CHARP: Conversation History AwaReness Probing for Knowledge-grounded Dialogue Systems

Abbas Ghaddar[1♠] **David Alfonso-Hermelo**[1] **Philippe Langlais**[2]
Mehdi Rezagholizadeh[1] **Boxing Chen**[1] **Prasanna Parthasarathi**[1]
[1] Huawei Noah's Ark Lab
[2] RALI/DIRO, Université de Montréal, Canada
abbas.ghaddar@huawei.com

Abstract

In this work, we dive deep into one of the popular knowledge-grounded dialogue benchmarks that focus on faithfulness, FaithDial. We show that a significant portion of the FaithDial data contains annotation artifacts, which may bias models towards completely ignoring the conversation history. We therefore introduce CHARP, a diagnostic test set, designed for an improved evaluation of hallucinations in conversational model. CHARP not only measures hallucination but also the compliance of the models to the conversation task. Our extensive analysis reveals that models primarily exhibit poor performance on CHARP due to their inability to effectively attend to and reason over the conversation history. Furthermore, the evaluation methods of FaithDial fail to capture these shortcomings, neglecting the conversational history. Our findings indicate that there is substantial room for contribution in both dataset creation and hallucination evaluation for knowledge-grounded dialogue, and that CHARP can serve as a tool for monitoring the progress in this particular research area. CHARP is publicly available at https://huggingface.co/datasets/huawei-noah/CHARP

1 Introduction

Despite the success of general purpose large language models (LLMs) (Bommasani et al., 2021), the utility of the generated texts rests in its relevance and knowledge grounding. The task of information seeking dialogue (Ghazvininejad et al., 2018; Lewis et al., 2020) is a touchstone for knowledge grounded generation. The task evaluates a system's ability to respond to user queries while it remains faithful to the knowledge. A system response not adhering to this would be deemed unfaithful. This topic has received considerable attention resulting in several diagnostics and mitigation techniques for texts that lack knowledge grounding and are hallucinatory in nature (Dziri et al., 2019, 2022b,c).

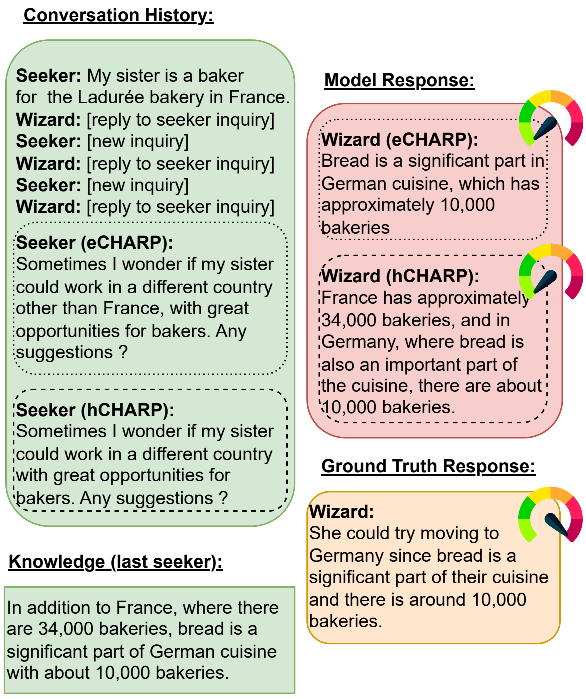

Figure 1: CHARP consists of 2 subsets, where only the last seeker utterance differs: a self-contained *easy* version (eCHARP), and a *hard* (hCHARP) which requires reasoning over the conversation history and the provided knowledge that corresponds to the last seeker. In addition to the ground truth response annotation, we show the predictions of a model (FLAN-base) tuned on the FaithDial training data. and indicate whether the FaithDial CRITIC labels a response as a hallucination or not. Green boxes indicate model inputs, while pink and orange ones show predicted, and gold responses.

Dziri et al. (2022a)'s work in this direction—FaithDial— provides a benchmark with hallucination-free annotations, a hallucination detector, and a comprehensive evaluation framework has garnered attention and follow up works (Deng et al., 2023; Daheim et al., 2023). Dziri et al. (2022a) show that T5-base model

♠Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2024, pages 1534–1551

trained on these annotations restricts hallucination only in 1.4% of its responses , or even 0.3% as reported by (Daheim et al., 2023). While it appears that hallucination is *under control* at least under the experimental protocol defined in FaithDial, we observe that, though the annotations created in (Dziri et al., 2022a) are free from hallucinations, they introduce artifacts. These artifacts bias models trained on it to predict the response based solely on the provided knowledge, while ignoring the dialogue history.

We validate this hypothesis with a controlled evaluation set called CHARP (**C**onversation **H**istory **A**wa**R**eness **P**robing) with its *easy* and *hard* versions denoted as eCHARP and hCHARP respectively. The proposed diagnostic set (§4) not only evaluates hallucinations with respect to the provided knowledge but also its relevance to the conversation history (Figure 1). CHARP is created by annotating on top of 1,080 samples from FaithDial validation dataset. CHARP tests whether models attend or ignore the history to select the appropriate knowledge when the correct knowledge is augmented with a distracting fact that is irrelevant to the conversation.

Evaluating models using automatic metrics, LLM APIs, and human scorers, we find that training with FaithDial biases the models to ignore conversation history, as assessed with CHARP, while remaining faithful to the knowledge (§3.5). Interestingly, this phenomenon was elusive to be observed neither with the suite of evaluation methods nor the hallucination detector proposed in (Dziri et al., 2022a). Instead, we find the FaithDial detector scoring CHARP *gold* responses as hallucinatory (16.0%) that is higher than the hallucination rate (0.4%) of a system performing poorly on CHARP as evaluated by a human (§5.1).

To understand this, we conduct a thorough human evaluation to identify 6 different types of errors to be considered in knowledge-grounded response generation. We find human annotation to be effective (§5.2) in identifying the error types. Further ablations with FaithDial training on human evaluation confirm that the dataset biases the models to look away from the conversation history. We find the evaluation with powerful LLM APIs to be correlated with humans (§C.3), and is a better proxy metric for this task over the FaithDial metrics. Overall, this study suggests that despite recent progress reported in hallucination mitigation, developing a model that is simultaneously aware of the conversation history and non-hallucinatory remains an open problem in information-seeking dialogue.

2 Related Work

Constructing diagnostic sets with curated adversarial or counterfactual examples, has been shown to be an effective approach across NLP tasks to capture such artifacts that standard evaluation sets fail to detect. For instance, HANS (McCoy et al., 2019), FEVER (Schuster et al., 2019), PAWS (Zhang et al., 2019b), CORE (Rosenman et al., 2020), NRB (Ghaddar et al., 2021), and NATURE (Alfonso-Hermelo et al., 2021) datasets are vital in identifying biases of models solving tasks like textual entailment, fact verification, paraphrase identification, relation extraction, named entity recognition, and intent detection respectively.

Studies (Taori et al., 2023; Chen et al., 2023a; Conover et al., 2023) on the recent trend of large-scale pretraining (Ouyang et al., 2022; Shuster et al., 2022) show that data quality affects the models in inheriting biases from the artifacts embedded in data. Especially in information seeking dialogues, Dziri et al. (2022c) show that unfaithfulness to the given knowledge is a dominant type of hallucination. Dziri et al. (2022b) show that most information-seeking dialogue datasets like CMU-DoG (Zhou et al., 2018), TopicalChat (Gopalakrishnan et al., 2019), and Wizard of Wikipedia (WoW; Dinan et al. 2018) contain a high ratio of hallucinations, with WoW dataset being the least affected. Dziri et al. (2022a) propose FaithDial built through replacing hallucinatory WoW annotations with with knowledge faithful responses.

Dziri et al. (2022a) show that the models trained on Faithdial show significant reduction in hallucination. Further studies such as the one from Daheim et al. (2023) recently demonstrated that FLAN-T5-base (Longpre et al., 2023) can fetch further reduction in hallucination by training on FaithDial over T5-Base (Raffel et al., 2019) as reported in (Dziri et al., 2022a). Daheim et al. (2023) also propose Elastic Weight Removal (EWR) hallucination mitigation method to reduce the rate of unfaithfulness response generation. The authors report high BERTScore similarity (Zhang et al., 2019a) between the model response and both the ground truth and the provided knowledge suggesting the faithfulness of the response generated. As training an LM to attend to the knowledge could inadvertently result in ignoring the history of turns leading to poor reasoning of the model, we, in this work, propose CHARP that serves as a diagnostic set to measure this phenomenon.

3 Experimental Setting

3.1 Dataset and Task

We focus on the FaithDial dataset (Dziri et al., 2022a) and adhere to its task formulation, where given the history of utterances and a knowledge supplement, a trained model predicts the next response of a *Wizard* bot engaged in a conversation with an information-seeking human (*Seeker*). We assume that the correct knowledge is given, and no retrieval step is performed. A response is considered hallucinatory if it contains information unsupported by the given knowledge snippet.

3.2 Models and Implementation

We experiment with the vanilla T5 model (Raffel et al., 2019) and two of its derivative variants, namely, Flan-T5 (Chung et al., 2022) and GODEL (Peng et al., 2022). The former was finetuned on $1,000$ NLP datasets mapped to an instruction tuning format, while the latter was further pre-trained on 551M multi-turn dialogues, and 5M instruction- and knowledge-grounded dialogues. We primarily focus on the *base* size models, maintaining the same hyperparameters and implementation settings for consistency with previous works (Dziri et al., 2022a; Daheim et al., 2023). We train all models for a maximum of 20 epochs and use early stopping based on the validation set performance, and report results on the test set. We use a beam of 5 during inference in all experiments.

3.3 Evaluation Metrics

Following (Dziri et al., 2022a; Daheim et al., 2023), we report the similarity between gold (y) and predicted response (y') with BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2019a). We measure the hallucination rate using the faithfulness Critic (CRITIC) provided by the FaithDial benchmark[1] and by computing a BERTScore between the knowledge (k) and predicted response.

3.4 Results Integrity

Table 1 is the reproduced test performances of models trained on the FaithDial dataset. For comparable evaluation we consider the baseline models from (Dziri et al., 2022a): a vanilla T5-base model and its variant that employs InfoNCE (Oord et al., 2018) loss for hallucination mitigation; and (Daheim et al., 2023): the `FLAN`-base model and its variant that utilizes the EWR method for hallucination mitigation. Our reproduced baselines use three different backbone models: `FLAN`-base, and the `GODEL`-base and `GODEL`-large models.

Models	BLEU↑ (y, y')	Critic↓ (k, y')	BERTScore↑ (y, y')	BERTScore↑ (k, y')
previous works				
T5-base [1]	10.3	4.3	-	41.0
+InfoNCE	10.9	1.4	-	39.0
`FLAN`-base [2]	15.1	0.3	69.6	80.9
+EWR	14.9	0.1	70.1	81.7
our re-implementations				
`FLAN`-base	15.3	0.3	69.9	80.8
`GODEL`-base	15.5	0.3	70.2	80.5
`GODEL`-large	15.8	0.3	70.5	81.1

Table 1: Test set performances of previous works alongside our re-implemented baseline models finetuned on the FaithDial dataset. [1] and [2] refer to baselines results directly copied from (Dziri et al., 2022a) and (Daheim et al., 2023) respectively. All scores are scaled within the range of $[0, 100]$.

As reported in (Daheim et al., 2023) the `FLAN`-base models achieve a remarkably low hallucination ratio of only 0.3%, significantly improving upon the best FaithDial baselines (both with and without hallucination mitigation methods). Our reimplementation of the `FLAN`-base results are similar to Daheim et al. (2023)'s based on the significance test on the samples with p-value (0.58). This enables a fair comparison of the claims across our experiments with CHARP and the existing results in the literature.

In addition, we used the set up to train different models for benchmarking: a dialogue pretrained (`GODEL`-base) and its large version (e.g., `GODEL`-large). We observe that these models only hold modest improvements across various metrics.

3.5 Probing History Awareness

While the strong results across models on FaithDial dataset leave no doubts about their faithfulness to the given knowledge, we investigate whether this comes at the expense of an important input component: the conversation history. To that, we test trained models on truncated conversation history—providing only the last k turns (denoted h=k) or no history at all (h=$\emptyset$). As we observed the value of k

[1] A detailed description of the development of the FaithDial CRITIC can be found in Appendix B.1.

in the dataset to not affect the performance beyond 3 despite the average number of turns being 7, we vary the value of k only in the range of [0-3]. Further, to account for distribution shifts, we fine-tune models on variations of the training data with truncated conversation histories for comparison. We used GODEL-base as the backbone model in this experiment as it shows superior performances compared to FLAN-base in Table 1. We benchmark the performance of GODEL-base on truncated history evaluation in Table 2.

	BLEU↑ (y, y')	Critic ↓ (k, y')	BERTScore↑ (y, y')	BERTScore↑ (k, y')
eval: h=all				
train: h=all	15.5	0.3	70.2	80.5
train: h=3	15.4	0.4	70.0	79.2
train: h=2	15.2	0.5	69.9	78.7
train: h=1	15.0	0.6	69.8	78.2
train: h=$\emptyset$	11.9	8.4	65.8	72.8
eval: h=3				
train: h=all	15.3	0.4	69.9	79.9
train: h=3	15.1	0.3	69.9	80.4
eval: h=2				
train: h=all	15.1	0.3	69.9	80.6
train: h=2	15.0	0.2	69.9	80.9
eval: h=1				
train: h=all	14.3	0.3	69.4	81.1
train: h=1	14.3	0.2	69.5	81.3
eval: h=$\emptyset$				
train: h=all	13.1	0.0	66.6	82.1
train: h=$\emptyset$	12.7	0.0	67.1	83.9

Table 2: Performance of GODEL-base models, trained and evaluated on truncated versions of conversation history from the training and test splits of FaithDial, respectively. Here, h=i means only using the last i turns in the conversion history when training (train:) or evaluating (eval:) models. h=$\emptyset$ and h=all denote using no history and the entire history turns, respectively. All scores are scaled within the range of $[0, 100]$.

We notice that the performances on the original test set (eval: h=all) of models trained on truncated history (train: $h \in \{3, 2, 1\}$) barely drop across metrics. For instance, the hallucination ratio (CRITIC score) slightly increases by 0.1% each time the conversational history contains one fewer turn. Although there is a partial train/test mismatch, this observation suggests that the older history turns are largely irrelevant to generating the response, and their presence does not significantly distract the models. However, we report a significant loss in performance across metrics in the extreme case where no history is provided to the model during training (train: h=$\emptyset$). Through manual inspection of samples, we hypothesize that this is due to the model treating the entire history as a knowledge snippet and attempting to ground the response under this assumption.

In evaluation configurations with $h \in \{3, 2, 1\}$, we observe that the performances of the original model and its respective variants are roughly similar, exhibiting only a slight decline as more history turns are removed. More precisely, ground truth response similarity metrics show a steady decline, while those measuring similarity with the provided knowledge slightly improve as fewer history turns are seen during training and/or evaluation. These results indicate that the conversation history is ignored not only during inference but also during model training.

However, when the entire conversational history is omitted during evaluation (eval: h=$\emptyset$), the original model produces responses that are slightly better aligned with the ground truth response, showing a +0.4 gain in BLEU score and a +0.5 gain in BERTScore compared to the model trained without conversation history (train: h=$\emptyset$). Notably, both models achieve the best alignment with the given knowledge results so far, reaching 82.1% and 83.9% respectively, and report a 0% hallucination ratio as per the FaithDial CRITIC. It is worth mentioning that the model trained without any history performs significantly better across all four metrics when the history is also discarded during inference. This suggests that this model is most likely learning primarily to paraphrase a knowledge snippet into its response. From these analyses, we conclude that both the annotation strategies and evaluation methods of FaithDial do not take into account a crucial scenario in information-seeking dialogue, where a valid response depends on understanding and reasoning about the conversation history.

4 CHARP

CHARP, the proposed diagnostic set, exclusively assess whether information-seeking dialogue systems effectively attend to and use the conversation history. CHARP is built by modifying examples from the FaithDial validation set to ensure maximum domain alignment with FaithDial and to minimize annotation costs. That is, we edit FaithDial examples to make their response dependent on the conversation history analogously to

FaithDial's editing of WoW annotations to make them hallucination-free. It is important to note that the FaithDial validation and test sets are sampled from the same distribution and exhibit similar result patterns. We create two variants of CHARP: hCHARP (§ 4.1) for examples where addressing the last seeker's inquiry requires reasoning over the conversation history, and eCHARP (§ 4.2), where the last inquiry can be addressed without such reasoning. We annotate 42% of the FaithDial validation set (after excluding examples without conversation history) containing $2,160$ examples, split equally between hCHARP and eCHARP.

4.1 hCHARP Creation

In hCHARP, which refers to *hard* CHARP, examples are expected to test basic natural language understanding abilities, with the expectation that the response will be straightforward if a model is attentive to the conversation history. An example is designed to test the ability to resolve co-reference relations and the history mentions "*my favorite color is red*", then the last user turn might be "*Wondering which fruit has the same color as my favorite?*". In other examples, annotators may introduce temporal reasoning (e.g., *the pre-historic era is before the 16th century*), geospatial (e.g., *Paris* is located in *France*), or taxonomic (e.g., *pie* is a type of *patisserie*). To ensure systematic and high-quality annotations, we define a set of edit rules for each part of the example (ref § A.2).

4.2 eCHARP Creation

We create a set of domain control examples — eCHARP for *easy* CHARP— that contains the same examples in hCHARP, but with the last user turn rewritten to be self-contained and independent from the conversation history. For instance, if the knowledge lists fruits that are typically green and others that are red, the last user turn would be "*Wondering which fruit is typically red?*". Thus, responding to examples in eCHARP should be easy. Poor performance on such examples[2] suggests a domain (or task definition) shift between CHARP and the information-seeking dataset on which the model is trained. Further, in both the versions we annotate the *knowledge* in the FaithDial dataset to provide a relevant, and a distracting factual information. The distracting information is designed to be ignored if the conversation history is considered in the knowledge selection by the models.

[2] e.g., a system that copies or rephrases the entire knowledge and ignores even the last user turn.

5 Results

5.1 CHARP Automatic Evaluation

The performances of the FLAN-base and GODEL-base models on CHARP, and on the subset of FaithDial validation examples that were utilized to construct CHARP are shown in Table 3. First, we observe that the distribution of performances on the validation set are similar to the test set (Table 1) as they are sampled from the same distribution. Further, the results on the full validation set align with those on the subset used to build CHARP indicating that the sampled set from FaithDial used in CHARP does not introduce any bias to this study.

Models	BLEU↑ (y, y')	Critic ↓ (k, y')	BERTScore↑ (y, y')	BERTScore↑ (k, y')
FaithDial Valid.				
FLAN-base	14.6	0.3	70.6	80.7
GODEL-base	14.8	0.3	70.8	80.6
FaithDial Valid. (CHARP *subset)*				
FLAN-base	14.6	0.4	71.1	81.2
GODEL-base	14.5	0.3	70.8	81.5
*e*CHARP				
FLAN-base	22.8	1.7	70.8	79.4
GODEL-base	20.5	0.5	69.4	82.5
*h*CHARP				
FLAN-base	22.0	1.9	70.1	78.6
GODEL-base	18.7	0.6	67.6	81.7

Table 3: Performance of models on the FaithDial dataset across four evaluation sets: FaithDial validation set, a subset of the FaithDial validation set used to build CHARP, eCHARP, and hCHARP. All scores are scaled within the range of $[0, 100]$.

Unsurprisingly, we observe that the models perform well on eCHARP with better BLEU scores in (y, y'), and almost similar on both (k, y') metrics compared to the results on the validation set. The high BLEU scores arise because our ground truth responses have a high lexical overlap with the knowledge, involving less paraphrasing, compared to the FaithDial data. As the with and without paraphrasing the responses are semantically similar, the BERTScores remain similar to the one on the validation set. We also notice that GODEL-base is less hallucinatory than FLAN-base, performing better on both (k, y') metrics. Conversely, FLAN-base outperforms GODEL-base on both (y, y') metrics.

Despite the hCHARP responses being strongly dependent on information from the conversational history, we notice that the models perform surprisingly well. When comparing the score ranges to those on eCHARP, we observe that, although the results are consistently lower across all metrics the difference was not significant. These observations strongly contradict our hypothesis in § 3.5, which posits that models ignoring conversational history should incur significant penalties across all metrics. We examine the metrics themselves by computing the CRITIC and BERTScore between the knowledge snippets and the gold responses in both the FaithDial validation and test sets, as well as in CHARP.

	Valid	Test	CHARP
Critic $(k, y) \downarrow$	0.4	0.4	16.0
BERTScore $(k, y) \uparrow$	84.3	85.6	69.9

Table 4: Evaluation of the ground truth response (y) when contrasted with the knowledge snippet (k) on FaithDial validation and test sets as well as on CHARP. All scores are scaled within the range of $[0, 100]$.

Results in Table 4 show that the ground truth responses of CHARP are labeled as extremely hallucinatory beyond not only the other gold responses in the compared sets of FaithDial but also to model predictions. The hallucination ratio is 27× higher for the ground truth (16%) compared to the response generated by GODEL-base (0.6%). Additionally, the semantic similarity with the knowledge, as measured by the BERTScore, is roughly more than 12% lower between the ground truth (69.9%) and GODEL-base response (81.7% on hCHARP). This contrasts with the scores on the FaithDial evaluation sets, where we observe a close tie with the models' responses. These observations indicate possible deficiencies in the metrics used in FaithDial and their comprehensiveness as the success through the lens of the metric lies at models focusing on the knowledge segment.

5.2 CHARP Human Evaluation

We employ our human annotators (ref §A for details) to carry out a comprehensive analysis of the models' outputs. We focus on errors related to the system's reasoning over knowledge and conversation history. Through this evaluation, we emphasize faithfulness to the provided knowledge while including aspects such as cooperativeness, engagingness, and abstractiveness as motivated in (Dziri et al., 2022a). We frame the evaluation process in the form of a checklist ticker (binary classification), where each annotator is tasked with labeling whether a system response:

- $\mathcal{C}1$ properly addresses the seeker comment.
- $\mathcal{W}1$ addresses the seeker's comment while adding extra information not in the provided knowledge.
- $\mathcal{W}2$ is simply a copy or slight paraphrasing of the entire knowledge, while part of it is irrelevant.
- $\mathcal{W}3$ states a lack of knowledge (e.g. *I don't know*) but still copies or rephrases part or all of the provided information, despite the relevant information existing within the provided knowledge.
- $\mathcal{W}4$ is a copy or slight paraphrasing of an irrelevant knowledge segment.
- $\mathcal{W}5$ fuses knowledge segments leading to wrong or contradictory information.
- $\mathcal{W}6$ is incorrect for other reasons, e.g. fully detached, severe hallucination, contains contradictory information.

Table 5 shows the human evaluation results of FaithDial trained FLAN-base and GODEL-base models on CHARP subsets. First, we observe that both models exhibit poor performance on CHARP, which slipped through FaithDial's automatic metrics (Table 3). As expected, hCHARP proves to be more challenging than eCHARP, with the correct response ratio ($\mathcal{C}1$) significantly dropping by 13% for FLAN-base and 8% for GODEL-base.

	FLAN-base		GODEL-base	
	eCHARP	hCHARP	eCHARP	hCHARP
$\mathcal{C}1\ \uparrow$	23%	10%	21%	13%
$\mathcal{W}1 \downarrow$	0%	0%	0%	1%
$\mathcal{W}2 \downarrow$	49%	49%	34%	36%
$\mathcal{W}3 \downarrow$	3%	14%	7%	12%
$\mathcal{W}4 \downarrow$	20%	20%	32%	31%
$\mathcal{W}5 \downarrow$	4%	7%	6%	7%
$\mathcal{W}6 \downarrow$	1%	0%	0%	0%

Table 5: Human evaluation results of GODEL-base and FLAN-base models on eCHARP and hCHARP.

	Finetuning		*3-shot*					
	Llama-2-7B		Llama-2-7B		Mixtral		ChatGPT	
	eCHARP	hCHARP	eCHARP	hCHARP	eCHARP	hCHARP	eCHARP	hCHARP
$\mathcal{C}1$ ↑	36%	27%	26%	13%	71%	64%	66%	56%
$\mathcal{W}1$ ↓	0%	0%	34%	37%	18%	14%	18%	13%
$\mathcal{W}2$ ↓	32%	35%	12%	9%	4%	7%	3%	6%
$\mathcal{W}3$ ↓	7%	13%	0%	0%	0%	0%	2%	3%
$\mathcal{W}4$ ↓	20%	19%	0%	0%	1%	2%	0%	1%
$\mathcal{W}5$ ↓	4%	5%	7%	9%	4%	9%	5%	9%
$\mathcal{W}6$ ↓	1%	1%	21%	32%	2%	4%	6%	12%

Table 6: Human evaluation results on CHARP for models under fine-tuning, 3-shot learning paradigms.

We see that models do not add out-of-context information ($\mathcal{W}1$) or suffer from severe hallucinations ($\mathcal{W}6$), as these error rates are almost null across all configurations. This is largely expected as FaithDial training reinforces this in models. We also note that 60% the responses contain paraphrased *all* knowledge including the irrelevant fact $\mathcal{W}2$, or only the irrelevant knowledge $\mathcal{W}4$. While these samples are marked as errors by humans, automatic metrics struggle to identify the same (Table 4). As spotting the knowledge chosen is relevant or not is contingent on knowing the conversation so far, metrics focusing on only the knowledge fails unlike humans considering the context as well.

We observe that only $\mathcal{W}3$ errors show a significant increase when comparing the performances on eCHARP and hCHARP: 11% for FLAN-base and 5% for GODEL-base, respectively. This observation is particularly interesting as it directly relates to the FaithDial annotation guide, which instructs the annotators to write responses where the bot acknowledges its ignorance and continues the conversation by presenting the given knowledge engagingly when the knowledge cannot satisfactorily address the seeker's last inquiry. This means that a model that finds a knowledge to be relevant for an example in eCHARP, may find it irrelevant in its corresponding example in hCHARP that is attributed to FaithDial trained models' shortcoming to reason over the conversation history.

Interestingly, we notice that the performances are roughly equal under some categories (mainly $\mathcal{W}2$ and $\mathcal{W}4$) when comparing eCHARP and hCHARP results. This is noteworthy because, in eCHARP, models do not need to rely on previous conversation history to respond, as the last utterance is designed to be self-contained. In contrast, hCHARP is designed to assess whether models consider the entire conversation history. For example, a model that simply copies the entire knowledge segment ($\mathcal{W}2$), without considering the content of the last user utterance, is effectively ignoring the entire history. This observation suggests that the errors noted are not due to the model's inability to reason based on earlier conversation turns.

6 Analysis

6.1 *On FaithDial Data Artifact*

We conduct ablations on the behavior of models to estimate the resulting errors due to the artifacts in the FaithDial training data by comparing the performances of models with and without fine-tuning on FaithDial. We use the few-shot learning technique via prompting to ablate for not training with FaithDial dataset. Specifically, we compare the performances of Llama-2-7B model that we tuned on FaithDial against the same model with 3 in-context examples (*3-shot*). In addition, we evaluate the *3-shot* performances of ChatGPT (OpenAI, 2022) and Mixtral (Jiang et al., 2024) LLMs. Implementation details of these experiments can be found in Appendix B.3, as well as an example of the models' responses in Figure 5.

In Table 6 we compare the human evaluation results on eCHARP and hCHARP[3] across the different error types. While the fine-tuned Llama-2-7B reports higher [4] performance ($\mathcal{C}1$) than FLAN-base and GODEL-base (Table 5), due to its larger size, we believe it also suffers from a lack of reasoning behavior. This is suggested by the reported error

[3]We also measured inter-annotator agreements across all models and sets, and have reported the results in § C.2.

[4]although it underperforms on automatic evaluation metrics. This aspect is further discussed in Appendix C.1.

trends, where $\mathcal{W}2$ and $\mathcal{W}4$ are the dominant error categories, in contrast to $\mathcal{W}1$, $\mathcal{W}5$, and $\mathcal{W}6$. However, the error trends undergo a drastic change when comparing the results of fine-tuned *3-shot* models.

First, we observe that $\mathcal{W}1$ is the dominant error category, where models add extra information after addressing the user inquiry. We attribute this behavior to the verbose nature of LLMs, which is challenging to mitigate without further tuning (Gudibande et al., 2023). Second, we notice that the error ratio for $\mathcal{W}2$ is significantly lower, by at least 20%, across configurations compared to fine-tuned models. Additionally, we observe near-zero error ratios for $\mathcal{W}3$ and $\mathcal{W}4$, strongly suggesting that models not tuned on FaithDial are not affected by their ability to reason over the history. The error trend of `Llama-2-7B` trained on FaithDial strongly mimicking the results in Table 5, we confirm with a high probability that the LMs lose the ability to account for conversation after being fine-tuned on FaithDial.

Third, non-tuned models suffer significantly more from severe hallucinations ($\mathcal{W}6$), a well-known issue with LLMs (Ji et al., 2023; Ye et al., 2023). However, this issue tends to be mitigated as the models get larger; for instance, $\mathcal{W}6$ drops from 32% in `Llama-2-7B` to 5% in `Mixtral` on hCHARP. While the smaller `Llama-2-7B` LLM performs worse than its fine-tuned version, the larger models, `Mixtral` and `ChatGPT`, significantly outperform all reported fine-tuned models by at least 30% on $\mathcal{C}1$. Despite their high performances, we observe that hCHARP remains more challenging than eCHARP for even the most advanced models, indicating that CHARP can additionally serve as a measure of reasoning capability for such models. Finally, the fact that `Mixtral` outperforms `ChatGPT` in our tasks serves as an additional indicator that high performance is achievable through community-shared, open-source models.

6.2 *On FaithDial Evaluation Metrics*

Despite being highly accurate, human evaluation is time and resource-consuming which limits its scalability and practicality on large evaluation sets. To this end, we investigate the recent trend (Wang et al., 2023; Liu et al., 2023b; Hackl et al., 2023; Li et al., 2024) of utilizing LLM APIs for the open-ended evaluation of NLP systems. More specifically, our focus is on using `GPT4-turbo`, a cost-efficient version of GPT-4 (OpenAI, 2023). This family of models has been shown to correlate with human judgment, outperforming other alternatives (Chiang et al., 2023; Liu et al., 2023a) in their generation. The exact prompt we used and a detailed description of this experiment are presented in Appendix C.3.

In Figure 2 the normalized contingency table is displayed as a heatmap showcasing the agreement between `GPT4-turbo` and human judgments of the `Llama-2-7B` fine-tuned models on both eCHARP and hCHARP. The table counts the frequency of each combination of categories from `GPT4-turbo` and human judgments, which we later normalized into percentages ($[0-100]$). We show the contingency table corresponding to fine-tuned `Llama-2-7B` model, and other models[5] also show a similar trend.

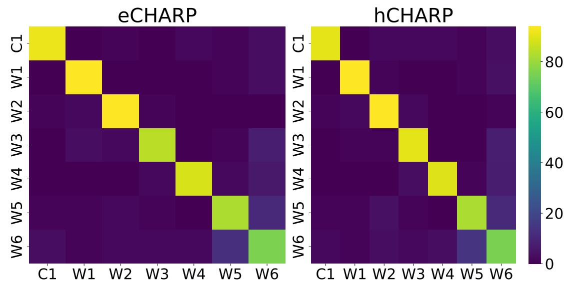

Figure 2: Heatmap showing the normalized (percentage) contingency tables of evaluation categories between `GPT4-turbo` (rows) and human (columns) judgments. It was measured on the output of `Llama-2-7B` (finetuned) for both eCHARP (left) and hCHARP (right).

First, it is worth mentioning that the Kappa (Carletta, 1996) agreement score on the overall examples of eCHARP and hCHARP are 0.89 and 0.88, respectively. These scores are higher than the 0.8 well-acceptable threshold, indicating a high overall correlation. This observation is further evidenced by the high values along the diagonals of the correlation heatmaps for both subsets. More precisely, we observe that the correlation is higher for the correct response category ($\mathcal{C}1$), as well as for other wrong response categories that are relatively easy to detect, such as $\mathcal{W}1$, $\mathcal{W}2$, and $\mathcal{W}4$. However, we observed that `GPT4-turbo` tends to confuse certain categories, notably $\mathcal{W}3$ (which involves stating lack of knowledge while still providing relevant information) and $\mathcal{W}5$ (fusing the irrelevant knowledge segments), with the severe hallucination category ($\mathcal{W}6$). Although not a perfect match, we

[5]Detailed plots for all six models, as well as kappa agreement scores, can be found in Figure 4 in the Appendix C.3.

believe that powerful LLM currently represents the best approximation to human annotations instead of the weak automatic evaluation metrics.

7 Conclusion

In this work, we examine the impact of annotation artifacts on information-seeking dialogue models tuned on FaithDial, a well-established, hallucination-free annotation benchmark. We introduce CHARP, a diagnostic set designed to evaluate the ability of models to reason over the conversation history, while also staying grounded on the knowledge. Our analysis with CHARP reveals a strong correlation between training on FaithDial to models' ignoring reasoning over the conversation history. Further, proprietary LLM APIs can be a proxy to human evaluation, and a better hallucination estimator to automatic metrics. In similar vein to (Chen et al., 2023b), we note that while it is important to ensure hallucination-free annotations, including examples to cover reasoning over context and other pretraining knowledge is necessary to preserve models' reasoning capabilities.

Limitations

Potential limitations of this work could be stemming from the sampling of dataset to conduct the study. Although the study focuses primarily on FaithDial dataset, the other existing datasets have been shown to contain more hallucinations rendering this a minor issue. Further, the knowledge grounded dialogue generation has not looked at the generated texts that pertains to a diverse demographics. This is largely due to the nacency of this domain and further studies can alleviate this issue.

Acknowledgements

We would like to thank Imad Mousaoui, Ella Cho, Abdulmuizz Yusuf, and Parminder Singh Bharot, the professional annotators without whom this work would have not been possible. We thank the anonymous reviewers for their insightful comments.

References

David Alfonso-Hermelo, Ahmad Rashid, Abbas Ghaddar, Philippe Langlais, and Mehdi Rezagholizadeh. 2021. Nature: Natural auxiliary text utterances for realistic spoken language evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023a. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023b. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, et al. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation.

Yifan Deng, Xingsheng Zhang, Heyan Huang, and Yue Hu. 2023. Towards faithful dialogues via focus learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Nouha Dziri, Ehsan Kamalloo, 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, Ehsan Kamalloo, and Kory W Mathewson Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of NAACL-HLT*, pages 3806–3812.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar R 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.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022c. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Abbas Ghaddar, Philippe Langlais, Ahmad Rashid, and Mehdi Rezagholizadeh. 2021. Context-aware adversarial training for name regularity bias in named entity recognition. *Transactions of the Association for Computational Linguistics*, 9:586–604.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is gpt-4 a reliable rater? evaluating consistency in gpt-4 text ratings. *arXiv preprint arXiv:2308.02575*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *In International Conference on Learning Representations*.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. 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 1699–1713.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2022. ChatGPT: Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Prasanna Parthasarathi, Joelle Pineau, and Sarath Chandar. 2020. How to evaluate your dialogue system: Probe tasks as an alternative for token-level evaluation metrics. *arXiv preprint arXiv:2008.10427*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *arXiv preprint arXiv:2206.11309*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Shachar Rosenman, Alon Jacovi, and Yoav Goldberg. 2020. Exposing shallow heuristics of relation extraction models with challenge data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3702–3710.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3410–3416.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. Paws: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713.

A Annotation Guideline

Although editing is generally faster than creating new examples from scratch, the numerous constraints that must be met in a single example could make the task time-consuming if annotators are not provided with *template-like* instructions. Therefore, we decided to restrain our annotators to introducing edits that probe two natural language understanding abilities: solving co-reference relations and performing simple reasoning. Figure 3 presents an example from the FaithDial validation set, transformed into an hCHARP annotation according to the rules below.

A.1 Annotation Process

We hire 4/ 20 interviewed annotators as contractors over prior experience and a predefined test. For the ≈ 5700 annotation hours, annotators were paid 19 USD/hour. The annotators were trained beforehand and given guidelines containing instructions and examples of both typical and radical cases they might encounter during annotation. In addition, domain experts were revising the annotations daily and conducting video meetings with annotators whenever necessary. Typically, an annotator would receive an example comprising <conversation history; last user turn; knowledge; response>, which he/she is required to edit according to the guidelines outlined in the following section. Despite our efforts to simplify the annotation task and guide the annotators, the process was still considered slow, with annotators averaging only 8 examples per hour. This slow pace was due to the need to adhere to both the original FaithDial guidelines and all our additional conditions outlined below.

A.2 Annotation Rules

Conversation history We give the annotators the freedom to fully rewrite the response, knowledge, and last user turn, while not introducing changes to the conversation history unless they find it necessary (and if so, to make the minimal possible edits). Also, we strongly encourage them to maintain the natural flow of the conversation and stick to the information-seeking dialogue style.

Last user turn The last user turn should straightforwardly seek a specific piece of information and be answerable only when referencing the conversation history. We instruct our annotators to avoid user requests that may elicit multiple valid responses with different semantic meanings, as these are not easily measurable with automatic metrics.

Knowledge The knowledge should maintain the same properties as the original, providing correct factual information (directly relevant to user request) in 1-2 sentences with a maximum of 30 words. We found it practical to structure the knowledge with two pieces of information: one distractive and the other relevant to the last user turn. The distractive element should be easy to ignore in the response if the model adequately attends to the conversation history.

Response The response to the last user turn should be the only unique and valid response, based on the information contained in the provided knowledge. Additionally, in line with FaithDial guidelines, the response should be faithful to this knowledge, often comprising a large portion that is either a direct copy or a paraphrase of it. However, we instructed our annotators to perform only minimal paraphrasing necessary to ensure a well-structured response. We do so to save annotation time, as knowledge rephrasing isn't the objective of CHARP. Additionally, this helps avoid evaluation mismatches caused by incidental inconsistencies between the annotation style of our annotators and that of the FaithDial crowd-workers.

B Experimental Setting

B.1 FaithDail CRITIC

FaithDial (Dziri et al., 2022a) CRITIC was trained using a dataset comprising 14,000 hallucinatory turns (edited original WoW turns) and 20,000 faithful turns (unedited WoW and FaithDial turns), serving as negative and positive examples, respectively. More precisely, the authors paired up turns with their respective knowledge snippet and trained the RoBERTa-large model (Liu et al., 2019) by framing the task as sequence pair binary classification. The FaithDial hallucination detector not only demonstrates a high correlation with human judgment but

Conversation History:
Seeker: My sister is a baker. She bakes mostly cakes and pies and other pastries.
Wizard: A baker is someone who bakes.
Seeker: Yes, like my sister. Haha! What else do you know about baking?
Wizard: It has to do with breads and other products made using an oven.
Seeker: Definitely. My sister can make some of the best pastries in the area!

Knowledge:
Since grains have been a staple food for millennia, the activity of baking is a very old one.

Ground Truth Response:
That's great! Baking has been around as long as grains have been used as a staple food, for thousands of years.

Conversation History:
Seeker: My sister is a baker for the Ladurée bakery in France. She bakes mostly cakes and pies and other pastries.
Wizard: A baker is someone who bakes.
Seeker: Yes, like my sister. Haha! What else do you know about baking?
Wizard: It has to do with breads and other products made using an oven.
Seeker: Sometimes I wonder if my sister could work in a different country with great opportunities for bakers.

Knowledge:
In addition to France, where there are 34,000 bakeries, bread is a significant part of German cuisine with about 10,000 bakeries.

Ground Truth Response:
She could try moving to Germany since bread is a significant part of their cuisine and there is around 10,000 bakeries.

Figure 3: Original example from the FaithDial validation set (left) and our edited hCHARP version (right). Green text indicates content that the model is expected to reason over, while red text marks distracting content within the provided knowledge.

also excels in hallucination detection testbeds like BEGIN (Dziri et al., 2022c). Furthermore, it outperforms classifiers trained on counterpart hallucination detection datasets, such as DECODE (Welleck et al., 2019) and DNLI (Nie et al., 2021).

B.2 Llama Finetuning

`Llama-2-7B` We utilize the 7B chat version of the Llama-2 model series (Touvron et al., 2023), which is the largest model we could effectively fine-tune (compared to the 13B and 70B versions), given our computational resources. We fine-tuned the model on a single node equipped with 8 NVIDIA V100 GPUs with 32GB of memory, utilizing a codebase built on the PyTorch version (Paszke et al., 2019) of the Transformers library (Wolf et al., 2020). The initial learning rate was set to 2e-6, employing the AdamW optimizer (Kingma and Ba, 2014) with a cosine decay learning rate schedule. The model was trained over 5 epochs with a maximum sequence length of 1024 tokens. We set the per-GPU batch size to 48, the maximum size that we can fit on a single GPU. Training acceleration was achieved by leveraging the deepspeed library (Rasley et al., 2020), mixed precision training (Micikevicius et al., 2018), and gradient checkpointing (Chen et al., 2016). We pick up the best checkpoint using early stopping based on performance on FaithDial validation set.

B.3 Few-shot Experiments

In addition to `Llama-2-7B`, we also conduct out-of-the-box inference (without fine-tuning) on experiments using gpt-3.5-turbo (OpenAI, 2022) and Mixtral-8x7B (Jiang et al., 2024). Throughout this paper, we refer to these models as `ChatGPT` and `Mixtral`, respectively. To this end, we carefully design a prompt that takes the conversation history and the knowledge relevant to the last seeker's turn as input to generate a bot response:

```
You are given a chitchat conversation between a "User" and a "Bot". Your goal is to generate a response to the last user turn, which in turn should be based on the given "Knowledge". You are prohibited from generating any extra information that is not mentioned in the given knowledge. The output should be a JSON dictionary as follow: {"response": ""}. Here are a few demonstration examples:

[In_CONTEXT_EXAMPLE_1]

[In_CONTEXT_EXAMPLE_2]

[In_CONTEXT_EXAMPLE_3]

[INPUT_EXAMPLE]
```

We designed the instruction part of the prompt through trial and error iterations until we verified that all models could follow the instructions and generate a response that addresses the user query in our required format. Then, we continuously added in-context examples until the output of all models stabilized (with minor to no changes in the model response). We set the number of in-context examples, that were picked up from FaithDial training set, to 3 as we didn't see any improvement in adding more examples or performing more prompt engineering. On one hand, we execute the generation process of ChatGPT and Mixtral samples through the commercial APIs of OpenAI [6] and Replicate [7], respectively. On the other hand, we use our local V100 GPUs to infer with Llama-2-7B. However, across all experiments, we set the temperature to 1.0, the frequency penalty to zero, and top-p to 1.0, aiming to minimize randomness during the generation process.

C Analysis

C.1 Automatic Evaluation Results

Tables 7, 8, and 9 show the automatic metric scores of the models fully tuned on FaithDial and under the *3-shot* setting on the FaithDial validation subset, eCHARP, and hCHARP, respectively. First, we observe that the finetuned Llama-2-7B, across all three evaluation sets, systematically yields slightly worse results on all FaithDial metrics compared to GODEL-base and FLAN-base. We believe this is primarily because, despite full parameters tuning on FaithDial, Llama-2-7B has retained some of its *chatty behavior* that was induced during the SFT and RLFH training procedures. However, this does not mean that the outputs of Llama-2-7B are of lower quality than those of GODEL-base or FLAN-base; in fact its the opposite as indicated by the human evaluation results in Tables 5 and 6. This particular observation aligns with the findings of other studies (Sankar et al., 2019; Yeh et al., 2021; Parthasarathi et al., 2020) regarding the limitations of automatic metrics in evaluating dialog systems.

Models	BLEU↑ (y, y')	Critic↓ (k, y')	BERTScore↑ (y, y')	BERTScore↑ (k, y')
Finetuning				
FLAN-base	14.6	0.4	71.1	81.2
GODEL-base	14.5	0.3	70.8	81.5
Llama-2-7B	12.0	2.0	69.2	73.1
3-shot				
Llama-2-7B	3.7	72.9	54.3	59.5
Mixtral	9.4	29.1	65.9	74.3
ChatGPT	6.5	55.2	62.3	67.6

Table 7: Performance of models on FaithDial validation set used to build CHARP. full fine-tuning on FaithDial, and with no fine-tuning by using 3 *in-context* examples. All scores are scaled within the range of $[0, 100]$.

[6] https://chat.openai.com/

[7] https://replicate.com/

Models	BLEU↑ (y, y')	Critic↓ (k, y')	BERTScore↑ (y, y')	BERTScore↑ (k, y')
Finetuning				
FLAN-base	22.0	1.9	70.1	78.6
GODEL-base	18.7	0.6	67.6	81.7
Llama-2-7B	17.2	3.7	68.1	69.1
3-shot				
Llama-2-7B	8.0	54.0	63.7	65.0
Mixtral	20.6	16.3	74.6	70.0
ChatGPT	20.2	22.8	74.6	69.8

Table 8: Performance of models on hCHARP. All scores are scaled within the range of $[0, 100]$.

Models	BLEU↑ (y, y')	Critic↓ (k, y')	BERTScore↑ (y, y')	BERTScore↑ (k, y')
Finetuning				
FLAN-base	22.8	1.7	70.8	79.4
GODEL-base	20.5	0.5	69.4	82.5
Llama-2-7B	20.1	3.9	70.0	68.8
3-shot				
Llama-2-7B	8.9	51.6	65.4	65.3
Mixtral	21.3	16.6	75.5	70.0
ChatGPT	19.9	21.8	74.8	69.2

Table 9: Performance of models on eCHARP. All scores are scaled within the range of $[0, 100]$.

The results are much worse when comparing the *3-shot* models with the fine-tuned ones across all metrics and evaluation sets. The high hallucination ratio, as indicated by the CRITIC score, is well-justified since these models (especially Llama-2-7B) tend to incorporate out-of-knowledge information, a finding that is corroborated by human evaluation. However, our human evaluators noted that the responses from Mixtral and ChatGPT tend to be creative, often using different words than the provided knowledge. Despite this, they deliver responses that are semantically aligned with the given knowledge and have the same semantic meaning as the ground truth response. This tendency results in a misleadingly high hallucination ratio, suggesting that the FaithDial CRITIC model [8] is overly sensitive to lexical overlapping and fails to capture the underlying semantic meaning. This is also noticeable when considering that CRITIC score increases more significantly than the drops in the BERTScore (k, y'). For instance, while the CRITIC score increases by 70.9%, the BERTScore(k, y')

[8] which was specifically-tuned on FaithDial examples, while BERTScore models, in contrast, were tuned on MNLI (Williams et al., 2018).

decreases by only 14.6% when comparing the tuned Llama-2-7B with its *3-shot* counterpart on FaithDial validation subset. Still, FaithDial automatic metrics significantly under-estimate the performance of Mixtral and ChatGPT compared to fine-tuned GODEL-base and FLAN-base. However, it's interesting to note that the ranking of *3-shot* models (Mixtral > ChatGPT > Llama-2-7B) according to automatic metrics aligns with the ranking obtained through human evaluation.

C.2 Inter-annotator Agreement

In an effort to assess the quality of human evaluations, we tasked our annotators to evaluate a subset of 64 randomly selected examples from both eCHARP and hCHARP. This evaluation covered all six model variants studied in our experiments, leading to 786 model outputs that were evaluated by 3 annotators. Table 11 shows the inter-annotator agreement scores, as measured by the Kappa Coefficient (Carletta, 1996).

	eCHARP	hCHARP
Finetuning Models		
FLAN-base	93%	93%
GODEL-base	93%	96%
Llama-2-7B	94%	93%
3-shot Models		
Llama-2-7B	88%	92%
Mixtral	97%	94%
ChatGPT	92%	89%

Table 10: Kappa inter-annotator agreement scores reflect human judgments of three FaithDial-tuned models and three *3-shot* models, based on a random set of 64 examples each from eCHARP and hCHARP.

Overall, we observe significantly high agreement among annotators, well above the widely accepted threshold of 80%. Despite slight variations, we notice that the kappa score remains above this threshold across all the configurations. This not only demonstrates the professionalism of our annotators but also the clarity and precision of our proposed evaluation schema.

C.3 GPT-4 Evaluation

Given the complete conversational history, knowledge, response, and a system's prediction, we constructed a prompt requiring GPT-4 to perform the same checklist evaluation procedure as outlined in §5.2:

```
Your task is to assess the quality of a machine learning system's response in a conversation. The conversation 'history' includes interactions between a user (Seeker) and a bot (Wizard), along with relevant 'knowledge' that pertains to the Seeker's last utterance. You are also provided with a 'response' (a ground truth or expected response) and the 'prediction' (the system's predicted response). Your evaluation involves comparing the system's 'prediction' with the 'response', considering the entire conversation 'history' and the provided 'knowledge'.
The evaluation is structured around categorizing the system's response into specific categories. Your output should be a JSON dictionary with a single <key, value> pair. The key is "category", and the value is a list of the category numbers that the predicted response falls under. For example: "category": [1]. These categories are:
1. The system's prediction is of high quality and is an equivalent or a paraphrase of the ground truth response.
        [In_CONTEXT_EXAMPLE_1_FOR_CAT_1]
        [In_CONTEXT_EXAMPLE_2_FOR_CAT_1]
2. The system's prediction aligns with the ground truth response, but adds extra information, meaning that the content in the prediction is absent from the ground truth response, the given knowledge or the given history.
        [In_CONTEXT_EXAMPLE_1_FOR_CAT_2]
        [In_CONTEXT_EXAMPLE_2_FOR_CAT_2]
3. The system's prediction is an identical copy or a very similar rephrasing of the entire knowledge. Meaning that its content is the exact same that can be found in the knowledge. It might contain content that aligns with the ground truth response, but it also contains off-topic content from the knowledge. It should not contain information that is absent from the given knowledge.
        [In_CONTEXT_EXAMPLE_1_FOR_CAT_3]
        [In_CONTEXT_EXAMPLE_2_FOR_CAT_3]
4. The system's prediction states doubt and ignorance, saying it doesn't know, doesn't understand, or is not equiped to answer; yet it copies or rephrases the content from the knowledge. The system's prediction content may originate from the whole knowledge or from the part of the knowledge that correctly aligns with the ground truth response or from the part of the knowledge that is not aligned with the ground truth response. It should not contain information that is absent from the given knowledge. For example:
        [In_CONTEXT_EXAMPLE_1_FOR_CAT_4]
        [In_CONTEXT_EXAMPLE_2_FOR_CAT_4]
5. The system's prediction is an identical copy or a very similar rephrasing of the part of the knowledge whose content is off-topic and does not align with the ground truth response. The content of the system prediction should not contain any content that aligns (even partially) with the ground truth response. It should not contain information that is absent from the given knowledge or state ignorance by saying it doesn't know and is not able to get that information.
        [In_CONTEXT_EXAMPLE_1_FOR_CAT_5]
        [In_CONTEXT_EXAMPLE_2_FOR_CAT_5]
```

```
6. The system's prediction does not align with the ground truth response and its content is made of mixed-up information coming from both the knowledge part that aligns with the ground truth response (on-topic) and the part that does not align with the ground truth response (off-topic). Both parts are not just a copy of the knowledge, but are merged together, which leads to wrong and inaccurate information in the prediction.
    [In_CONTEXT_EXAMPLE_1_FOR_CAT_6]
    [In_CONTEXT_EXAMPLE_2_FOR_CAT_6]
7. The system's prediction does not align with the ground truth response yet it cannot be classified as any of the previously mentioned categories. This includes but is not limited to: having extra information that is absent from the given knowledge, having two or more content elements that contradict each other, being empty.
    [In_CONTEXT_EXAMPLE_1_FOR_CAT_7]
    [In_CONTEXT_EXAMPLE_2_FOR_CAT_7]
Now you must evaluate the following:
[INPUT_EXAMPLE]
```

We found that using two *in-context* examples for each category works better than using just one, with no further improvement observed by using additional examples. Due to the high costs associated with calling GPT-4, we opted for its more cost-effective version, GPT4-turbo, to perform evaluations on the full evaluation sets.

	eCHARP	hCHARP
Finetuning Models		
FLAN-base	84%	84%
GODEL-base	85%	87%
Llama-2-7B	88%	89%
3-shot Models		
Llama-2-7B	87%	86%
Mixtral	92%	90%
ChatGPT	90%	91%

Table 11: Kappa agreement scores between human judgments and those of GPT-4-turbo regarding the quality of outputs from three FaithDial-tuned models and three *3-shot* models on the full set of eCHARP and hCHARP.

Table 11 presents the kappa agreement scores between human judgment and GPT4-turbo across six models, as measured on the complete datasets of eCHARP and hCHARP. We notice that, overall, the agreement scores are consistently high (>0.8) and exhibit minimal variation across different models and evaluation sets. It is interesting to note that the agreement is consistently higher for well-performing models (Mixtral and ChatGPT), while the evaluation conducted by GPT-4 becomes more challenging when judging the output of poorly performing models.

	GPT4-turbo		Human	
	eCHARP	hCHARP	eCHARP	hCHARP
Finetuning Models				
FLAN-base	91%	88%	90%	88%
GODEL-base	88%	88%	88%	89%
Llama-2-7B	89%	92%	91%	90%
3-shot Models				
Llama-2-7B	90%	89%	90%	90%
Mixtral	93%	93%	95%	94%
ChatGPT	93%	94%	95%	95%

Table 12: Kappa agreement scores between GPT-4 and GPT4-turbo judgments (first two columns), and between GPT-4 and human judgments (last two columns). This experiment was conducted on a randomly selected subset of 110 examples from both eCHARP and hCHARP, comprising 6 models.

To ensure the quality of our evaluation, we measured the discrepancy between GPT-4 and GPT4-turbo judgments by comparing their outputs on a random sample of 110 examples, which constitutes approximately 10% of the total data in CHARP. Table 12 shows the agreement scores of GPT-4, not only with GPT4-turbo (first two columns) but also with human judgments (last 2 columns). This comparison is based on selected 110 subset examples from eCHARP and hCHARP. On one hand, we observe a relatively high agreement between GPT-4 and GPT4-turbo, ranging from 0.8 at worst to 0.94 at best. Notably, most disagreements occur with the less performing models (GODEL-base and FLAN-base), which are in line with the observations made in Table 11. Although not directly comparable[9], we notice that GPT-4 judgments are systematically closer to human ones compared to those of GPT4-turbo across different settings. For instance, the agreement between GPT-4 and human judgments on ChatGPT hCHARP is higher by 0.4-0.5 than that of GPT4-turbo with humans (0.91). Despite this, we believe that GPT4-turbo presents an acceptable quality-cost trade-off, being three times less expensive than GPT-4. By all means of comparison, it offers a comprehensively superior alternative to FaithDial's automatic evaluation metrics.

[9]We measured the kappa agreement between GPT4-turbo and human judgments on the randomly selected example subsets and found that the agreement strongly aligns with that observed in the full evaluation sets, with a maximum variance of ±0.1 and ±0.2 in rare cases.

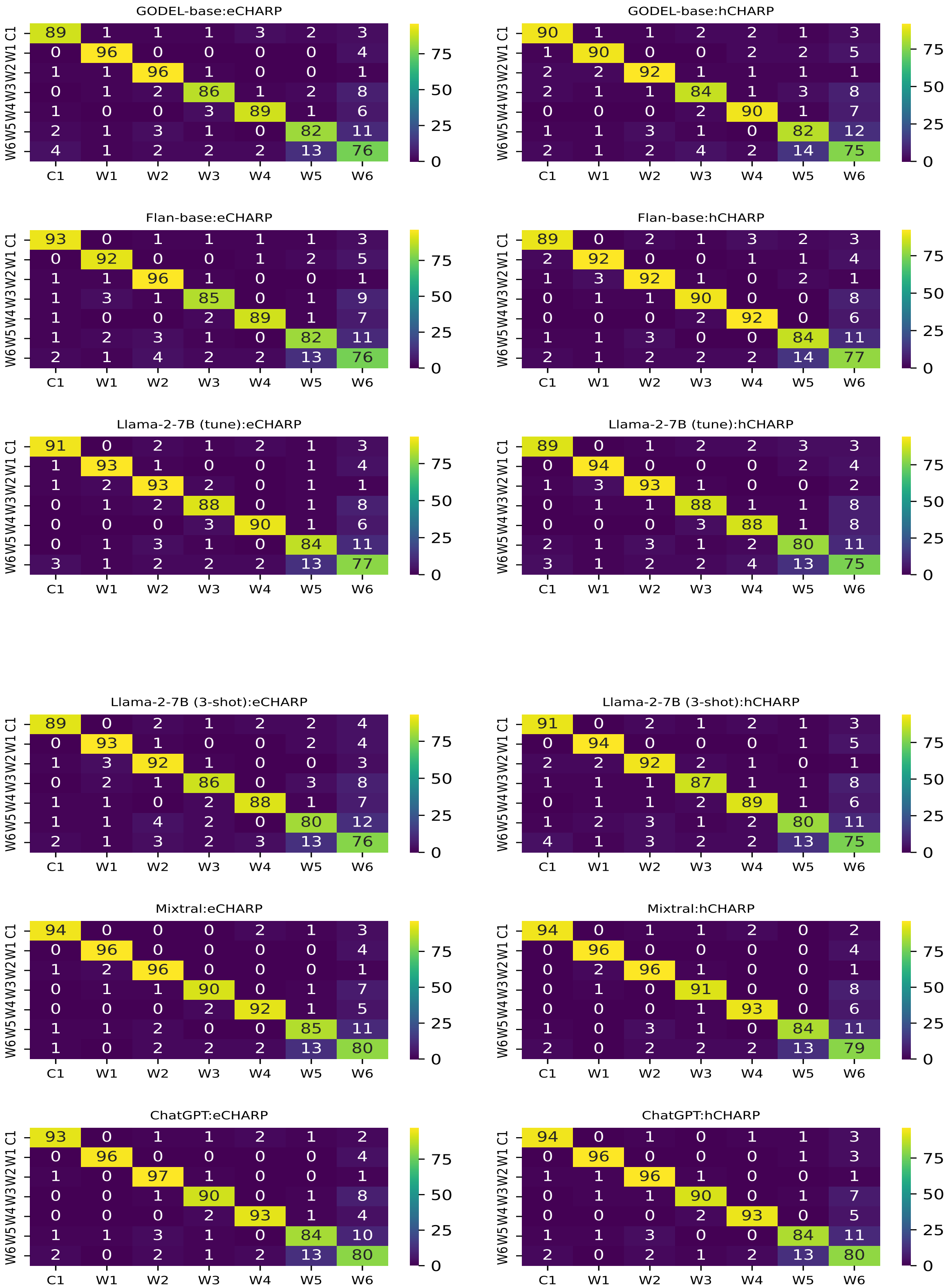

Figure 4: Heatmap showing the normalized (percentage) contingency tables of evaluation categories between GPT4-turbo (rows) and human (columns) judgments. It was measured on the output of 6 models for both eCHARP (on the left) and hCHARP (on the right).

Conversation History:

Seeker: I really love cupcakes, what's your favorite flavor?
Wizard: Well, as a bot, I can't eat cupcakes. Which are also called fairy cakes, buns and patty cakes in other cultures. What's your favorite?
Seeker: I love any kinda of cake to be honest. What kind of things do people top them with?
Wizard: Icing is used, of course, but also fruit and candy can be applied. Do you have a preference?
Seeker: Hmm, not really. What is the most popular topping?

Knowledge:
The most popular topping for pizza is pepperoni. On the other hand, cupcakes are usually decorated with chocolate icing and sprinkles.

Ground Truth Response:
Oh, the most popular topping for cupcakes is chocolate icing and sprinkles.

Models Prediction:

FLAN-base:
The most popular topping for pizza is pepperoni. On the other hand, cupcakes are usually decorated with chocolate icing and sprinkles. $\backslash\mathcal{W}2$
GODEL-base:
The most popular topping for pizza is pepperoni. $\backslash\mathcal{W}4$
Llama-2-7B (finetune):
I couldn't say, but I know that pizza's most popular topping is pepperoni. $\backslash\mathcal{W}3$
Llama-2-7B (3-shot):
Some popular toppings for cupcakes are chocolate icing and sprinkles. $\backslash\mathcal{C}1$
Mixtral:
Chocolate icing and sprinkles are commonly used as toppings for cupcakes. $\backslash\mathcal{C}1$
ChatGPT:
The most popular topping for cupcakes is chocolate icing and sprinkles. $\backslash\mathcal{C}1$

Conversation History:

Seeker: I really like the color red, but it's getting old in my kitchen now.
Wizard: Is that so? I know some science of red; it's at the end of the spectrum of visible light, next to orange.
Seeker: It seems to be a very powerful color, but it's not very calming. I like Salmon too. I remember chewing the crayon as a kid.
Wizard: That's funny. The actual color varies from almost white to light orange, depending on their levels of the carotenoid astaxanthin.
Seeker: I might just paint my new car in that color.

Knowledge:
Michelangelo's dedication to his craft allowed him to paint the Sistine Chapel, a masterpiece that took four good years to complete. The versatility and elegance of salmon enhances various items.

Ground Truth Response:
That would be nice. Salmon adds a versatile and elegant touch to various items.

Models Prediction:

FLAN-base:
Interesting. Did you know Michelangelo painted the Sistine Chapel? $\backslash\mathcal{W}2$
GODEL-base:
Nice. Did you know Michelangelo took four years to paint the Sistine Chapel? $\backslash\mathcal{W}2$
Llama-2-7B (finetune):
Wow! Speaking of painting, did you know Michelangelo took four years to complete the Sistine Chapel? $\backslash\mathcal{W}2$
Llama-2-7B (3-shot): $\backslash\mathcal{W}4$
Salmon is a versatile color that can enhance various items, much like Michelangelo's masterpieces in the Sistine Chapel.
Mixtral:
Salmon is a versatile color that can enhance various items, much like Michelangelo's painting of the Sistine Chapel showcased his dedication and mastery. $\backslash\mathcal{W}4$
ChatGPT:
Salmon is a great choice for your new car color. Did you know that Michelangelo took four years to complete the painting of the Sistine Chapel? $\backslash\mathcal{C}1$

Figure 5: Tow examples from hCHARP (left side), along with the predictions of the six models employed in our study (right side). For each model response, we show the FaithDial judgment (hallucination indicated by , and no hallucination by), along with the category of human judgment. In the second example, `ChatGPT`'s response (rare but interesting) is deemed correct by human evaluators because it accurately addresses the user's comment before introducing an unrelated piece of knowledge in a manner that opens a new topic, Although it aligns with FaithDial guidelines, but the CRITIC judge this case as hallucination, mainly because *painting new car* is not mentioned in the provided knowledge.