high Recall and K-Precision metrics. However, such a model will be penalized heavily when evaluated for faithfulness w.r.t. irrelevant knowledge.

Instruction-Following Models According to the most human aligned and easy to use metrics (i.e., Recall and K-Precision), we conclude that GPT-3.5 outperforms other models on majority of the datasets in correctness w.r.t information need. However, when analyzing the faithfulness w.r.t relevant knowledge, Flan-T5 is shown to be the best model in all three datasets. Moreover, our further analysis on the models’ faithfulness w.r.t irrelevant knowledge demonstrates that models struggle to correctly identify whether the provided knowledge is relevant or not.

Limitations It is worth mentioning that the experiments for evaluating the faithfulness of the models are conducted in a modified setting, where a relevant or irrelevant passage is provided in the prompt on purpose. This is different from the real-world scenario, where the retrieved passages can contain a mix of relevant and irrelevant knowledge.

Finally, it should also be noted that beyond qualitative investigation, we did not explore a wide range of prompts for the tasks studied in this work. Recent work has shown that the performance of instruction-following models can vary greatly depending upon the provided prompt ([Zhao et al., 2021](#); [Liu et al., 2023b](#)). We leave it to future works to investigate better prompts for instruction-following models in a retrieval-augmented setting.

7 Conclusion

We extensively study the capability of instruction-following models to correctly and faithfully respond to questions in three QA settings (natural, multi-hop, and conversational). First, we uncover various issues with using traditional metrics, like F1 score, to evaluate the correctness of models. Through correlation with human judgement, we find that LLM-based metrics (e.g. GPT-4) and token-level Recall are promising metrics for evaluating the correctness w.r.t information need. Moreover, our further faithfulness analysis shows that LLM-based metrics like LLMCritic (GPT-4) and lexical-based K-Precision are more aligned with human judgements in evaluating the faithfulness of

the models given the relevant knowledge.

Overall, we find that GPT-3.5 is better at providing correct responses for all tasks, whereas Flan-T5 comes out on top for faithfulness. However, all models struggle to accurately respond with “I don’t know” given an irrelevant passage when explicitly instructed to do so.

While Recall and K-Precision are the most human judgement aligned and widely-accessible alternative metrics, they are easy to hack. Therefore, we encourage the community to come up with more reliable metrics.

References

- Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topicqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.
- Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. [Open-domain question answering goes conversational via question rewriting](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online.
- Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Petr Baudis and Jan Sedivý. 2015. Modeling of the question answering task in the yodaqa system. In *Conference and Labs of the Evaluation Forum*.
- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. [Semantic parsing on Freebase from question-answer pairs](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.
- Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. [Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.
- Samuel R. Bowman. 2023. [Eight things to know about large language models](#).
- 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](#).
- Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. [Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality](#).
- Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabezudo, Arash Eshghi, Ioannis Papaioannou, Verena Rieser, and Ioannis Konstas. 2023. [The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 947–959, Toronto, Canada. Association for Computational Linguistics.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. [Deep Reinforcement Learning from Human Preferences](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi

- Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).
- Christopher Clark and Matt Gardner. 2018. [Simple and effective multi-paragraph reading comprehension](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.
- Nouha Dziri, Ehsan Kamaloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. [Faithdial : A Faithful Benchmark for Information-Seeking Dialogue](#). *Transactions of the Association for Computational Linguistics*, 10:1473–1490.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. [Faith and Fate: Limits of Transformers on Compositionality](#). ArXiv:2305.18654 [cs].
- Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. [Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022b. [On the origin of hallucinations in conversational models: Is it the datasets or the models?](#) In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. [The curious case of neural text de-generation](#). In *International Conference on Learning Representations*.
- Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. [Q²: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O’Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. [Opt-impl: Scaling language model instruction meta learning through the lens of generalization](#).
- Gautier Izacard and Edouard Grave. 2021. [Leveraging passage retrieval with generative models for open domain question answering](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. [Atlas: Few-shot learning with retrieval augmented language models](#).
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Ehsan Kamaloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. [Evaluating open-domain question answering in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. [Internet-augmented language models through few-shot prompting for open-domain question answering](#).
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. [Latent retrieval for weakly supervised open domain question answering](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. [Lost in the middle: How language models use long contexts](#).
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#). *ACM Comput. Surv.*, 55(9).
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. [G-eval: Nlg evaluation using gpt-4 with better human alignment](#).
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint*.
- Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y.-Lan Boureau. 2022. [Reducing conversational agents’ overconfidence through linguistic calibration](#). ArXiv:2012.14983 [cs].
- Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen-tau Yih. 2021. [Neurips 2020 efficientqa competition: Systems, analyses and lessons learned](#). In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 86–111. PMLR.
- Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. [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.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022a. [Training language models to follow instructions with human feedback](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022b. [Training language models to follow instructions with human feedback](#). ArXiv:2203.02155 [cs].
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Ashwin Paranjape, Omar Khattab, Christopher Potts, Matei Zaharia, and Christopher D Manning. 2022. [Hindsight: Posterior-guided training of retrievers for improved open-ended generation](#). In *International Conference on Learning Representations*.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.
- Peng Qi, Haejun Lee, Tg Sido, and Christopher Manning. 2021. [Answering open-domain questions of varying reasoning steps from text](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3599–3614, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Mari-beth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang,

- Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsim-poukelli, Nikolai Grigorev, Doug Fritz, Thibault Sotiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. [Scaling Language Models: Methods, Analysis & Insights from Training Gopher](#). ArXiv:2112.11446 [cs].
- 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](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and D. Reitter. 2021a. Measuring attribution in natural language generation models. *ArXiv*, abs/2112.12870.
- Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021b. [Increasing Faithfulness in Knowledge-Grounded Dialogue with Controllable Features](#). 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 704–718, Online. Association for Computational Linguistics.
- Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. [CoQA: A conversational question answering challenge](#). *Transactions of the Association for Computational Linguistics*, 7:249–266.
- Devendra Sachan, Mostofa Patwary, Mohammad Shoeibi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021. [End-to-end training of neural retrievers for open-domain question answering](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6648–6662, Online. Association for Computational Linguistics.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. [Multi-task prompted training enables zero-shot task generalization](#). In *International Conference on Learning Representations*.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. [Replug: Retrieval-augmented black-box language models](#).
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguerre-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. [LaMDA: Language Models for Dialog Applications](#). ArXiv:2201.08239 [cs].
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#).

- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#).
- Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. [Large language models are not fair evaluators](#).
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. [Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. [Multi-passage BERT: A globally normalized BERT model for open-domain 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 5878–5882, Hong Kong, China. Association for Computational Linguistics.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021. [Answering complex open-domain questions with multi-hop dense retrieval](#). In *International Conference on Learning Representations*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. [Opt: Open pre-trained transformer language models](#).
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#).

A Experimental Details

A.1 Instruction Model Details

To generate text, we use a high temperature to avoid repetitiveness in sampling, but further leverage top-p sampling (Holtzman et al., 2019) to avoid sampling words with very low frequency (which may lead to incoherent text being generated). The values used for all generation parameters are listed below:

- Top-p: $p = 0.95$
- Temperature: $t = 0.95$
- Seed: $s = 0$
- Min. new tokens: $\min_{token} = 1$
- Max. new tokens: $\max_{token} = 50$

A.2 Retriever Details

While the retriever remains constant for each task, the number of retrieved passages provided to instruction-following models and fine-tuned FiD varies. Instruction-following models are constrained by the input context size, hence, they receive fewer passages than fine-tuned FiD. For the conversational QA task, including the conversation history in the prompt further reduces the number of passages that can be incorporated into the input context. Despite the varying context sizes of different instruction-following models, we provide a consistent number of retrieved passages (denoted by K) for each model within a specific task to maintain fair comparison. The details are as follows:

- open-domain QA (NQ): $K = 8$
- multi-hop QA (HotpotQA): $K = 8$
- conversational QA (TopiOCQA): $K = 4$

Unlike instruction-following models, FiD is not restricted by input context size. We use the default settings for each dataset — 100 passages for NQ, 50 passages for TopiOCQA, and up to 18 passages for HotpotQA. For HotpotQA, the top 100 reasoning chains produced by the retriever are duplicated to generate the final passage set.

B Prompts details

In Section 4.2, we introduce LLM-based evaluations to evaluate the correctness of a model response w.r.t. the user’s information need. To accomplish this, we use the prompt template shown in Figure 7, and map “yes” to 1 and “no” to 0. Similarly, Section 5.1 introduces the LLMCritic

System prompt: You are CompareGPT, a machine to verify the correctness of predictions. Answer with only yes/no.

You are given a question, the corresponding ground-truth answer and a prediction from a model. Compare the “Ground-truth answer” and the “Prediction” to determine whether the prediction correctly answers the question. All information in the ground-truth answer must be present in the prediction, including numbers and dates. You must answer “no” if there are any specific details in the ground-truth answer that are not mentioned in the prediction. There should be no contradicting statements in the prediction. The prediction may contain extra information. If the prediction states something as a possibility, treat it as a definitive answer.

Question: {Question}
Ground-truth answer: {Reference answer}
Prediction: {Model response}

CompareGPT response:

Figure 7: The prompt template used for correctness evaluation.

System prompt: You are CompareGPT, a machine to verify the groundedness of predictions. Answer with only yes/no.

You are given a question, the corresponding evidence and a prediction from a model. Compare the “Prediction” and the “Evidence” to determine whether all the information of the prediction is present in the evidence or can be inferred from the evidence. You must answer “no” if there are any specific details in the prediction that are not mentioned in the evidence or cannot be inferred from the evidence.

Question: {Question}
Prediction: {Model response}
Evidence: {Reference passage}

CompareGPT response:

Figure 8: The prompt template used for faithfulness evaluation.

Evaluation method for calculating the faithfulness of the models w.r.t. relevant knowledge. To run this evaluation, we used the prompt shown in Figure 8.

Furthermore, we conducted other experiments to study the answer abstinence of the models in Section 5.3. The template used in these experiments is shown in Figure 9.

Category	Subcategory	Count	Percentage
Enumeration of Reference Answers	Enumeration of Reference Answers	21	7.17
Granularity Discrepancies	Temporal granularity discrepancy	4	1.37
Granularity Discrepancies	Spatial granularity discrepancy	10	3.41
Incomplete Reference Answers	List of Named Entities	13	4.44
Incomplete Reference Answers	Open-ended Questions	41	13.99
Incorrect Gold Answers	Incorrect Gold Answers	4	1.37
Intrinsic Ambiguity in Questions	Ambiguous Questions	12	4.10
Semantic Equivalence	Multinomial Entities	1	0.34
Semantic Equivalence	Synonymous Answers	8	2.73
Semantic Equivalence	More Elaborate Answers	163	55.63
Sufficient Subset	Sufficient subset	10	3.41
Symbolic Equivalence	Symbolic Equivalence	6	2.05

Table 7: Percentage share and exact counts of F1 failure cases by sub-category. See [Section 4.3](#) for more details.

```

Please answer the following question given
the following passages. If the answer is not
in the passages or cannot be inferred from
the passages, respond as "I don't know".
- title: {Passage title}
{Passage text}

- title: {Passage title}
{Passage text}
...
Question: {Question}
Answer:

```

Figure 9: The prompt template used for faithfulness w.r.t irrelevant knowledge.

C Failure Cases of Metrics

Lexical-based metrics [Figure 4](#) presents an overview of the F1 metric failures; the exact percentages and counts can be found in [Table 7](#).

GPT4-Eval To better understand how GPT4-Eval fails compared to F1, we took the subset of annotated failure cases (described in [Section 4.3](#)) where GPT-4Eval also predicts 0; In total, we found 70 instances out of overall 296 samples. [Figure 10](#) shows the distribution of failure subcategories for the GPT-4Eval subset. We observe that a higher proportion of failures are caused by open-ended questions, whereas more elaborated answers and enumeration of reference answers are less penalized by GPT4-Eval compared to the remaining failures shown in [Table 7](#). Moreover, all other subcategories now have a higher proportion due to the gap left by more elaborate answers and enumeration of reference answers. To illustrate the new findings, we include a few samples in [Figure 11](#).

D Human Evaluation

[Section 4](#) and [Section 5](#) describe the human evaluation procedures for both correctness of the responses w.r.t. information need and faithfulness of the models w.r.t. relevant knowledge.

[Table 8](#) demonstrates the quantitative results on the 100 samples picked for human evaluation using all studied correctness metrics. Similarly, the faithfulness results on the 50 samples are presented in [Table 9](#).

E Failure Cases of Models in Faithfulness w.r.t Irrelevant Knowledge

Results illustrated in [Table 6](#) show that models sometimes perform differently given relevant or irrelevant knowledge. [Figure 12](#) demonstrates the failure examples of the studied models in all three QA datasets. It can be observed that given an irrelevant passage, models (especially Alpaca) do not refrain from answering. Moreover, failure examples presented in [Figure 13](#) show that GPT-3.5 has difficulty in generating responses even when the correct information is available in the provided knowledge.

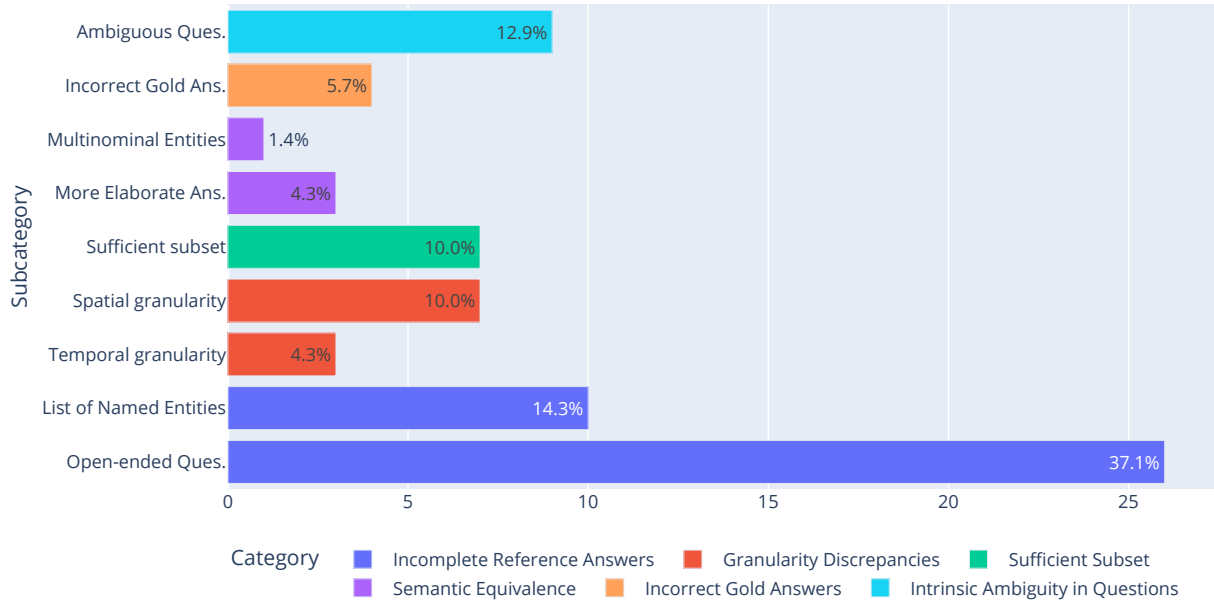


Figure 10: Distribution of failure cases of GPT4-Eval by sub-category. It struggles the most with *Open-ended Questions*.

Dataset	Model	EM	F1	Recall	Recall (S)	METEOR	Rouge-L	BertS (F1)	BEM	GPT4-Eval	Human Eval
NQ	FiD	66.00	<u>70.97</u>	72.83	72.0	58.76	<u>70.33</u>	94.88	75.18	72.0	82.0
	GPT-3.5	1.0	21.21	87.10	83.00	38.45	19.77	84.62	91.74	89.00	93.00
	Flan-T5	<u>65.00</u>	72.19	<u>77.73</u>	<u>75.00</u>	<u>58.56</u>	71.05	<u>94.74</u>	<u>80.36</u>	<u>81.00</u>	<u>89.00</u>
	Alpaca-7B	11.0	26.51	59.87	51.0	30.07	26.44	85.52	67.82	64.0	74.0
HotpotQA	FiD	<u>55.00</u>	<u>68.71</u>	68.73	63.0	<u>52.61</u>	<u>68.52</u>	<u>94.53</u>	74.78	70.0	77.0
	GPT-3.5	8.0	27.25	<u>78.83</u>	77.00	39.91	26.25	85.52	89.19	<u>81.00</u>	<u>82.00</u>
	Flan-T5	65.00	83.58	84.67	<u>76.00</u>	62.62	83.31	96.01	<u>87.72</u>	86.00	94.00
	Alpaca-7B	21.0	41.95	69.0	62.0	42.22	41.89	88.43	78.04	68.0	77.0
TopiOCQA	FiD	37.00	61.45	<u>63.55</u>	<u>44.00</u>	<u>54.91</u>	60.55	92.58	69.36	57.0	75.0
	GPT-3.5	4.0	44.85	79.41	46.00	57.18	41.94	89.93	82.59	84.00	88.00
	Flan-T5	<u>29.00</u>	<u>52.88</u>	63.0	44.0	50.32	<u>52.01</u>	90.90	67.39	<u>60.00</u>	<u>77.00</u>
	Alpaca-7B	7.0	32.86	44.24	22.0	36.18	31.98	87.26	52.42	41.0	52.0

Table 8: Results on 100 samples per dataset that were manually annotated for correctness w.r.t. information need. The best score is highlighted in bold and the second best is specified with underline. Traditional evaluation metrics like EM and F1 vary significantly from human evaluation in their judgments.

Dataset	Model	K-F1 ↑	K-Precision ↑	K-BertS (F1) ↑	FaithCritic ↓	Q2 (F1) ↑	Q2 (NLI) ↑	LLMCritic ↑	Human Eval ↑
NQ	GPT-3.5	24.57	76.53	86.70	75.85	54.44	60.77	94.87	<u>88.46</u>
	Flan-T5	4.81	96.15	81.01	19.28	<u>39.06</u>	<u>40.83</u>	89.74	92.31
	Alpaca-7B	<u>9.89</u>	<u>80.64</u>	<u>83.01</u>	<u>33.37</u>	35.54	37.23	82.05	83.33
HotpotQA	GPT-3.5	15.91	84.7	84.05	54.23	54.93	63.21	100.00	<u>96.81</u>
	Flan-T5	3.03	95.74	78.61	29.81	33.84	36.88	<u>100.00</u>	100.00
	Alpaca-7B	<u>11.50</u>	<u>87.55</u>	<u>83.12</u>	<u>49.09</u>	<u>45.63</u>	<u>51.95</u>	87.23	93.62
TopiOCQA	GPT-3.5	29.52	<u>80.80</u>	87.69	74.64	<u>65.25</u>	<u>70.92</u>	97.96	<u>88.78</u>
	Flan-T5	24.09	91.94	<u>86.70</u>	<u>59.61</u>	69.16	73.97	93.88	93.88
	Alpaca-7B	<u>24.58</u>	77.76	86.31	57.56	50.89	56.62	89.8	79.59

Table 9: Faithfulness results on 50 manually annotated samples. The best score is highlighted in bold and the second best is specified with underline.

Error Type: Incomplete Reference Answers - Open-ended Questions

Dataset: TopiOCQA

Question: does she (J.K. Rowling) donate to any charity?

Reference answer: Yes - In 2002, Rowling contributed a foreword to “Magic”, an anthology of fiction published by Bloomsbury Publishing, helping to raise money for the National Council for One Parent Families.

Alpaca’s response: Yes, she does. She is a member of both English PEN and Scottish PEN, and has donated the proceeds from several of her books to various charities.

Error Type: Incomplete Reference Answers - List of Named Entities

Dataset: Natural Questions

Question: what kind of animals live in the great victoria desert

Reference answer: malleefowl; water-holding frog; great desert skink; sandhill dunnart; dingo; sand goanna; crest-tailed mulgara; southern marsupial mole; chestnut-breasted whiteface; perentie

Flan-T5’s response: lizard.

Error Type: Intrinsic Ambiguity in Questions

Dataset: Natural Questions

Question: who played in the stanley cup finals last year

Reference answers: Nashville Predators, Pittsburgh Penguins

GPT-3.5’s response: The Washington Capitals and the Vegas Golden Knights played in the Stanley Cup Finals last year (2018).

Figure 11: Examples where GPT-4Eval predictions differ from human judgement (alongside $F_1 < 0.3$).

Dataset: NQ

Question: points on a sphere or angles in a circle are measured in units called

Reference answer: radians

Responses:

Alpaca: radians

Flan-T5: I don't know

GPT-3.5: degrees

Irrelevant knowledge: Title: Inscribed angle

inscribed angle theorem states that an angle θ inscribed in a circle is half of the central angle 2θ that subtends the same arc on the circle. Therefore, the angle does not change as its vertex is moved to different positions on the circle...

Question: who played the colorado kid in rio bravo

Reference answer: Ricky Nelson

Responses:

Alpaca: The role of the Colorado Kid was played by John Dehner in the 1959 film Rio Bravo.

Flan-T5: Jim J. Bullock

GPT-3.5: I don't know.

Irrelevant knowledge: Title: Jim J. Bullock

Jim J. Bullock James Jackson Bullock (born February 9, 1955), sometimes credited as Jm J. Bullock, is an American actor and comedian of stage, television and motion pictures... Bullock became a notable entertainment figure in the 1980s

Dataset: HotpotQA

Question: Ralph Hefferline was a psychology professor at a university that is located in what city?

Reference answer: New York City

Responses:

Alpaca: Columbia, Missouri.

Flan-T5: Columbia

GPT-3.5: I don't know.

Irrelevant knowledge: Title: University of Missouri

The University of Missouri (also, Mizzou, or MU) is a public land-grant research university located in Columbia, Missouri, U.S. It was founded in 1839 ...

Dataset: TopiOCQA

Question:

User: what is the location of mars in the solar system

Agent: Mars is the fourth planet from the Sun

User: is it known by any other names?

Reference answer: Red Planet

Responses:

Alpaca: No, it is known by its official name, Mars.

Flan-T5: Mars is also known as the Red Planet.

GPT-3.5: I don't know.

Irrelevant knowledge: Title: Haboob

Global dust storms on Mars have been compared to haboobs on Earth.

Figure 12: Failure examples of the models in generating *I don't know* in the presence of irrelevant knowledge.

Dataset: NQ

Question: points on a sphere or angles in a circle are measured in units called

Reference answer: radians

Responses:

Alpaca: 1. radians

Flan-T5: radians

GPT-3.5: I don't know.

Relevant knowledge: Title: Position angle

... It is calculated in a plane that contains the sphere center and the great circle, where is the angular distance of two points viewed from the center of the sphere, measured in **radians**...

Question: who had a baby at 100 in the bible

Reference answer: 1. Sarah, 2. Abraham

Responses:

Alpaca: 100 year old Sarah had a baby with Abraham.

Flan-T5: I don't know.

GPT-3.5: Abraham had a baby at 100 in the Bible.

Relevant knowledge: Title: The Bible and humor

... Macy says that in speaking of humor in the Bible he usually begins with asking how it might feel to personally get the message **Abraham and Sarah** received, putting them into the absurd situation of **a ninety year old having a baby**...

Dataset: HotpotQA

Question: Roger O. Egeberg was Assistant Secretary for Health and Scientific Affairs during the administration of a president that served during what years?

Reference answer: 1969 until 1974

Responses:

Alpaca: 1969-1974.

Flan-T5: 1969 until 1974.

GPT-3.5: I don't know.

Relevant knowledge:

Title: Roger O. Egeberg

... His other roles included Assistant Secretary for Health and Scientific Affairs in the Department of Health, Education, and Welfare (now the United States Department of Health and Human Services) during the **Nixon administration** and Dean of the University of Southern California's medical school.

- Title: Richard Nixon

Richard Milhous Nixon (January 9, 1913 – April 22, 1994) was the 37th President of the United States **from 1969 until 1974**, ...

Dataset: TopiOCQA

Question:

User: how much did the blues brothers movie cost to make

Reference answer: \$17.5 million

Responses:

Alpaca: 17.5 million

Flan-T5: \$17.5 million

GPT-3.5: I don't know.

Relevant knowledge: Title: The Blues Brothers (film)

Principal photography began in July 1979, with the film's budget still not settled. For the first month, things ran smoothly on and off the set. When Weiss saw the supposedly final **\$17.5 million** budget, he reportedly joked, "I think we've spent that much already." ...

Figure 13: Failure examples of the models in generating informative responses in the presence of relevant knowledge.