Dial BEINFO *for Faithfulness*: Improving Factuality of Information-Seeking Dialogue via Behavioural Fine-Tuning

Anonymous ACL submission

Abstract

Factual faithfulness is a crucial requirement in information-seeking dialogue: the system should respond to the user queries so that the responses are meaningful and aligned with the knowledge provided to the system. However, most modern large language models (LLMs) suffer from hallucinations, that is, they generate responses not supported by or even contradicting the knowledge source. To mitigate the issue and increase faithfulness of information-seeking dialogue systems supported by the LLMs, we introduce BEINFO, a simple yet effective method that applies 'behavioural tuning' on the LLMs to aid information-seeking dialogue. Relying on three standard information seeking dialogue datasets, we show that models tuned with BEINFO become considerably more faithful to the knowledge source both for datasets and domains seen during BEINFO-tuning, as well as on unseen domains, when applied in a zero-shot manner. In addition, we present a 'real-life' case study on conversations with real users, showcasing that the models with 3B parameters (e.g., Flan-T5) tuned with BEINFO demonstrate strong performance on data from real 'production' conversations: when tuned on a limited amount of such realistic in-domain dialogues, they surpass much larger LLMs used 'off-the-shelf', both on automatic and human evaluation metrics.

1 Introduction

Pretrained large language models (LLMs), being able to generate natural and grammatical text and respond coherently to user queries, are the mainstay of modern NLP (Naveed et al., 2023). They have demonstrated their capabilities in a plethora of tasks where the general world knowledge, which can be learnt via pretraining directly from the data, is required (Touvron et al., 2023; Hoffmann et al., 2022). However, reliance only on the content from the pretraining data also means that the model's

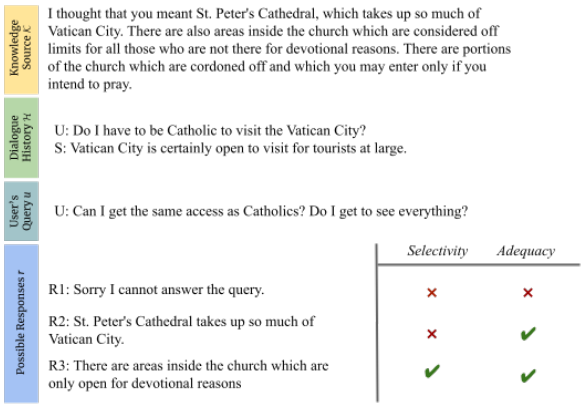

Figure 1: An example of an information-seeking dialogue based on the DoQA dataset (Campos et al., 2020). Potential responses R1, R2, R3 at the bottom illustrate different issues with two crucial aspects of factual faithfulness: selectivity and response adequacy.

responses might be generic or not be up to date, especially for queries responses to which change across time such as *Who is the current prime minister of the United Kingdom?* An even more prominent issue is *hallucination* (Zhang et al., 2023), a phenomenon often observed even with the most powerful LLMs: the models are prone to output incoherent, irrelevant and/or even factually incorrect or unsupported statements (Naveed et al., 2023).

A widely used method to ground and control the content of the output of an LLM is *retrieval-augmented generation* (RAG; Lewis et al., 2020), where the input to the model is complemented with a retrieved external knowledge source relevant to the user's query. However, even with the use of RAG, the model's output can be unpredictable and not fully controllable: they still sometimes do not adhere to the knowledge source and hallucinate (Shuster et al., 2021), which can decrease their applicability in user-facing scenarios, as well as raise concerns of their safety (Daheim et al., 2023).

The problem of adherence to the knowledge sources is especially important in the context of

information-seeking dialogue (Saeidi et al., 2018). The core of this task is to maintain a conversation with the user and respond to their queries based on the provided knowledge source. Figure 1 presents an example of information-seeking dialogue between the user and the system and potential responses of the system. Orthogonally to improving retrieval systems themselves (Wang et al., 2023a; Mo et al., 2023), prior work has attempted to combat hallucinations with task arithmetic (Daheim et al., 2023), conditioning generation on special control tokens (Rashkin et al., 2021), and by incorporating a token-level critic which judges the faithfulness of the generated response (Dziri et al., 2021). However, the proposed approaches requires either training an additional model or using complex inference processes such as context-aware decoding (Shi et al., 2023).

In this work, we propose **BeInfo**, a simple yet effective method that applies '**be**havioural fine-tuning' of LLMs to increase faithfulness of the generated responses for **info**rmation-seeking dialogue supported by the LLMs. The model is tuned on a reasonably sized collection of publicly available dialogue data with the true knowledge source(s) extended with randomly sampled facts from a large knowledge base. Intuitively, this should teach the model to become more selective in the information it uses to generate the response and 'prepare' its expected behaviour (hence the term 'behavioural tuning') for the intended task of knowledge-grounded dialogue. The tuned model can either be used 'as is' or as a starting point to fine-tune it further to a specific domain.

First, we assess the effectiveness of BeInfo on three standard datasets for information-seeking dialogue: FaithDial, TopiOCQA and DoQA. Our results demonstrate that BeInfo leads to consistent improvements in factual faithfulness across several standard evaluation metrics, also with on par or larger lexical overlap between the generated and golden responses. The improvements are especially pronounced when models tuned with BeInfo are applied in a zero-shot manner to unseen datasets and domains, indicating the usefulness of behavioural tuning for the task. We then present a case study focused on conversations with real users: the main result demonstrates that combining BeInfo with a small number of in-domain dialogues can substantially increase dialogue factuality even in specialized dialogue domains. The code for BeInfo is available online at: [URL-ANONYMOUS].

2 Methodology

Task Definition. The aim of information-seeking dialogue is to provide the user with information they need based on one or more knowledge sources, which are typically retrieved from a large knowledge base. More formally, given the knowledge source $\mathcal{K}$, the dialogue history $\mathcal{H}$ and the user's query u, the system should output the response r which is factually faithful to $\mathcal{K}$. Here, we follow Rashkin et al. (2021) and Dziri et al. (2022a)'s direct definition of *faithfulness*: the response should not contain any information which either contradicts $\mathcal{K}$ or is not supported by $\mathcal{K}$.

Behavioural Tuning for Faithfulness. An effective model for faithful information-seeking dialogue needs to perform two actions correctly: **1)** select the correct part of information provided in $\mathcal{K}$ (termed *selectivity*) to base the generated response on and **2)** provide the response, with the requirement to (i) inform the user when $\mathcal{K}$ contains no information relevant to u, or (ii) ask for clarification (termed *(response) adequacy*);[1] see Figure 1 again. BeInfo aims to improve on both desiderata via behavioural fine-tuning (Ruder, 2021) of any instruction-tuned LLM.

To instill the capability for information-seeking dialogue into the model, we perform behavioural tuning on the combination of (i) conversational QA and (ii) information-seeking dialogue datasets. In both tasks, the response has to be generated based on some knowledge source $\mathcal{K}$, making them suitable for faithful response generation. Further, beyond tuning on related tasks, we propose to augment the datasets to steer the model towards the selectivity and adequacy behaviour, as follows.

For selectivity, ground truth $\mathcal{K}$ provided in the dataset is extended with additional knowledge sources $\mathcal{K}'$ which are irrelevant to user query u, serving as negative examples or distractors. Intuitively, distractors mimic the presence of information irrelevant to u in $\mathcal{K}'$, this way promoting the model's selectivity. We augment ground truth knowledge source $\mathcal{K}$ with n distractors; they are randomly sampled from the knowledge base of the

[1] Put simply, in our setup response adequacy discerns between **1)** the case when the model does have the correct information in the knowledge source and should provide it versus **2)** the case when the model is certain that it cannot provide a correct answer to the user query or it does not even understand the query and requires further clarification to be able to react in the next turn. There might be other, finer-grained options of response adequacy beyond the two simple cases investigated here, but we leave those investigations to future research.

corresponding dataset[2].

For response adequacy, we augment the fine-tuning datasets with dialogues without any relevant $\mathcal{K}$ provided, making them unanswerable for the system. To construct such dialogs, for a dialogue history $\mathcal{H}$ and a corresponding user query u we randomly sample unrelated knowledge sources $\mathcal{K}'$. During fine-tuning, the response r is substituted with a special response signifying that the combination of $\mathcal{H}$ and u cannot be answered based on provided $\mathcal{K}'$. In our experiments, we augment the original dataset with 10% unanswerable dialogues.

Further Task-Specific Fine-Tuning. The output of the 'general' behavioural fine-tuning step is a 'behaviour-specialised' LLM for factually faithful information seeking dialogue. It can be used directly 'as is', or as a starting point for further task-specific tuning, as illustrated in Figure 2.

3 Experimental Setup

Training Setup. In order to leverage inductive biases of instruction-tuned models, the input for BEINFO includes the following: (i) instructions to respond as factually accurately as possible, (ii) augmented knowledge source which includes: ground truth $\mathcal{K}$ and $n = 4$ distractors $\mathcal{K}'$ for 'answerable' dialogues, and 5 randomly sampled $\mathcal{K}'$-s for unanswerable dialogues and (iii) dialogue history which combines all the previous turns (the set $\mathcal{H}$) and the current user query u. An example input and instruction text are shown in Appendix A. The models are then trained in a standard sequence-to-sequence fashion with cross-entropy loss. The output is either ground truth responses for answerable dialogues, where knowledge source $\mathcal{K}$ contains the information to address user's query, or a predefined response *'Could you please clarify or rephrase the query?'* if the dialogue is unanswerable. Training the models using BEINFO proceeds at turn level: dialogue history at every turn is used as input.

Datasets. To perform behavioural fine-tuning, we use a standard dataset for information seeking dialogue, FaithDial (Dziri et al., 2022a), and an established conversational QA dataset, TopiOCQA (Adlakha et al., 2022). Generalisation capabilities of the models after the BEINFO tuning are evaluated on another domain and dataset (i.e., this could be seen as 'zero-shot' from the domain adaptation perspective). For this, unless explicitly stated otherwise, we rely on a multi-domain conversational QA dataset, DoQA (Campos et al., 2020). The key statistics of the datasets are in Table 1, with further details and data analyses in Appendix B.

	FaithDial	TopiOCQA	DoQA
Domains	Open Wikipedia-based	Open Wikipedia-based	3 (Cooking, Travel, Music)
# dialogues	4,094 / 764 / 791	3,509 / 205 / 206	1,037 / 200 / 1,200
# turns	36,809 / 6,851 / 7,101	45,450 / 2,514 / 2,502	4,612 / 911 / 5,394
Avg. turns	9	13	4.48
Avg. length of questions	17.25	6.92	12.99
Avg. length of responses	20.29	11.38	10.43

Table 1: Overall statistics of the used dialogue datasets. The number of conversations and turns are provided for train / dev / test splits of the datasets.

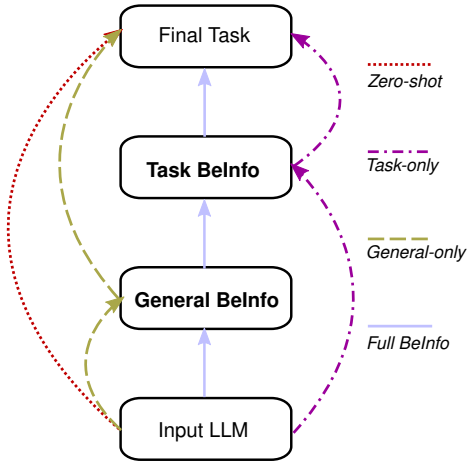

Figure 2: An overview of different fine-tuning and inference setups for LLMs with and without BEINFO (§3).

Models. Prior work (Dziri et al., 2022a) has demonstrated that instruction-tuned models such as the Flan series (Chung et al., 2022) are a very strong baseline for factuality in information-seeking dialogue. Thus, we use them as a base for the proposed method.[3] In the experiments, we use Flan-T5 (Chung et al., 2022) (BASE, LARGE and XL) and Tk-Instruct-3B (Wang et al., 2022). All the backbone models were pretrained on a large number of tasks with instructions, which yields faster specialisation of the models to information-seeking dialogue, especially when, as in our setup, the input/prompt includes a short description of the task.

Fine-Tuning and Inference Setups. The LLMs can be used directly in the final task in a fully *zero-shot* manner or via in-context learning as 'black boxes': this is a typical usage of very large models in dialogue tasks. We can also conduct BEINFO

[2]We use random sampling due to its simplicity and leave the strategies based on semantic similarity as future work.

[3]We again note that BEINFO can be applied on top of any generative model.

tuning of 'smaller LLMs' via different regimes: (i) fine-tuning directly on the task data but with augmented knowledge sources (if available) (i.e., *task-only* BEINFO); (ii) fine-tuning only on the available data from other dialogue datasets and porting the tuned model to the task in a zero-shot fashion (i.e., *general-only* BEINFO- an example is tuning on FaithDial and TopiOCQA and using the model for DoQA, or vice versa); (iii) finally, we can run a stage of general BEINFO followed by in-task BEINFO (termed *full* BEINFO). An overview of the different setups is provided in Figure 2.

Evaluation Metrics. We rely on automated metrics to measure *lexical similarity* of the generated responses and ground truth responses: BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). To measure *semantic similarity* between generated and gold responses, we use BERTScore (Zhang et al., 2019).[4] To evaluate *faithfulness*, we use BERTScore and token-level precision between the generated response and the knowledge source $\mathcal{K}$. We denote BERTScore between ground truth and generated responses as "BERTS" and one between the knowledge source $\mathcal{K}$ and generated responses as "$\mathcal{K}$-BERTS". In both cases we use BERTScore-F1. Token-level precision between the generated response and knowledge source $\mathcal{K}$ ($\mathcal{K}$-Precision; Adlakha et al., 2023) measures the proportion of tokens in generated response which occur in $\mathcal{K}$. Prior work (Adlakha et al., 2023) demonstrates that $\mathcal{K}$-Precision has the highest correlation with human (as well as GPT-4-elicited) faithfulness judgements among different automated metrics.

Hyperparameters and Training Details. BEINFO was implemented using HuggingFace Transformers library (Wolf et al., 2020). The models were trained with AdamW (Loshchilov and Hutter, 2019). With BEINFO, we tune for 5 epochs, the learning rate is 5e-5; when tuning the model to a specific dataset, we run it for 10 epochs with the learning rate of 5e-6. We use the warm-up rate of 0.1 and linear decay, with the default weight decay rate of 0.01. Beam search is run with the beam size of 10.

4 Results and Discussion

Faithfulness on Unseen Data. One of the main aims of behavioural fine-tuning with BEINFO is to increase the factual faithfulness of responses in zero-shot domain transfer, on unseen data in any domain. Therefore, we start by presenting the results of the variant tuned with BEINFO on FaithDial plus TopiOCQA, where inference is run on the dataset unseen during BEINFO tuning: DoQA (i.e., general BEINFO from Figure 2). The results are presented in Table 2. They confirm that BEINFO substantially improves faithfulness while either improving or only minimally affecting the similarity between generated responses and the gold response. Importantly, the improvements hold across different model sizes: Flan-T5 BASE, LARGE and XL with 250M, 780M and 3B parameters, respectively.

Model	BLEU	ROUGE	BERTS	$\mathcal{K}$-BERTS	$\mathcal{K}$-Precision
Flan-T5$_{BASE}$	22.89	34.46	61.60	67.75	90
+BEINFO	22.76	34.04	61.71	77.55	100
Flan-T5$_{LARGE}$	26.16	39.57	64.61	71.38	93.86
+BEINFO	26.34	38.55	63.19	75.55	100
Flan-T5$_{XL}$	28.66	41.99	65.89	67.21	94.12
+BEINFO	26.65	39.39	64.60	80.19	100

Table 2: Results on DoQA without any in-task BEINFO tuning. The models are tuned on a combination of FaithDial and TopiOCQA. The results are averaged across three domains in DoQA – *Cooking, Travel and Movies*. Full results are presented in Appendix C.

Using a Smaller Dataset for BEINFO Tuning. The previous results from Table 2 show BEINFO's effectiveness when tuned on two reasonably sized datasets, FaithDial with 36,809 turns, and TopiOCQA with 45,450 turns. Now, we test the opposite direction: fine-tuning BEINFO on a smaller-scale dataset like DoQA (4,612 turns) and evaluating zero-shot on FaithDial. Besides further testing the versatility of the approach, we also probe sample efficiency of the approach and its adaptability to smaller datasets and computational budgets.

Results in Table 3 suggest that tuning the models with BEINFO even on smaller datasets without any subsequent in-task tuning consistently improves the factuality of generated responses. Especially large gains were observed for larger models, both for faithfulness and semantic similarity between the generated responses and the ground truth, indicating the potential for sample efficiency of BEINFO. Similar trends were observed when evaluating on TopOCQA instead of FaithDial; see Appendix D.

Different Instruction-Tuned Models. Previous results have already verified that BEINFO can be applied to Flan models of different sizes, and we now evaluate its impact on another instruction-based model: Tk-Instruct-3B. We fine-tune the models again on FaithDial and TopiOCQA and evaluate their performance on DoQA's *Travel* domain test

[4] Similarly to Daheim et al. (2023), we use *deberta-large-mnli* as an underlying model for computing the score.

Model	BLEU	ROUGE	BERTS	$\mathcal{K}$-BERTS	$\mathcal{K}$-Precision
Flan-T5$_{\text{BASE}}$	4.15	19.5	53.78	42.17	0
+BEINFO	5.39	21.04	54.68	70.03	27.78
Flan-T5$_{\text{LARGE}}$	5.01	20.02	54.56	61.77	0
+BEINFO	9.27	29.29	61.75	86.58	6.67
Flan-T5$_{\text{XL}}$	5.26	22.21	56.13	65.52	6.67
+BEINFO	10.2	30.76	62.78	88.50	100

Table 3: Zero-shot results on FaithDial. The models are tuned on DoQA.

Model	BLEU	ROUGE	BERTS	$\mathcal{K}$-BERTS	$\mathcal{K}$-Precision
Flan-T5$_{\text{XL}}$	25.88	41.68	66.91	66.42	100
+BEINFO	23.28	36.22	63.02	81.77	100
Tk-Instruct-3B	20.23	31.60	58.47	69.45	100
+BEINFO	29.19	42.56	66.24	70.58	97.8

Table 4: Zero-shot results on DoQA *Travel* domain. The models are tuned on FaithDial + TopiOCQA.

set. While the absolute scores, as expected, do differ between different underlying models, the results in Table 4 indicate the positive effect of BEINFO also on Tk-Instruct-3B.

BEINFO with Task-Specific Fine-Tuning. We have demonstrated that the models tuned with BEINFO largely improve factual faithfulness on unseen datasets and domains (i.e., the general BEINFO setup). Here, we study whether these models can serve as an effective starting point for continued task-specific fine-tuning. To this end, we first tune the models with BEINFO on the combination of FaithDial and TopiOCQA as before, and then continue fine-tuning/specialising the model on a single dataset (e.g., FaithDial or TopiOCQA): the *full* setup from Figure 2.[5]

Figure 3 demonstrates that already *task-only* BEINFO yields strong performance, while models with BEINFO perform on par or better on average than the models which were tuned to a specific dataset *both* on semantic similarity of generated responses and factual faithfulness. While prior work (Daheim et al., 2023) typically optimised one aspect (e.g., semantic similarity) at the expense of the other (faithfulness), and vice versa, here we show that through the use of knowledge distractors BEINFO achieves competitive performance on both aspects and retains the cross-dataset generalisation ability.

BEINFO versus Catastrophic Forgetting. Further, one issue which might arise from further specialising a model to a given task/dataset is a well-known phenomenon of *catastrophic forgetting*: pretrained language models are prone to forgetting

[5]We present the results only with Flan-T5$_{\text{BASE}}$ as preliminary experiments with larger model sizes demonstrated similar relative trends.

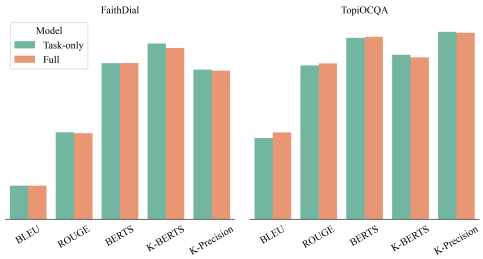

Figure 3: Results of task-specific tuning on FaithDial (left) and TopiOCQA (right). *'Task-only'* denotes Flan-T5 tuned directly on FaithDial or TopiOCQA, again with knowledge distractors. *'Full'* denotes the model first tuned with BEINFO on both datasets and then further tuned on each of the datasets; see Figure 2.

Model	BLEU	ROUGE	BERTS	$\mathcal{K}$-BERTS	$\mathcal{K}$-Precision
Task-only: FaithDial	27.29	41.71	75.31	64.80	73.38
General-only	38.75	69.57	80.40	72.24	81.24
Full: FaithDial	33.87	55.85	78.46	74.18	79.63
Task-only: TopiOCQA	36.24	68.64	80.94	73.38	83.63

Table 5: Results on TopiOCQA when the BEINFO model is further fine-tuned on FaithDial after the original FaithDial + TopiOCQA fine-tuning. 'Task-only: TopiOCQA' denotes direct tuning on TopiOCQA, which serves as an upper bound in this experiment.

previously learnt knowledge or skills when tuned on new data (De Cao et al., 2021; Yuan et al., 2023). To evaluate whether the models would retain their ability to respond faithfully to examples considerably different from the ones seen during fine-tuning, we evaluate the models tuned on FaithDial on TopiOCQA.[6] The scores in Table 5 demonstrate that even after continued fine-tuning on FaithDial the model retains high faithfulness scores on TopiOCQA (cf. $\mathcal{K}$-BERTS and $\mathcal{K}$-Precision). At the same time, degradation in scores for similarity to ground truth responses shows that further tuning largely influences the style/form of the responses. The average response length in FaithDial in considerably larger than that in TopiOCQA (see Appendix B), meaning that further tuning on FaithDial leads the model to generate longer responses not matching the gold responses in TopiOCQA. In other words, these results show that further fine-tuning might influence the surface form of the responses but not the desired skill to respond faithfully gained with BEINFO. In practice, a general model tuned with BEINFO on a wide range of tasks/domains and then specialised to one of them would still retain its ability to respond faithfully for *any* of the domains seen in the general 'behavioural tuning' step.

[6]We focus on TopiOCQA as the true responses in the dataset are more grounded in the knowledge source $\mathcal{K}$ (see Appendix B).

5 Evaluating BEINFO on Real Conversations

Experimental Setup. To probe the potential of BEINFO for boosting real user-facing production systems, we rely on a small internal dataset of 200 fully anonymised dialogs with real users in the hotel reservation domain (termed HOTEL-200 henceforth); the dialogues concern hotel bookings and FAQ-s about its various facilities. It is crucial to evaluate the models on examples also collected from real user-system communication, as the language use is considerably different to some established datasets such as DoQA or FaithDial compiled via crowdsourcing work. For instance, the average length of the user query in HOTEL-200 is only 6.35 tokens, while it is 17.25 in FaithDial or 13 in DoQA (cf., Table 1).

As the data comes from real conversations, there are no gold responses which could be used for automated evaluation. Thus, we resort to evaluation of correctness/factual faithfulness with an LLM: here, we use GPT4 (termed *GPT4-Eval* henceforth) as its judgements were shown to be most correlated with human judgements (Adlakha et al., 2023).[7] For *GPT4-Eval* we prompt GPT4 to act as the evaluator providing it with natural language instructions, knowledge source $\mathcal{K}$, conversation history $\mathcal{H}$ with user query u and the system-generated response. In the instructions we request the model to rate generated responses on a 7-point Likert-scale for faithfulness, available in Appendix E.

We compare the following models and their configurations: (i) GPT4 itself as the model responding to user query u, (ii) Falcon-40B (Almazrouei et al., 2023) as a strong open-source LLM,[8] (iii) Flan-T5$_{\text{XL}}$ tuned with BEINFO, under the three different regimes illustrated before in Figure 2 (*general-only, task-only, full*). For the *general-only* and the first stage of the *full* BEINFO, we again rely on the combination of FaithDial and TopiOCQA datasets.

To obtain data for the task-specific tuning stage, we collect 2,000 examples from the same conversational system, then generate 'silver' responses via GPT4 and treat the silver responses as true outputs for task-specific fine-tuning.[9]

GPT-4	Falcon-40B	XL-original	XL+BEINFO (g)	XL+BEINFO (t)	XL+BEINFO (f)
4.63	3.60	3.55	3.98	4.46	**4.81**

Table 6: Averaged *GPT4-Eval* scores (higher is better) on the HOTEL-200 dataset. XL denotes the Flan-T5$_{\text{XL}}$ model taken-off-the-shelf (XL-original) or fine-tuned via three different regimes of BEINFO (t=task-only; g=general-only; f=full).

Results and Discussion. The main results are reported in Table 6. While the zero-shot BEINFO approach with Flan-T5$_{\text{XL}}$ achieves a reasonably high average faithfulness score in absolute terms, it is still far from that of GPT4, which serves as an upper bound zero-shot system. Most importantly, the progress in scores reveals the importance of various BEINFO fine-tuning stages. Even the *general-only* fine-tuning stage without seeing a single in-domain training example yields an average score which is substantially higher than that of the original Flan-T5$_{\text{XL}}$ as well as higher than the score obtained by the 40B Falcon model. Further, the scores indicate the importance of being able to fine-tune smaller models with in-domain data: the 3B model tuned with the full BEINFO even outperforms GPT4 on *GPT4-Eval*, and it also obtains strong performance with *task-only* BEINFO.

These results further support our hypothesis that BEINFO actually 'behaviourally prepares' the models to respond to user's queries in a factually faithful manner and tuning on further task-specific data only amplifies its impact as it gets further adapted to the domain. Put simply, behavioural fine-tuning via BEINFO performs structural (or behavioural) adaptation, while further task-specific fine-tuning combines the behavioural adaptation with (semantic) domain adaptation.

Ablation: Distributions of Scores. We further study the actual distributions of *GPT4-Eval* scores for the four models variants of Flan-T5$_{\text{XL}}$ and compare it against the distribution obtained by GPT-4. The distributions are shown in Figure 4. As only a small fraction of responses is labelled with intermediate scores (1,2,3,5), the core differences lie in relative distribution of *perfect*, *poor* and *'not great, not terrible'* responses (scores 6,0 and 4, respectively).[10] The model tuned with task-only BEINFO rarely provides wrong facts but mostly responds

[7] As running evaluation with large models such as GPT4 behind proprietary APIs incurs large costs (Adlakha et al., 2023), we only evaluate the outputs for a smaller dataset where other means of evaluation cannot be used.

[8] Falcon-40B was an open-source large language model with state-of-the-art results at the time of the experimentation.

[9] Note that here we use the GPT4 model for three different purposes: (i) as an evaluator; (ii) as an actual baseline system; (iii) as a 'silver data generator'.

[10] Score 4 usually corresponds to the system responding with a generic clarification question or notifying the user that the information is not available.

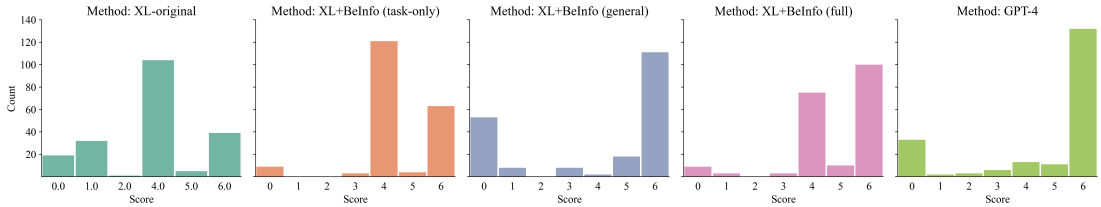

Figure 4: Distribution of *GPT4-Eval* scores of 4 variants based on Flan-T5$_{XL}$ and GPT-4. See Table 12 in Appendix E for the interpretation of the individual scores.

with *not great, not terrible* responses which do not mislead the user but might not be helpful. On the other hand, the model tuned with the general-only BEINFO and GPT-4 both yield responses that typically fall into the extreme categories. In other words, the responses are either perfect (score 6) or will provide the user with wrong information (score 0), which is not desirable for a user-facing system. The model tuned with the full BEINFO combines the benefits of behavioural tuning with the use of in-domain data: the model produces the least factually unfaithful responses (score 0) while maintaining the ability to respond with information relevant to the user's query (a large number of responses with scores 6). In sum, 'pre-tuning' the model with the general-only BEINFO stages raises faithfulness of the model by extracting relevant information from the knowledge source $\mathcal{K}$ while further tuning on task-specific data further helps avoid providing misleading or irrelevant information to the user.

Faithfulness versus Abstractiveness. Increasing faithfulness of a model to the underlying knowledge source $\mathcal{K}$ can lead the model to respond with large *extracted spans* of text from $\mathcal{K}$. Ideally, the responses should be abstractive but factually faithful: in other words, they should transmit the information provided in the knowledge source $\mathcal{K}$ but use different means of expression of it. As in prior work (Dziri et al., 2022a; Daheim et al., 2023), we use the Density metric proposed by Grusky et al. (2018) to measure abstractiveness. This measures average length of spans copied from the knowledge source. We focus on $\mathcal{K}$-BERTScore to measure faithfulness in this experiment: the average length of the knowledge source $\mathcal{K}$ ($\approx$120 words on average) is relatively large with respect to the length of generated responses ($\approx$ 18–25 words) making $\mathcal{K}$-BERTScore suitable for this case.

Figure 5 illustrates the trade-off between faithfulness and abstractiveness for Flan-T5$_{XL}$ under different fine-tuning setups on the HOTEL-200 dataset. The results demonstrate that general-only

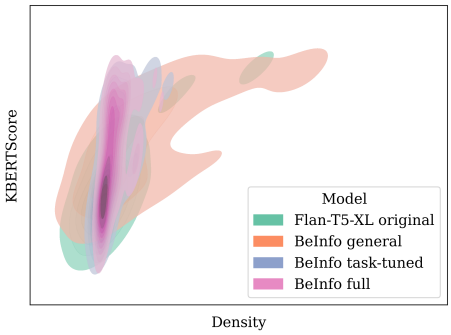

Figure 5: Density and $\mathcal{K}$-BERTScore on HOTEL-200 illustrating the trade-off between faithfulness (y-axis) and abstractiveness (x-axis) for Flan-T5$_{XL}$ for different setups: (i) XL-original: 'off-the-shelf' Flan-T5$_{XL}$; (ii) BEINFO general-only: Flan-T5$_{XL}$ tuned with BEINFO on FaithDial and TopiOCQA without any in-task data; iii) BEINFO task-only: Flan-T5$_{XL}$ finetuned only on task-specific data; iv) BEINFO full. Numeric results are provided in Appendix F.

fine-tuning with BEINFO improves the model's factuality but increases the extractiveness of the responses. Tuning on task-specific data helps to raise the abstractiveness of the responses. Further analyses and comparisons (cf. results in Appendix F) demonstrate that Flan-T5$_{XL}$ tuned with the full BEINFO is on par with GPT-4 and better than a considerably larger Falcon-40B model.

Human Evaluation. In addition to automatic metrics, we also conduct human evaluation on HOTEL-200 with two annotators. They were tasked to rate each response on factuality using the same Likert-scale as used for *GPT4-Eval* (see Appendix E). Three models were assessed: XL+BEINFO-general, XL+BEINFO-task-specific and XL+BEINFO-full. Average human factuality scores were 3.64, 4.62 and 4.95, respectively. This further proves the effectiveness of behavioural tuning for improved factuality. To further assess relevance of automatic *GPT4-Eval*, we also compute Pearson's correlation coefficient ρ between human judgements and GPT4-Eval scores. This results in strong positive correlation with $\rho = 0.52$, indicating that *GPT4-*

Eval can be used as a reasonable automatic proxy.

6 Related Work

Mitigating Hallucinations in Information-Seeking Dialogue has achieved increased interest recently with the omnipresence of large language models (Wang et al., 2023b; Chuang et al., 2023; Daheim et al., 2022, 2023; Zhang et al., 2023). Previous methods can be largely divided into those which increase factuality of pretrained models via further training or modification of the generation procedure. The former includes, e.g., tuning the models with contrastive learning (Sun et al., 2023) or a special *focus learning* loss which reduces hallucinations on token level (Deng et al., 2023). The latter includes, e.g., conditioning generation process on special control tokens (Rashkin et al., 2021), task arithmetic (Daheim et al., 2023) or training a critic network which can detect problematic tokens and replace them (Dziri et al., 2021). Other approaches have been developed to specifically improve faithfulness with respect to retrieved knowledge source in decoding. One proposed option is to do context-aware decoding (CAD; Shi et al., 2023) where generative probabilities are contrasted between those based only on user query and those based on the user query *and* the knowledge source. The aim is to force LLMs to rely more on the knowledge source than the model's internal knowledge from pretraining. In contrast to CAD, Chuang et al. (2023) propose to contrast generation probabilities from different layers of LLMs to promote factual knowledge in the resulting output probabilities.

Improving Faithfulness via Supervised Tuning. Task-specific supervised fine-tuning could be seen as an option to improve faithfulness of the model's responses (Zhang et al., 2023). Prior work (Cao et al., 2023; Chen et al., 2023) has demonstrated that fine-tuning on higher-quality data improves the model's factuality on benchmarks such as TruthfulQA (Lin et al., 2022). In contrast, supervised fine-tuning on the data which includes numerous irrelevant or factually inconsistent responses can lead the model to amplifying the noise in the training data. A recent analysis from Dziri et al. (2022b) has shown that over 60% of responses in three standard datasets for information-seeking dialogue (WoW, Dinan et al., 2018; CMU-DoG, Zhou et al., 2018; and TopicalCHAT, Gopalakrishnan et al., 2019) contain hallucinations, making them unsuitable for supervised fine-tuning aimed at improving factuality. To resolve this, Dziri et al. (2022a) released a corrected version of WoW where the responses were fixed to be factually consistent with the knowledge source. As behavioural fine-tuning heavily relies on the quality of the underlying data, we have carefully selected and resorted to FaithDial and TopiOCQA in the first stage of BEINFO with highest factual faithfulness of their ground truth responses (see Appendix B for further details).

7 Conclusion and Future Work

We presented BEINFO, a simple yet effective method that applies behavioural fine-tuning of large language models underlying information-seeking dialogue systems, with the goal of improving factuality of system responses. Instruction-tuned models are fine-tuned on a collection of publicly available dialogue data for two related tasks, conversational question answering and information-seeking dialogue, where the model must use the correct knowledge source among several 'knowledge distractors' and provide a factually correct and adequate response. The main results indicated the effectiveness of BEINFO both in in- and cross-dataset setups. In addition, we demonstrated that further tuning on task-specific data might yield further gains in terms of faithfulness as well as reducing extractiveness, also in experiments with real conversations from a production-ready dialogue system.

This work leads up to several potential directions of future work. Firstly, BEINFO is orthogonal to other existing approaches to improving faithfulness. For instance, a combination of CAD (Shi et al., 2023) and BEINFO could further improve factuality of responses. Secondly, BEINFO was evaluated on information-seeking dialogue. Another interesting direction could be to applying it to other language generation tasks where faithfulness to the knowledge sources is crucial, such as summarisation. Furthermore, the effectiveness of the approach can be also tested on other instruction-tuned models (e.g., T0, Sanh et al., 2021) and models of larger sizes, e.g., Flan-UL2 and beyond.[11]

The code and models will be made available online at [URL], allowing the research community to build stronger models for factually faithful information-seeking dialogue.

[11] Due to a large number of experiments coupled with computational constraints and feasibility, we focus on models that do not go beyond 3B parameters.

Limitations

The experiments could be further extended by altering how the knowledge distractors $\mathcal{K}'$ are sourced. Firstly, the impact of the number n of knowledge distractors $\mathcal{K}'$ on faithfulness performance should be further studied. Also, another extension on this front concerns different heuristics of how $\mathcal{K}'$ is sampled. Namely, in our experiments they were sampled at random, while getting $\mathcal{K}'$ which are semantically similar or distant from the true knowledge source $\mathcal{K}$ or user query u might further impact performance.

In the experiments we focus on three widely used datasets for information seeking dialogue and two instruction-tuned models. BEINFO can be further extended to other datasets such as CoQA (Reddy et al., 2019), MultiDoc2Dial (Feng et al., 2021) or the DSTC9 (Kim et al., 2020) extension of MultiWOZ 2.1 (Eric et al., 2020). The evaluation on production-ready dialogues, due to associated costs of evaluation, is conducted on 200 dialogues, and we plan to run a larger-scale analysis, also spanning other dialogue domains, in future work.

We also tested whether BEINFO can be used with parameter-efficient finetuning (PEFT) to reduce its computational cost. Our preliminary experiments proved that BEINFO can be effectively combined with PEFT. However, as PEFT techniques are out of the scope of the paper and their use is orthogonal to the main experiments reported in this work, we leave out the preliminary results and focus on full fine-tuning as our main setup.

Given that BEINFO uses instruction-tuned models and 'behaviourally' tunes them with a predefined instruction, additional experimentation could be conducted on how the wording of the instruction influences the performance and whether one can induce higher factuality by just changing the instruction text.

Finally, the work on improving knowledge retrieval systems as done e.g. by Mo et al. (2023) is out of scope of this work, and we focus on reducing hallucinations of LLMs in information-seeking dialogue directly, without the intervention to the knowledge retrieval component.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv preprint arXiv:2307.16877*.

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics.

Yihan Cao, Yanbin Kang, and Lichao Sun. 2023. Instruction mining: High-quality instruction data selection for large language models. *arXiv preprint arXiv:2307.06290*.

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

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

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

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation. *arXiv preprint arXiv:2303.17574*.

Nico Daheim, David Thulke, Christian Dugast, and Hermann Ney. 2022. Controllable factuality in document-grounded dialog systems using a noisy channel model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1365–1381, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–

6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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)*, pages 4554–4566, Toronto, Canada. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

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, Andrea Madotto, Osmar Zaïane, 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.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anushree Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Interspeech 2019*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

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.

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. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*.

Fengran Mo, Jian-Yun Nie, Kaiyu Huang, Kelong Mao, Yutao Zhu, Peng Li, and Yang Liu. 2023. Learning to relate to previous turns in conversational search. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA. Association for Computing Machinery.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

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.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. 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.

Sebastian Ruder. 2021. Recent Advances in Language Model Fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13618–13626.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

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, Krima 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. 2022. 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Zhangdie Yuan, Songbo Hu, Ivan Vulić, Anna Korhonen, and Zaiqiao Meng. 2023. Can pretrained language models (yet) reason deductively? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1447–1462, Dubrovnik, Croatia. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

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, Brussels, Belgium. Association for Computational Linguistics.

```
Please answer the following user query given
the information and the conversation.

INFORMATION:
I trim the stem and remove the outer leaves till
they snap to get to the fresh inner core and steam
them the night or morning before grilling so they
are cold and moist. I prefer steaming because I
want all of the nutrients to remain in the
artichoke. I cut them in half for the grill, remove
the choke and brush them with grapeseed oil where
they come into contact with the grill. First,
facing down grill them till they feel hot on top
then; flip them over to keep the yummy inner side
tender. fill the cavity with garlic butter and...
lemon if you wish. I prefer the brown color to the
lemon flavor.

User: How do I best grill an artichoke?
Agent: Cut them in half for the grill, remove the
choke and brush them with grapeseed oil where they
come into contact with the grill.
User: Are there other ways to cook it?
Agent:
```

Figure 6: Example with the prompt used in BEINFO.

A Example of Input

An example with instructions used in BEINFO is shown in Figure 6. The used prompt is similar to the one which proved successful for conversational question-answering in Adlakha et al. (2023).

B Additional Dataset Statistics and Characteristics

We present overall statistics of the datasets used for BEINFO and evaluation in Table 1.

Additionally, we analyse the characteristics of factual faithfulness of the true responses with respect to the knowledge source. The results in Table 7 demonstrate that the responses in FaithDial (Dziri et al., 2022a) are semantically most similar to their knowledge source, which is in line with the dataset collection procedure aimed to make the dataset more factual than the original responses. Similarity of contextual semantic token representations (BERTS-F1) is reversely correlated to lexical overlap between the response and knowledge source.

As BEINFO is aimed at improving the model's general factual faithfulness, the results suggest that FaithDial (Dziri et al., 2022a) and TopiOCQA (Adlakha et al., 2022) are best used for behavioural tuning and DoQA for testing the out-of-distribution capabilities of the model. The former two have a large semantic but not literal overlap between the knowledge source and the corresponding golden response, meaning that the behavioural tuning will not lead to model learning to '*copy-paste*' from the knowledge source to the response.

		FaithDial	TopiOCQA	DoQA
$(y, \mathcal{K})$	$\mathcal{K}$-BERTS-F1	67.31	62.91	52.48
	$\mathcal{K}$-Precision	46.23	80.67	97.73
y	Avg. length	17.17	10.89	13.29

Table 7: BERTScore-F1 and K-Precision between the ground truth knowledge source $\mathcal{K}$ and gold response y. Average length is calculated as an arithmetic mean of number of whitespaced words in a response.

C Per-Domain Performance on DoQA

Tables 8 – 10 present per-domain results of BEINFO (general-only) on DoQA.

Model	BLEU	ROUGE	BERTS	$\mathcal{K}$-BERTS	$\mathcal{K}$-Precision
Flan-T5$_{\text{BASE}}$	23.77	34.19	61.54	67.80	100.0
+BEINFO	23.96	34.65	61.74	79.54	100.0
Flan-T5$_{\text{LARGE}}$	27.35	39.79	64.83	71.78	100.0
+BEINFO	28.17	40.57	64.07	77.77	100.0
Flan-T5$_{\text{XLARGE}}$	32.16	42.99	65.93	68.19	100.0
+BEINFO	28.76	42.68	65.97	81.65	100.0

Table 8: Zero-shot results on DoQA *Cooking*.

Model	BLEU	ROUGE	BERTS	$\mathcal{K}$-BERTS	$\mathcal{K}$-Precision
Flan-T5$_{\text{BASE}}$	21.37	34.23	60.99	70.69	69.70
+BEINFO	21.93	34.51	61.52	73.99	100.0
Flan-T5$_{\text{LARGE}}$	23.64	37.34	62.99	72.47	81.58
+BEINFO	25.57	38.99	63.22	71.52	100.0
Flan-T5$_{\text{XLARGE}}$	27.94	41.30	64.83	67.01	82.35
+BEINFO	27.90	39.27	64.81	77.14	100.0

Table 9: Zero-shot results on DoQA *Movies*.

Model	BLEU	ROUGE	BERTS	$\mathcal{K}$-BERTS	$\mathcal{K}$-Precision
Flan-T5$_{\text{BASE}}$	23.52	34.96	62.27	64.76	100.0
+BEINFO	22.40	32.95	61.88	79.12	100.0
Flan-T5$_{\text{LARGE}}$	27.50	41.59	66.02	69.90	100.0
+BEINFO	25.27	36.10	62.29	77.35	100.0
Flan-T5$_{\text{XLARGE}}$	25.88	41.68	66.91	66.42	100.0
+BEINFO	23.28	36.22	63.02	81.77	100.0

Table 10: Zero-shot results on DoQA *Travel*.

D Zero-Shot Results on TopiOCQA

The results on TopiOCQA when the smaller dataset DoQA is used for BEINFO fine-tuning are presented in Table 11.

Model	BLEU	ROUGE	BERTS	$\mathcal{K}$-BERTS	$\mathcal{K}$-Precision
Flan-T5$_{\text{BASE}}$	19.10	43.44	63.72	68.17	100.0
+BEINFO	16.08	31.41	58.87	68.85	100.0
Flan-T5$_{\text{LARGE}}$	23.26	42.0	63.64	75.83	100.0
+BEINFO	24.47	37.16	62.31	76.33	100.0
Flan-T5$_{\text{XL}}$	22.41	42.52	63.79	77.43	100.0
+BEINFO	27.13	40.59	62.58	76.89	100.0

Table 11: Zero-shot results on TopiOCQA when DoQA is used for BEINFO fine-tuning.

6	Only information asked for (perfect)
5	Information asked for, but also provided additional information that is relevant to the supporting facts (good)
4	Follow-up or generic question (No specific information is asked for), agent asked for clarification (not great not terrible)
3	Information asked for, but also provided additional information that is irrelevant to the supporting facts (not bad)
2	Transfer the customer to the correct customer service department (ok)
1	No information asked for, but provided additional information that is either relevant or irrelevant to the supporting facts (bad)
0	Information provided is not coming from the supporting facts (terrible), or transfer customers to the wrong queue (poor)

Table 12: Likert-scale for evaluating faithfulness automatically via GPT4.

E Evaluating Faithfulness with *GPT4-Eval*

The 7-point Likert scale used to evaluate faithfulness via GPT4 (i.e., the *GPT4-Eval* evaluation metric) is provided in Table 12.

F Results for Faithfulness vs. Abstractiveness

The results for factual faithfulness and abstractiveness on real conversations for Flan-T5$_{XL}$ tuned with BEINFO and larger language models are shown in Figure 7. Results demonstrate that BEINFO approximates a much smaller model to the performance of GPT-4 while overcoming the performance of a much larger open-source model, Falcon-40B. The exact numbers are shown in Table 13.

Model	Density (↓)	Coverage (↓)	$\mathcal{K}$-BERTScore (↑)
Flan-T5-XL	4.03	0.47	83.51
BEINFO (t)	2.35	49.18	84.64
BEINFO (g)	12.10	0.73	88.33
BEINFO (f)	2.32	0.60	86.30
Falcon-40B	5.72	0.46	84.25
GPT-4	2.01	0.64	87.49

Table 13: Results for faithfulness and abstractiveness on real user conversations. We use: a) $\mathcal{K}$-BERTScore to measure faithfulness of the model to the knowledge source $\mathcal{K}$; b) Density and Coverage (Grusky et al., 2018) to measure abstractiveness of the responses. (t)=task-tuned; (g)=general-only; (f)=full.

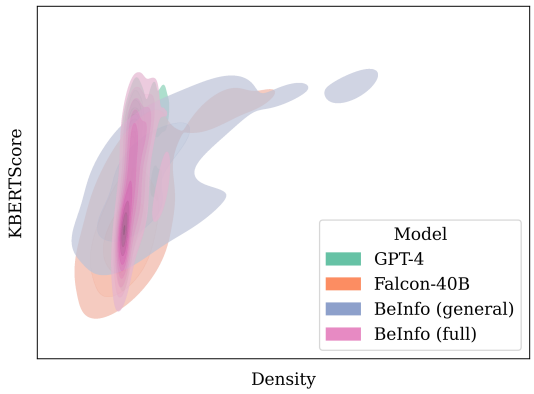

Figure 7: Density and $\mathcal{K}$-BERTScore illustrating the trade-off between faithfulness (y-axis) and abstractiveness (x-axis).