Elastic Weight Removal for Faithful and Abstractive Dialogue Generation

Anonymous ACL submission

Abstract

Generating factual responses is a crucial requirement for dialogue systems. To promote more factual responses, a common strategy is to ground their responses in relevant documents that inform response generation. However, common dialogue models still often hallucinate information that was not contained in these documents and is therefore unfaithful. In this work, we propose to alleviate such hallucinations by ‘subtracting’ the parameters of a model trained to hallucinate from a dialogue response generation model in order to ‘negate’ the contribution of such hallucinated examples from it. Extensive automatic and human evaluation shows favourable results when compared to state-of-the-art methods that combine the distributions of multiple models, such as DExperts (Liu et al., 2021), and others that change the training procedure, such as Quark (Lu et al., 2022a). Finally, we show how we can not only reduce hallucinations but also discourage extractive responses, which are often a consequence of reducing hallucinations by encouraging copy-pasting of document spans. We will publicly release our code for reproducibility and facilitating further research.

1 Introduction

Current-day large language models (LLMs) impressively generate coherent, grammatical, and seemingly meaningful text, but are prone to hallucinating incorrect information. While grounding them in relevant documents can alleviate this (Shuster et al., 2021), models still tend to generate information that conflicts these documents, which would again be classified as hallucination (Dziri et al., 2022a). This raises major safety concerns. Such hallucinations could impair student learning, or proliferate convincing-but-inaccurate news articles. Therefore, ensuring trustworthiness is crucial for the safe deployment of LLMs at scale, particularly in high-stakes domains.

$\mathcal{K}$: The Flash first appeared in “Showcase” #4 (October 1956) [...]
$\mathbf{u}_T$: What comic series is he from?

$\mathbf{u}_{T+1}$	F	A
He first appeared in “Showcase” #4 (November 1956).	✗	✗
He first appeared in “Showcase” #4 (October 1956).	✓	✗
His first appearance was in Showcase #4 in October 1956.	✓	✓

Figure 1: Constructed example of responses $\mathbf{u}_{T+1}$ that are i) hallucinated (words contradicting the knowledge $\mathcal{K}$ in red); ii) faithful but not abstractive (longest copied n-gram in blue); and iii) both **F**aithful and **A**bstractive based on Wizard-of-Wikipedia (Dinan et al., 2019).

Modelling solutions to mitigate hallucination often take inspiration from methods used to discourage other undesirable behaviours in LLMs, for example, contradictions (Keskar et al., 2019), repetitions (Lu et al., 2022a), or toxicity (Ilharco et al., 2023). One group of methods achieves this by fine-tuning an LLM conditioned on special tokens (Niu and Bansal, 2018; Keskar et al., 2019), which can be assigned to model generations by a learned reward model during training (Lu et al., 2022a). Another re-weights the predictive distribution with models that are specialised for positive or negative behaviour (Liu et al., 2021; Daheim et al., 2022), called ‘experts’ or ‘anti-experts’ respectively. While successful, these methods are either inefficient to train, as a large number of generations needs to be sampled during training, or inefficient in inference, as multiple models have to be stored and evaluated. In this work, we explore a different family of methods (Choubey et al., 2021; Ilharco et al., 2023) that uses modular deep learning (Ponti et al., 2021; Pfeiffer et al., 2023) by interpolating parameters without altering the model architecture. This is efficient during inference, because only one interpolated model needs to evaluated, and for training the models that are interpolated no new data needs to be sampled during the training procedure. Concretely, a new model is obtained as the weighted difference between a pretrained LLM and a model finetuned from it, for example, as an anti-expert (Ilharco et al., 2023). One drawback of this strategy is that parameters are weighted uniformly even though they might have differing contributions to hallucinations. Furthermore, it might result in catastrophic interference between the specialised models (McCloskey and Cohen, 1989). To address this, we propose Elastic Weight Removal (EWR), a novel method for parameter interpolation that weights the importance of each parameter by using the Fisher Information Matrix (FIM) as a measure of importance, similar to previous works in continual learning (Kirkpatrick et al., 2017), sample-efficient learning (Ponti et al., 2019), or merging models for different tasks (Matena and Raffel, 2022). In our experiments, we show how this can be used to discourage hallucinations by first training an anti-expert on synthetically created data and then interpolating it with the baseline model.

We compare our method with state-of-the-art methods for removing hallucinations and other undesired behaviours, which we adapt to removing hallucinations. Namely, we adapt Quark (Lu et al., 2022a), DExperts (Liu et al., 2021), and task arithmetic (Choubey et al., 2021; Ilharco et al., 2023). Our findings show consistent improvements in faithfulness, which can be combined with those of others, such as CTRL (Rashkin et al., 2021). Oftentimes, an increase in faithfulness comes at an increase in extractiveness from copy-pasting document spans into the response. Based on this insight, we finally highlight how EWR can be extended to reducing hallucinations and extractiveness at the same time. Our results are confirmed using a human evaluation with the Attributable to Identified Source (AIS) framework (Rashkin et al., 2023). We will release the code for all methods and metrics in a comprehensive framework.

2 Background

The goal of dialogue response generation is to continue a dialogue $\mathbf{u}_1^T := (\mathbf{u}_1, \ldots, \mathbf{u}_T)$ of T turns by generating a new turn $\mathbf{u}_{T+1}$. Here, each turn $\mathbf{u}_t$ is just a sequence of N_t tokens $[\mathbf{u}_t]_1^{N_t} \in \mathcal{V}^{N_t}$ from the model vocabulary $\mathcal{V}$. In document-grounded response generation, $\mathbf{u}_{T+1}$ is grounded in one or more documents $\hat{\mathcal{K}} \subseteq \mathcal{K}$ from a document knowledge base $\mathcal{K}$, meaning that $\hat{\mathcal{K}}$ informs the information content of $\mathbf{u}_{T+1}$. Therefore, $\mathbf{u}_{T+1}$ should also faithfully reflect it. This means that neither contradicting nor unverifiable information should be added. In this work, we assume that $\hat{\mathcal{K}}$ is given.

A common strategy for generating $\mathbf{u}_{T+1}$ is using language generators that model the distribution

$$p_{\boldsymbol{\theta}}(\mathbf{u}_{T+1} \mid \mathbf{u}_1^T, \hat{\mathcal{K}}) = \prod_{n=1}^{N_{T+1}} p_{\boldsymbol{\theta}}([\mathbf{u}_{T+1}]_n \mid [\mathbf{u}_{T+1}]_1^{n-1}, \mathbf{u}_1^T, \hat{\mathcal{K}}), \quad (1)$$

parameterised by weights $\boldsymbol{\theta}$, for next-token prediction paired with a search algorithm like beam search. We focus on different methods of obtaining $\boldsymbol{\theta}$ while maintaining the same model architecture.

2.1 Parameter Combination for Faithful Generation

Previous works have explored combining model parameters with different goals, for example, to increase robustness (Gao et al., 2022) but also to promote or discourage different behaviours by merging specifically trained model instances (Ilharco et al., 2023). In this work, we use it to discourage hallucinations in dialogue models. By letting $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$, where $\boldsymbol{\theta}_i \in \mathbb{R}^d$, denote the parameters of a set of models that should be merged and $\boldsymbol{\lambda}_i \in \mathbb{R}^d$ their respective scaling factors, many such methods can be expressed by:

$$\boldsymbol{\theta}' = \sum_{i=1}^{N} \frac{\boldsymbol{\lambda}_i \odot \boldsymbol{\theta}_i}{Z}, \quad (2)$$

where $\odot$ denotes element-wise multiplication and Z can be used to re-scale parameters.

One such method is task arithmetic (Ilharco et al., 2023), which bases on the idea that essential information about a task can be captured by the change of the parameter values between pretrained initialisation $\boldsymbol{\theta}_0$ and the finetuned $\boldsymbol{\theta}_{\text{ft}}$, called task vector. Given this information, the behaviour needed for this task can be added to the model $\boldsymbol{\theta}_0$ by adding a task vector and also removed by subtracting it. Concretely, the task vector can be expressed as:

$$\boldsymbol{\tau} := \boldsymbol{\theta}_{\text{ft}} - \boldsymbol{\theta}_0. \quad (3)$$

Then, task arithmetic (Ilharco et al., 2023) uses the following for model combination:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}_0 + \sum_i \lambda_i \boldsymbol{\tau}_i, \quad (4)$$

where the scalar λ_i promotes the behaviour captured by $\boldsymbol{\tau}_i$ if $\lambda_i > 0$ and discourages it if $\lambda_i < 0$.

We will use the latter to discourage hallucinations by training a model to hallucinate and then discouraging its behaviour through subtraction. We will refer to such a model as 'anti-expert' ($\boldsymbol{\theta}_{\mathrm{AE}}$) and then use the following task arithmetic:

$$\begin{aligned}\boldsymbol{\theta}' &= \boldsymbol{\theta}_0 - \lambda \cdot \boldsymbol{\tau} \\ &= \boldsymbol{\theta}_0 - \lambda \cdot (\boldsymbol{\theta}_{\mathrm{AE}} - \boldsymbol{\theta}_0) \\ &= (1+\lambda) \cdot \boldsymbol{\theta}_0 - \lambda \cdot \boldsymbol{\theta}_{\mathrm{AE}}. \end{aligned} \quad (5)$$

We would expect a model parameterised by $\boldsymbol{\theta}'$ to hallucinate less than one parameterised by $\boldsymbol{\theta}_0$.

We could also add an expert model $\boldsymbol{\theta}_{\mathrm{E}}$, for example, trained on abstractive data which significantly rewrites the documents content:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}_0 - \lambda_{\mathrm{AE}} \cdot (\boldsymbol{\theta}_{\mathrm{AE}} - \boldsymbol{\theta}_0) + \lambda_{\mathrm{E}} \cdot (\boldsymbol{\theta}_{\mathrm{E}} - \boldsymbol{\theta}_0). \quad (6)$$

Setting $\lambda = \lambda_{\mathrm{AE}} = \lambda_{\mathrm{E}}$ is equivalent to using Contrastive Parameter Estimation (CaPE; Choubey et al., 2021) with the following simplified update:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}_0 + \lambda \cdot (\boldsymbol{\theta}_{\mathrm{E}} - \boldsymbol{\theta}_{\mathrm{AE}}). \quad (7)$$

We will discuss how to train $\boldsymbol{\theta}_{\mathrm{AE}}$ and $\boldsymbol{\theta}_{\mathrm{E}}$ later.

Both task arithmetic and CaPE use scalars λ for parameter combination and therefore assume equal parameter importance. Intuitively, though, only a subset of parameters might be responsible for hallucinations. For example, anomalous encoder–decoder attention patterns correlate strongly with hallucinations (Raunak et al., 2021; Guerreiro et al., 2023, *inter alia*). Hence, only these specific parameters might be required to change. Moreover, composing multiple task vectors might lead to catastrophic interference (Ansell et al., 2022). Next, we show how parameters can be weighed individually which we hope will improve task arithmetic.

3 Elastic Weight Removal

In our proposed method, Elastic Weight Removal (EWR), we use the Fisher Information matrix (or Fisher) to combine models with importance-weighted scaling factors for each parameter. Thereby, we aim to preserve positive behaviour in the model fine-tuned for dialogue response generation while removing the most important parameters in the anti-expert task vector, which lead to hallucinated generations. We take inspiration from prior works that successfully use the Fisher for similar parameter-specific scaling, for example, against catastrophic forgetting (Kirkpatrick et al., 2017), for merging checkpoints of the same model trained independently on different tasks (Matena and Raffel, 2022), or preconditioning updates in stochastic optimization (Amari, 1998; Martens, 2020). We refer the reader to prior works (Schraudolph, 2002; Martens, 2020; Kunstner et al., 2019) for more information about theoretical properties of the Fisher. Of practical importance is that the Fisher has size d^2 for a neural network model with d parameters. Therefore, it is commonly approximated by its diagonal (Matena and Raffel, 2022, inter alia). The diagonal can be estimated efficiently by summing or averaging the squared gradients of the model over the training data. Here, the label is sampled from the model at each step instead of taking the annotated token (cf. Kunstner et al. (2019)). For a model $p_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{x})$ this means calculating: $\mathbf{f}_{\boldsymbol{\theta}} = \frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}}[\nabla \log p_{\boldsymbol{\theta}}(\mathbf{y}' \mid \mathbf{x})]^2$, where $\mathbf{y}' \sim p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{x})$ is sampled from the model.

We start by taking Equation (2) and setting $\boldsymbol{\lambda}_0$, which scales pre-trained parameters $\boldsymbol{\theta}_0$, to $\lambda_0 \cdot \mathbf{f}_{\boldsymbol{\theta}}$ (note that λ_0 is equal to 1 in Equation (5) for task arithmetic). Similarly, for each task vector $\boldsymbol{\tau}_i$, we replace the scalar factor λ_i with $\lambda_i \cdot \mathbf{f}_{\boldsymbol{\tau}_i}$. This way, we can still control the influence of each model with a scalar hyper-parameter, while the diagonal Fisher estimate controls individual parameters. Since the entries in $\mathbf{f}$ can have different magnitudes than the entries in $\boldsymbol{\theta}$, we use a scaling constant Z. Then, our parameter combination is defined as:

$$\boldsymbol{\theta}' = \frac{\lambda_0 \cdot \mathbf{f}_{\boldsymbol{\theta}_0} \cdot \boldsymbol{\theta}_0 + \sum_{i=1}^{N} \lambda_i \cdot \mathbf{f}_{\boldsymbol{\tau}_i} \cdot \boldsymbol{\tau}_i}{Z}, \quad (8)$$

One choice is to set $Z := \lambda_0 \cdot \mathbf{f}_{\boldsymbol{\theta}_0} + \sum_i |\lambda_i| \cdot \mathbf{f}_{\boldsymbol{\tau}_i}$, similar to Matena and Raffel (2022). Then, using only a hallucination anti-expert $\boldsymbol{\theta}_{\mathrm{AE}}$, we can rewrite the update as:

$$\boldsymbol{\theta}' = \boldsymbol{\theta}_0 - \frac{\lambda_{\mathrm{AE}} \cdot \mathbf{f}_{\boldsymbol{\tau}_{\mathrm{AE}}}}{\lambda_0 \cdot \mathbf{f}_{\boldsymbol{\theta}_0} + \lambda_{\mathrm{AE}} \cdot \mathbf{f}_{\boldsymbol{\tau}_{\mathrm{AE}}}} \boldsymbol{\theta}_{\mathrm{AE}}. \quad (9)$$

Therefore, $\mathbf{f}_{\boldsymbol{\theta}_0}$ and $\mathbf{f}_{\boldsymbol{\tau}_{\mathrm{AE}}}$ determine how much each parameter should be changed—parameters with large $\mathbf{f}_{\boldsymbol{\theta}_0}$ are preserved and parameters with large $\mathbf{f}_{\boldsymbol{\tau}_1}$ are changed more due to their contribution to negative behaviour. When an expert model is added, as well, it is only possible to obtain a similar rewrite when the sign of the corresponding α_i is flipped in the denominator, i.e. $Z := \lambda_0 \cdot \mathbf{f}_{\boldsymbol{\theta}_0} + \sum_i (-\lambda_i) \cdot \mathbf{f}_{\boldsymbol{\tau}_i}$. We have found this to be more stable empirically. However, it can introduce divisions by 0 which can be avoided by adding a small constant. Finally, we have found calculating the Fisher at $\boldsymbol{\tau}$ to perform well empirically, even though calculating it at $\boldsymbol{\theta}_{\mathrm{AE}}$ or $\boldsymbol{\theta}_{\mathrm{E}}$, respectively, is theoretically better grounded. Next, we describe how we train the expert and anti-expert models. Pseudocode for EWR is shown in Appendix A.1.

3.1 Training Data for (Anti-)Experts

We use different strategies to create hallucinated examples $\mathcal{D}^{\mathrm{AE}}$. For Wizard-of-Wikipedia (WoW), we use all examples from Faithdial (Dziri et al., 2022a) which humans rated as hallucinations according to the BEGIN taxonomy (Dziri et al., 2022c). Since such annotations often do not exist for other data, we try lightweight data augmentation techniques to artificially create hallucinated data. We find that replacing the ground-truth documents to randomly sampled ones performs similar to using human hallucination annotations. Potentially, this forces the model to hallucinate, as the input does not contain the correct information for the response. We use this strategy for all other datasets than WoW. CaPE and DExperts (which we introduce in detail in the following Section 4.2) also use a faithfulness expert in addition to a hallucination anti-expert. For training this expert, we use responses that are assigned an entailment token when training CTRL, because such examples are unlikely to contain hallucinations.

To create a dataset of abstractive examples $\mathcal{D}^{\mathrm{E}}$, we use the density and coverage metrics introduced in Grusky et al. (2018). Coverage measures the ratio of unigrams from the grounding documents that appear in the response and density measures the average length of copied text spans. Intuitively, we would like to have low density, because this indicates paraphrasing, but such examples might be hallucinated. Therefore, we pick examples that also have high coverage to ensure that the information from the document is used. We do this by splitting the dataset into buckets and assigning low, medium, and high density or coverage tokens to them, similar to Keskar et al. (2019), and taking the high density examples. Future work can explore further methods for data augmentation.

4 Experiments

We experiment on multiple datasets outlined in Section 4.1. We compare EWR to CaPE and task arithmetic, as well as a set of other unlearning methods, which we apply for faithful dialogue generation for the first time. Furthermore, we compare to state-of-the-art methods for faithful dialogue generation. We list these baselines in Section 4.2. Crucially, parameter combination can be added independently on top of many of the other baselines.

All experiments are implemented using Huggingface transformers (Wolf et al., 2020) and models are initialised from publicly available Flan-T5 checkpoints (Longpre et al., 2023), which we have found to perform substantially better than previously introduced encoder-decoder models like BART (Lewis et al., 2020) or T5 (Raffel et al., 2020). We organise our experiments using Sisyphus (Peter et al., 2018) and release configuration files to reproduce our results. Further experimental details, such as learning rate or number of epochs, are given in Appendix B.1. We use beam search with a beam size of 10 for decoding.

4.1 Datasets

We evaluate all methods on Wizard-of-Wikipedia (Dinan et al., 2019, WoW), an open-domain dataset for information-seeking dialogue where turns are grounded in Wikipedia snippets. WoW contains a *seen* and an *unseen* split. Furthermore, we use the DSTC9 (Kim et al., 2020) extension of MultiWoZ 2.1 (Eric et al., 2020), which augments the original dialogues by turns that are grounded in short FAQ documents. For further experiments, we use DSTC11 (Zhao et al., 2023; Kim et al., 2023), which extends DSTC9 to multi-document settings, and FaithDial (Dziri et al., 2022a), which is a dehallucinated subset of WoW. Statistics are shown in Appendix B.2.

4.2 Baselines

CTRL (Keskar et al., 2019) introduces a sequence of control tokens $\mathbf{c}$ to steer the model towards desirable generations:

$$p_{\boldsymbol{\theta}}(\mathbf{u}_{T+1} \mid \mathbf{u}_1^T, \hat{\mathcal{K}}, \mathbf{c}). \tag{10}$$

Rashkin et al. (2021) adapt the model in Equation (10) to promote faithfulness in document-grounded dialogue by introducing *entailment*, *lexical overlap* and *first-person* tokens. We employ the first two. Entailment indicates whether the response is entailed by the documents, determined by an MNLI model, and lexical overlap splits the responses into three buckets according to low,

Model	WoW$_{\text{seen}}$ BLEU(↑) $(\mathbf{y}, \hat{\mathbf{y}})$	Critic(↓)	Q^2(↑)	BERT(↑) $(\mathbf{y}, \hat{\mathcal{K}})$	F1(↑)	Dens.(↓)	DSTC9 BLEU(↑) $(\mathbf{y}, \hat{y})$	Critic(↓)	Q^2(↑)	BERT(↑) $(\mathbf{y}, \hat{\mathcal{K}})$	F1(↑)	Dens.(↓)
Flan-T5	18.5	24.3	76.2	84.4	78.6	12.4	18.5	6.2	62.3	61.3	45.2	1.73
+ TA	19.1	19.4	75.9	82.2	74.4	11.1	18.5	2.5	79.6	63.6	53.9	2.80
+ EWR	18.1 (↓-0.4)	18.1 (↓-6.2)	78.0 (↑1.8)	86.2 (↑1.8)	80.8 (↑2.2)	13.5 (↑1.1)	**20.0** (↑1.5)	4.3 (↓-1.9)	78.4 (↑16.1)	64.4 (↑3.1)	55.6 (↑10.4)	3.22 (↑1.49)
CaPE	18.8	13.2	78.2	83.7	75.9	11.2	17.3	2.3	72.5	63.3	52.6	2.63
+ EWR	19.0 (↑0.2)	9.4 (↓-3.8)	78.7 (↑0.5)	88.2 (↑4.5)	83.0 (↑7.1)	13.6 (↑2.4)	16.7 (↓-0.6)	2.6 (↑0.3)	79.2 (↑6.7)	64.3 (↑1.0)	54.0 (↑1.4)	2.76 (↑0.13)
CTRL	**19.5**	10.3	83.9	87.8	82.3	13.9	17.6	5.3	79.8	64.5	57.8	3.30
+ TA	19.3	8.9	82.7	87.0	81.2	13.0	18.0	**1.2**	**89.5**	**66.5**	**63.6**	4.53
+ EWR	18.4 (↓-0.8)	**5.7** (↓-4.6)	86.8 (↑2.9)	91.3 (↑3.5)	87.7 (↑5.4)	16.3 (↑2.4)	19.4 (↑1.7)	2.3 (↓-3.0)	85.3 (↑5.5)	65.5 (↑1.0)	60.6 (↑2.8)	3.80 (↑0.5)
DExperts	18.0	14.8	79.6	87.0	82.2	14.3	17.1	2.9	74.9	63.6	55.7	2.83
Quark	17.2	7.9	**91.9**	**92.6**	**90.2**	18.6	19.0	5.7	73.1	62.7	49.8	2.03
Noisy Channel	18.4	24.0	78.6	85.0	79.8	13.1	18.6	5.1	67.1	62.7	48.4	2.18

Table 1: Main results on WoW$_{\text{seen}}$ and DSTC9 indicating: i) performance in dialogue generation comparing true $\hat{\mathbf{y}}$ and predicted $\mathbf{y}$ responses (BLEU); ii) faithfulness of predicted response y to ground-truth knowledge $\hat{\mathcal{K}}$ (Critic, Q^2, BERT, F1); 3) abstractiveness (Dens.). We report several baselines adapted for faithful generation and show how Task Arithmetic (TA) and Elastic Weight Removal (EWR, ours) can be deployed on top of vanilla pre-trained models, like Flan-T5, or on top of other methods like CTRL. Relative improvements and degradations are indicated in green and red, respectively.

medium, and high lexical overlap. CTRL is trained on examples from all three buckets and both entailment labels but only conditioned on desired ones at inference time (high-overlap and entailment).

Quark (Lu et al., 2022a) uses a similar strategy as CTRL for unlearning. The difference is that not only the original training data but also model generations which are taken after each epoch are augmented with special tokens and used for training. Noting this similarity to CTRL, we therefore employ the same tokens to adapt it to faithful dialog generation, allowing for a direct comparison.

DExperts (Liu et al., 2021) makes use of an expert and anti-expert model in order to reduce toxicity. The expert model is trained to generate non-toxic text and the anti-expert to generate toxic text. However, instead of combining models in parameter space, as in our method, they are combined at inference time as a density ratio:

$$p(\mathbf{u}_{T+1} \mid \mathbf{u}_1^T, \hat{\mathcal{K}}) \propto \quad (11)$$
$$p_{\boldsymbol{\theta}_0}(\mathbf{u}_{T+1} \mid \mathbf{u}_1^T, \hat{\mathcal{K}}) \cdot \frac{p_{\boldsymbol{\theta}_{\text{E}}}(\mathbf{u}_{T+1} \mid \mathbf{u}_1^T, \hat{\mathcal{K}})}{p_{\boldsymbol{\theta}_{\text{AE}}}(\mathbf{u}_{T+1} \mid \mathbf{u}_1^T, \hat{\mathcal{K}})}.$$

Tokens with high expert probability are encouraged and tokens with high anti-expert probability are discouraged. We use the same expert and anti-expert models as in CaPE to adapt it to faithful dialog generation and fairly compare both methods.

Noisy Channel Model (Daheim et al., 2022) introduce a noisy channel model for document-grounded dialogue:

$$p(\mathbf{u}_{T+1} \mid \mathbf{u}_1^T, \hat{\mathcal{K}}) \propto \quad (12)$$
$$p_{\boldsymbol{\theta}_1}(\hat{\mathcal{K}} \mid \mathbf{u}_1^T, \mathbf{u}_{T+1}) \cdot p_{\boldsymbol{\theta}_2}(\mathbf{u}_{T+1} \mid \mathbf{u}_1^T).$$

Here, $p_{\boldsymbol{\theta}_1}(\hat{\mathcal{K}} \mid \mathbf{u}_1^T, \mathbf{u}_{T+1})$ can be seen as a faithfulness and $p_{\boldsymbol{\theta}_2}(\mathbf{u}_{T+1} \mid \mathbf{u}_1^T)$ as a fluency expert. We use their reranking method to rescore generations obtained from our baseline model.

4.3 Metrics

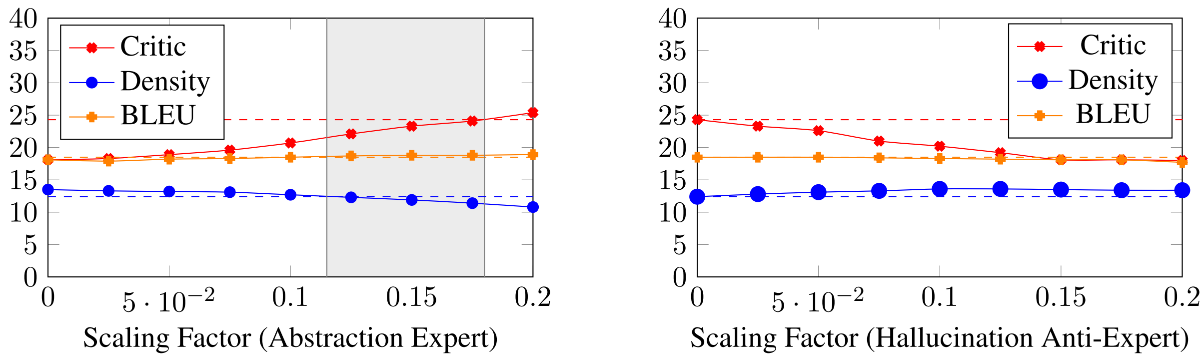
We measure the lexical similarity of the generated and the ground-truth responses with the sacrebleu (Post, 2018) implementation of BLEU (Papineni et al., 2002). To evaluate faithfulness, we employ the hallucination critic introduced by Dziri et al. (2022a)[1], which classifies responses as hallucinated or not, Q^2 (Honovich et al., 2021), which uses a question generation and question answering pipeline, as well as token-level F1 and BERTScore (Zhang* et al., 2020)[2]. To measure abstractiveness, we again use Density (Grusky et al., 2018). Further details are found in Appendix B.3.

5 Results

We first introduce our main results on WoW and DSTC9 in Section 5.1. Then, we characterise trade-offs between faithfulness and abstractiveness in Section 5.2 before discussing the controllability of model interpolation in Section 5.3. Finally, we discuss ablations on various datasets in Section 5.4 and report human evaluation results in Section 6.

[1] https://huggingface.co/McGill-NLP/roberta-large-faithcritic.

[2] We use the *deberta-large-mnli* checkpoint.



(a) Faithfulness–Abstractiveness Trade-Off

(b) Faithfulness–Performance Trade-Off

Figure 2: Metrics for EWR with Flan-T5$_{\text{base}}$ on WoW$_{\text{seen}}$. (a) Faithfulness and abstractiveness can be traded-off by varying both the influence of the abstractivness expert (a) and hallucination anti-expert (b).

5.1 Main Results on Faithfulness

We start with results for de-hallucinated models using Flan-T5$_{\text{base}}$ in Table 1. Results with Flan-T5$_{\text{large}}$ are found in the Appendix C.1 and show a similar trend: subtracting anti-experts from various base models can improve faithfulness at minor degradation in other metrics. Increases in faithfulness from EWR are often stronger than from task arithmetic, except for Flan-T5$_{\text{base}}$ on DSTC9, especially in terms of BERT and token-level F1, but can also lead to decreased BLEU. EWR on top of CTRL provides state-of-the-art performance in faithfulness, comparable to strong baselines like Quark. While the additional faithfulness expert used in CaPE generally improves over using only an anti-expert, we observe fast degradation in terms of BLEU and BertScore on DSTC9, potentially stemming from comparatively small amounts of expert training data after partitioning the dataset.

CTRL and Quark confirm the effectiveness of control tokens and iteratively applying them to model generations during training. DExperts and noisy channel reranking are mostly outperformed by EWR, task arithmetic, and CaPE, except for Flan-T5$_{\text{base}}$ on WoW. This is notable, as they require keeping multiple models but all others use just one at inference time. Nevertheless, the performance of noisy channel model reranking increases with beam size (Daheim et al., 2022) which we keep identical for all methods.

Improvements of CTRL and Quark are much more conspicuous in WoW than DSTC9. We attribute this to the fact that in DSTC9, the ground-truth documents are FAQs, in which the question might not be as important for the control tokens. Furthermore, gold responses contain follow-up

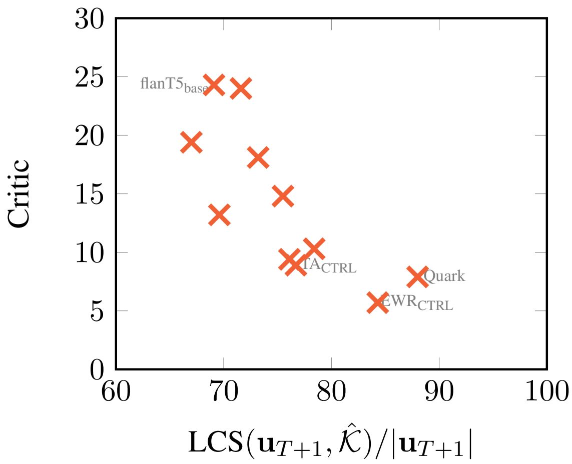


Figure 3: Improvements in faithfulness (Critic) tend to incur an increase in extractiveness (LCS) on WoW.

questions at every turn, which might decrease the effectiveness of the special tokens and might affect automatic metrics.

Nevertheless, our results in Table 1 also illustrate that increased faithfulness comes at the cost of increased extractiveness, as measured by Density. We investigate this further in the next subsection.

5.2 Faithfulness–Abstractiveness Trade-Off

As our main experiments show that improvements in faithfulness also increase extractiveness, we now outline experiments using an additional abstractiveness expert to reduce this effect. Figure 2 a highlights our results on WoW using Flan-T5$_{\text{base}}$, when only varying the scaling factor of the abstraction expert. From the plot, it emerges that we can control the trade-off between faithfulness and abstractiveness to improve over the baseline in both dimensions, in the interval indicated by the greyed

Model	BLEU(↑) $(y, \hat{y})$	Critic(↓)	Q^2(↑)	BF1(↑) $(y, \hat{\mathcal{K}})$	F1(↑)
			WoW$_{\text{unseen}}$		
Flan-T5$_{\text{base}}$	18.1	22.7	74.0	84.8	78.7
+ TA	**18.8**	19.2	75.7	82.8	75.0
+ EWR	17.4 (↓-0.7)	**17.7** (↓-5.0)	**78.4** (↑4.4)	**86.9** (↑2.1)	**81.6** (↑2.9)
			DSTC11		
Flan-T5$_{\text{base}}$	7.9	76.6	49.7	54.6	37.1
+ TA	8.0	60.0	51.0	59.9	**43.6**
+ EWR	**9.6** (↑1.7)	**41.1** (↓35.5)	**57.3** (↑7.6)	**60.0** (↑5.4)	38.6 (↑1.5)
			FaithDial		
Flan-T5$_{\text{base}}$	15.1	0.3	**66.4**	80.9	73.7
+ TA	**15.3**	**0.1**	57.5	77.3	67.6
+ EWR	14.9 (↓-0.2)	**0.1** (↓-0.2)	**66.4** (-0.0)	**81.7** (↑0.8)	**75.0** (↑1.3)

Table 2: EWR improves faithfulness on unseen topics (WoW$_{\text{unseen}}$), multi-document corpora (DSTC11), and datasets with cleaned ground-truth annotations (FaithDial).

area. To further quantify this trade-off, which has also been described in related works (Dziri et al., 2022a; Daheim et al., 2022; Aksitov et al., 2023), we use the ratio of the length of the longest common subsequence between $\mathbf{u}_{T+1}$ and $\hat{\mathcal{K}}$ and the length of $\mathbf{u}_{T+1}$ (LCS). We plot the dependency of LCS and Critic in Figure 3 for Flan-T5$_{\text{base}}$-based models on WoW. There is a clear trend towards more extractiveness with increased faithfulness but a better Critic score does not always imply an increase in LCS.

5.3 Scaling Factors & Controllability

Next, we assess how much control EWR provides over faithfulness scores within an acceptable range of BLEU, which measures overall performance. Figure 2 b highlights that there is a larger region of factors along which faithfulness constantly improves within a narrow range of BLEU scores. However, corresponding to the previously discussed trade-off, density increases with faithfulness, indicating that the scaling factor also controls how much of the knowledge is copied into the response.

5.4 Generalisation to Additional Datasets

In this section, we study the performance of EWR in challenging settings, namely on: i) unseen topics that require generalisation (WoW unseen), ii) multi-document corpora (DSTC11), and iii) cleaned training and test data that does not contain hallucinations in ground-truth annotations (FaithDial). We report the results in Table 2.

In summary, we observe the following: 1) EWR shows improvements in all settings, especially in terms of generalisation and in a multi-document setting. Furthermore, we can even improve faithfulness metrics when training and evaluating on the cleaned FaithDial dataset. 2) task arithmetic can improve results on multi-document corpora and some metrics on the unseen set but fails to improve BERT F1 and F1 on WoW unseen and FaithDial.

Model	WoW A (↑)	WoW C (↑)	WoW P (↑)	DSTC9 A (↑)	DSTC9 C (↑)	DSTC9 P (↑)
Flan-T5$_{\text{base}}$	72.3	1.74	1.19	89.7	2.83	1.71
+ EWR$_{\text{abs}}$	75.1	1.62	1.25	94.7*	2.41	1.49
CTRL	85.5*	1.58	1.12	94.7*	2.72	1.42
+ TA	88.8*	1.58	1.16	97.0*	2.63	1.40
+ EWR	96.8†	1.50	1.08	98.0†	2.50	1.36
Quark	93.1†	1.51	1.05	86.0	2.89	1.66

Table 3: Human evaluation on 218 examples annotated by 3 expert annotators each. We measure attributability (A), Co-cooperativeness (C), and paraphrasing (P). * indicates significance wrt. Flan-T5$_{\text{base}}$ and † wrt. to the next best method with $p < 0.05$.

6 Human Evaluation

In addition to the automatic evaluation, we conduct a human evaluation on WoW and DSTC9 with the help of three expert annotators [3], using the Attributable to Identified Source (AIS) framework (Rashkin et al., 2023). First, we ask them to score responses as attributable (A) only if all their content can be attributed to the knowledge that grounds the dialogue response. Furthermore, we ask annotators to rate cooperativeness (C), i.e. the ability of the model to connect with and follow up on user turns on a 3-point Likert scale. Here, 1 indicates a response that does not cooperate with the dialogue, 2 a response that brings the dialogue forward, and 3 a response that acknowledges the previous utterances and responds with a follow-up question. Lastly, annotators rate paraphrasing (P) on a binary scale, where 2 indicates non-trivial paraphrasing of the knowledge and 1 substantial copying. Detailed instructions can be found in Appendix B.4.

Table 3 shows the results for the A, C, and P categories with agreements of $0.61, 0.51, 0.53$, respectively, in terms of Fleiss' κ. Generally, we observe that human evaluation results for attributability confirm results based on automatic faithfulness metrics as they display similar patterns. In particular, all methods improve over vanilla Flan-T5, with CTRL and Quark performing similarly on average and outperforming each other on the two different datasets. Task arithmetic and EWR give improvements over

[3] All annotators are graduate students in NLP and paid above minimum wage.

CTRL on both datasets. Most notably, $\mathrm{EWR}_{\mathrm{CTRL}}$ improves over all other methods, including task arithmetic and Quark, by a statistically significant margin in human evaluation.

Our results also emphasize the trade-off between faithfulness and both paraphrasing (which reflects abstractiveness) and cooperativeness. Increased attributability often leads to a decrease in both other criteria. Nevertheless, EWR with a faithfulness anti-expert and an abstraction expert, labelled $\mathrm{EWR}_{\mathrm{abs}}$, improves both paraphrasing and attributability on WoW and attributability on both datasets compared to vanilla Flan-T5. While $\mathrm{EWR}_{\mathrm{abs}}$ does not outperform this baseline in paraphrasing on DSTC9, we believe that this stems from the way the expert dataset $\mathcal{D}^{\mathrm{E}}$ is constructed, related to the comparatively less strong performance of Quark and CTRL. As the ground-truth responses in DSTC9 contain longer follow-up questions, it is likely that density-based binning does not pick up nuances, such as the difference between non-paraphrased responses and follow-up questions independent from the knowledge.

7 Related Work

Hallucination in LMs The impressive abilities of LMs are offset by the potential for generating hallucinated text (Ji et al., 2022; Thoppilan et al., 2022; Bang et al., 2023; Qin et al., 2023; Choi et al., 2023), which sparked an increasing interest in tackling this problem in the context of grounded language generation (Ji et al., 2022), encompassing several tasks such as data-to-text generation (Wiseman et al., 2017; Parikh et al., 2020), machine translation (Wang and Sennrich, 2020; Raunak et al., 2021), summarisation (Durmus et al., 2020; Kang and Hashimoto, 2020), generative question answering (Li et al., 2021), and dialogue generation (Dziri et al., 2021, 2022c; Rashkin et al., 2021; Ji et al., 2022; Razumovskaia et al., 2022). Different studies aim to address the issue of hallucination by either developing automatic metrics to detect it (Wiseman et al., 2017), or by identifying potential causes, such as out-of-domain generalisation, noisy training data, and exposure bias (Kang and Hashimoto, 2020; Raunak et al., 2021; Wang and Sennrich, 2020; Dziri et al., 2021).

For neural dialogue models it has been shown that retrieving relevant knowledge can reduce –but not completely eliminate– hallucinations (Shuster et al., 2021). Therefore, different methods have been proposed to tackle it, such as token-level critics (Dziri et al., 2021), or control token- (Rashkin et al., 2021) and reranking-based methods (Daheim et al., 2022). Lastly, as hallucinations in training data can greatly exacerbate those in models (Dziri et al., 2022b), a hallucination-free dialogue benchmark has been proposed (Dziri et al., 2022a).

Controllable text generation Different works steer model behaviour by controlled generation, for example by combining models at decoding time (Liu et al., 2021) or in parameter space (Ilharco et al., 2023), conditioning on reward tokens assigned to model generations in training (Lu et al., 2022a) or the initial training data (Keskar et al., 2019; Niu and Bansal, 2018). Finally, different methods constrain text generation with logical constraints (Lu et al., 2021, 2022b) or by forcing specific words to appear (Pascual et al., 2021).

8 Conclusion & Future Work

We introduce Elastic Weight Removal (EWR), a novel method for steering the behaviour of language generation models by combining their parameters with those of (anti-)experts, weighted by Fisher Information. We show how EWR can be used to reduce hallucinations in document-grounded dialogue response generation across different settings. We compare it to other state-of-the-art methods, many of which we adapt to faithful response generation for the first time. Automated metrics and human evaluation show that EWR improves faithfulness over multiple baselines, and can furthermore provide complementary improvements with them. Moreover, we show that faithfulness comes at the expense of abstraction. Therefore, we combine an abstraction expert with the hallucination anti-expert to promote responses that are both more faithful and abstractive than the baseline.

The main contribution of this work is that it outlines an unexplored way of promoting faithfulness in document-grounded dialogue by using experts and anti-experts not at inference time—and thereby incurring significant overhead—but rather to navigate the parameter space towards an improved set of parameters without altering the model architecture. This opens up many potential areas for future work, such as controlling for further dimensions, or developing more sophisticated data augmentation techniques to create data for (anti-)experts.

9 Limitations

One limitation of our work is that we assume the ground-truth knowledge $\hat{\mathcal{K}}$ to be given. This assumption does not hold in general, when a dialogue system is used, because for a new user query it is unknown. We might then expect that our method stays more faithful to the retrieved knowledge, too, but could generate erroneous responses to the user query if this knowledge is incorrect.

A further limitation is the scale at which we conduct experiments, which do not go beyond 1B parameters due to the large number of baselines that we evaluate on multiple corpora. On the other hand, models used in production are often significantly larger, often having tens of billions of parameters.

Connected to this, many of such models are now trained using parameter-efficient finetuning techniques, which either introduce a new subset of model parameters that are trained, while all existing ones are kept fixed, or train a subset of model parameters. Our method should be amenable to this setting, because the task vector will also be 0 for parameters that are not trained. However, we did not experiment using parameter-efficient finetuning techniques in this work.

Finally, we only evaluate a small set of (data augmentation) techniques for creating hallucinated and abstractive data and future work could evaluate more such methods.

While we only study english datasets, we expect the techniques to be similarly applicable for other languages.

Ethics and Broader Impact Statement

Our work relies on LLMs to generate responses in dialogue. Since such LLMs are prone to producing errors, it can not be guaranteed that our methods also do not produce erroneous outputs, such as hallucinations, or output toxic or biased data. However, this work aims to mitigate hallucinations and therefore we think that there is no direct ethical concern.

References

Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yun-Hsuan Sung. 2023. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models. *CoRR*, abs/2302.05578.

Shun-ichi Amari. 1998. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. 2023. ChatGPT goes to law school. *Available at SSRN*.

Prafulla Kumar Choubey, Alexander R. Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Fatema Rajani. 2021. CaPE: Contrastive parameter ensembling for reducing hallucination in abstractive summarization.

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.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

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.

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.

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.

Yingbo Gao, Christian Herold, Zijian Yang, and Hermann Ney. 2022. Revisiting checkpoint averaging for neural machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 188–196, Online only. Association for Computational Linguistics.

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.

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784, Toronto, Canada. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858.

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.

Seokhwan Kim, Spandana Gella, Chao Zhao, Di Jin, Alexandros Papangelis, Behnam Hedayatnia, Yang Liu, and Dilek Z Hakkani-Tür. 2023. Task-oriented conversational modeling with subjective knowledge track in dstc11. In *Proceedings of The Eleventh Dialog System Technology Challenge*, pages 274–281.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. 2019. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, 32.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chenliang Li, Bin Bi, Ming Yan, Wei Wang, and Songfang Huang. 2021. Addressing semantic drift in generative question answering with auxiliary extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 942–947.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. 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 6691–6706, Online. Association for Computational Linguistics.

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.

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

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022a. QUARK: Controllable text generation with reinforced unlearning. In *Advances in Neural Information Processing Systems*.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, Noah A. Smith, and Yejin Choi. 2022b. NeuroLogic a*esque decoding: Constrained text generation with lookahead heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 780–799, Seattle, United States. Association for Computational Linguistics.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

James Martens. 2020. New insights and perspectives on the natural gradient method. 21(1):5776–5851.

Michael S Matena and Colin Raffel. 2022. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier.

Tong Niu and Mohit Bansal. 2018. Polite Dialogue Generation Without Parallel Data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jan-Thorsten Peter, Eugen Beck, and Hermann Ney. 2018. Sisyphus, a workflow manager designed for machine translation and automatic speech recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 84–89, Brussels, Belgium. Association for Computational Linguistics.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. 2023. Modular deep learning.

Edoardo M. Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021. Parameter space factorization for zero-shot learning across tasks and languages. *Transactions of the Association for Computational Linguistics*, 9:410–428.

Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019. Towards zero-shot language modeling. 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 2900–2910, Hong Kong, China. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver?

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, pages 1–64.

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.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Evgeniia Razumovskaia, Ivan Vulić, and Anna Korhonen. 2022. Data augmentation and learned layer aggregation for improved multilingual language understanding in dialogue. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2017–2033, Dublin, Ireland. Association for Computational Linguistics.

Nicol N Schraudolph. 2002. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738.

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.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552, Online. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. 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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: evaluating text generation with BERT. In *International Conference on Learning Representations*.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023. "what do others think?": Task-oriented conversational modeling with subjective knowledge. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 309–323, Prague, Czechia. Association for Computational Linguistics.

Appendix

A Details on Method

A.1 Pseudocode

Algorithm 1 outlines the steps for using EWR to reduce hallucinations while promoting abstractiveness. Concretely, a dialogue response generation

Algorithm 1 Pseudocode for removing hallucinations and promoting abstraction with EWR. Note that we apply $(\cdot)^2$ element-wise.

Input Dialogues $\mathcal{D}$, hallucinated anti-expert dataset $\mathcal{D}^{\mathrm{AE}}$, abstractive expert dataset $\mathcal{D}^{\mathrm{E}}$, initial parameter set $\boldsymbol{\theta}$
Output $\boldsymbol{\theta}'$

$\boldsymbol{\theta}_0 \leftarrow \text{finetune}(\boldsymbol{\theta}, \mathcal{D})$
$\boldsymbol{\theta}_{\mathrm{AE}} \leftarrow \text{finetune}(\boldsymbol{\theta}_0, \mathcal{D}^{\mathrm{AE}})$
$\boldsymbol{\tau}_{\mathrm{AE}} \leftarrow \boldsymbol{\theta}_{\mathrm{AE}} - \boldsymbol{\theta}_0$
$\boldsymbol{\theta}_{\mathrm{E}} \leftarrow \text{finetune}(\boldsymbol{\theta}_0, \mathcal{D}^{\mathrm{E}})$
$\boldsymbol{\tau}_{\mathrm{AE}} \leftarrow \boldsymbol{\theta}_{\mathrm{E}} - \boldsymbol{\theta}_0$
$\mathbf{f}_{\boldsymbol{\theta}_0} \leftarrow \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} (\nabla \log p_{\boldsymbol{\theta}_0}(u_{T+1} \mid u_1^T, \hat{\mathcal{K}}))^2$
$\mathbf{f}_{\boldsymbol{\tau}_{\mathrm{AE}}} \leftarrow \frac{1}{|\mathcal{D}^{\mathrm{AE}}|} \sum_{\mathcal{D}^{\mathrm{AE}}} (\nabla \log p_{\boldsymbol{\tau}_{\mathrm{AE}}}(u_{T+1} \mid u_1^T, \hat{\mathcal{K}}))^2$
$\mathbf{f}_{\boldsymbol{\tau}_{\mathrm{E}}} \leftarrow \frac{1}{|\mathcal{D}^{\mathrm{E}}|} \sum_{\mathcal{D}^{\mathrm{E}}} (\nabla \log p_{\boldsymbol{\tau}_{\mathrm{E}}}(u_{T+1} \mid u_1^T, \hat{\mathcal{K}}))^2$
$\boldsymbol{\theta}' \leftarrow \frac{\lambda_0 \cdot \mathbf{f}_{\boldsymbol{\theta}_0} \cdot \boldsymbol{\theta}_0 - \lambda_{\mathrm{AE}} \cdot \mathbf{f}_{\boldsymbol{\tau}_{\mathrm{AE}}} \cdot \boldsymbol{\tau}_{\mathrm{AE}} + \lambda_{\mathrm{E}} \cdot \mathbf{f}_{\boldsymbol{\tau}_{\mathrm{E}}} \cdot \boldsymbol{\tau}_{\mathrm{E}}}{Z}$

model is trained first. Then, an anti-expert and expert model are trained on hallucinated and abstractive (and not hallucinated) data, respectively. Both models are the subtracted and added to the dialogue response generation model, respectively, but weighted by Fisher information. The Fisher information is estimated by its diagonal with a squared gradient approximation over the training, where labels are sampled. We have found that calculating this by parameterising the model with the task vectors $\boldsymbol{\tau}_{\mathrm{AE}}$ and $\boldsymbol{\tau}_{\mathrm{E}}$ performs empirically well, but it is theoretically better motivated to calculate it at the anti-expert $\boldsymbol{\theta}_{\mathrm{AE}}$ or expert $\boldsymbol{\theta}_{\mathrm{E}}$, respectively. Both strategies provided similar performance in our experience.

B Details on Experiments & Evaluation

B.1 Further Experimental Details

All models, with the exception of Quark and (anti-)experts, which we train for 5 epochs, are trained for 10 epochs using an initial learning rate of $6.25e{-5}$, linear learning rate decay without warmup, and a batch size of 32, following prior work (Daheim et al., 2022). We take checkpoints after each epoch and pick the one with smallest validation loss. For Task Arithmetic and EWR we do a grid search to determine the scaling factors on a validation set on WoW, FaithDial, and DSTC9. For DSTC11 we did not perform such a grid set because we only had a validation but not a test set, and the hyperparameters seemed to be consistent across datasets. We chose 1.0 for Task Arithmetic and 0.15 for EWR for all experiments with only a hallucination anti-expert, since these factors performed best. We use Flan-T5$_{\mathrm{base}}$ and Flan-T5$_{\mathrm{large}}$ with 250M and 780M parameters, respectively. We use the checkpoints that are available on the huggingface hub under `https://huggingface.co/google/flan-t5-base` and `https://huggingface.co/google/flan-t5-large`. All experiments are performed on NVIDIA A100 or V100 GPUs and each model takes at most half a day to finetune.

All code for reproducing the experiments will be made publicly available in a comprehensive software repository under Apache License 2.0 [4].

Dataset	#train	#val	#test
WoW (Dinan et al., 2019)	83247	4444	{4356, 4380}
DSTC9 (Kim et al., 2020)	19184	2673	1981
FaithDial (Dziri et al., 2022a)	18357	3417	3539
DSTC11 (Zhao et al., 2023)	14768	2129	-

Table 4: Dataset statistics showing the number of train, validation, and test examples counted in number of utterances. For WoW test, we first show the seen and then unseen split in curly brackets. For DSTC11, the test set was not available yet at the time of writing.

B.2 Further Details on Datasets

In this section we provide details on the splits of all used datasets. The statistics are shown in Table 4. For Wizard-of-Wikipedia, we have used the train, dev and both test splits (seen and unseen). For DSTC11 we have only used validation split, because the test set was not yet available at the time of our experiments.

For the hallucination anti-expert model, the training data is exactly the same size as for the document-grounded response generation model, just with the knowledge switched out. For all expert models we subsample the data according to the assigned control tokens which depend on the used metric and NLI model.

All datasets are in English and might therefore represent predominantly the demographics of english-speaking countries. WoW was collected by crowdsourcing dialogues in a roleplaying game. DSTC9 was collected by asking crowdworkers to fill in dialogues from MultiWoZ 2.1 (Eric et al., 2020). DSTC11 was collected using crowdworkers on Amazon MTurk, who stem from the USA, Canada, and Great Britain (Zhao et al., 2023). Finally, FaithDial was created by asking crowdwork-

[4] `https://www.apache.org/licenses/LICENSE-2.0`

ers, also on Amazon MTurk, to clean dialogues from WoW (Dziri et al., 2022a).

B.3 Further Details on Used Metrics

We evaluate BLEU (Papineni et al., 2002) on the corpus-level using the sacrebleu package (Post, 2018). Other metrics are calculated on an example-level and averaged to obtain a global score. Concretely, for critic model taken from Dziri et al. (2022a), this means that we classify each utterance as hallucination or not, with 1 indicating hallucination and 0 otherwise. The score is averaged over these classifications and can therefore be seen as calculating the percentage of hallucinated examples in the model predictions. The model used for this is finetuned from RoBERTA (Liu et al., 2019) and released as part of Dziri et al. (2022a). It is openly available on the huggingface hub and can be found under `https://huggingface.co/McGill-NLP/roberta-large-faithcritic`. For Q^2 (Honovich et al., 2021), a pipeline of steps is performed for each generated example to arrive at a score. First, answer candidates are determined for the generated response, which often correspond to spans of entities. Then, questions are generated for each answer candidate and answered based on the knowledge documents. If the answer is the same by string match, a score of 1 is assigned. If there is no string match, a score of 1 is assigned if an NLI model judges one answer to entail the other, and a score of 0 otherwise. Questions are also filtered, and if no valid question is found, entailment between the knowledge and the generated response is calculated as a fallback. We base our implementation on the open-source implementation found in `https://github.com/orhonovich/q-squared` which was released with Honovich et al. (2021) and will open-source our reproduction under Apache License 2.0.

Our adoption of density (Grusky et al., 2018) calculates the average squared length of extractive spans that were copied from the knowledge documents into the generated response. We average the densities of all predictions. Similarly, F1 calculates the token-level overlap between generated response and document, and we again take the average over predictions. Again, all the implementations of these metrics will be made publicly available by us.

For BertScore (Sun et al., 2022), we use the open-source implementation found at `https://github.com/Tiiiger/bert_score` and use the 'deberta-large-mnli' checkpoint, which was recommended at the time of implementation.

B.4 Details on Human Evaluation

In this section, we detail the instructions and recruitment for our human evaluation. All of the annotators are graduate students in NLP from one of the authoring institutions and are all paid well above minimum-wage. All annotators voluntarily agreed to participating in our study and were informed, and agree to, that no personal data would be released and only the human judgements would be stored. The annotators were instructed to score 218 randomly sampled examples generated with different models from WoW and DSTC9 according to three criteria: Faithfulness, Coherence, and Paraphrasing, abbreviated with F, C, and P, respectively, in Table 3. The instructions for Faithfulness follow the well-established Attributable to Identified source framework (AIS) (Rashkin et al., 2023). We follow the exact definitions from their work and show these as guidelines to the annotators, who were instructed to carefully read the paper. This is feasible, because all annotators have graduate-level knowledge of NLP. Following the frame work, we instructed users to only annotate interpretable responses, others were to be left out. Then, a score of one should be assigned if the conditions in (Rashkin et al., 2023, Definition 8) are met. We repeat the definition here verbatim for completeness and refer the reader to their work for more information about the framework.

Definition 1. *AIS, full definition (Rashkin et al., 2021) A pair (s, c), where s is a sentence and c_l, t is a pair consisting of a linguistic context and a time, is Attributable to Identified Sources (AIS) iff the following conditions hold:*

1. *The systems provides a set of parts P of some underlying corpus K, along with S.*
2. *s in the context c is interpretable (i.e., $E(c, s) \neq NULL$.*
3. *The explicature $E(c, s)$ is a standalone proposition.*
4. *The pair $(E(c, s), t)$ is attributable to P.*

The pair $E(c, s), t)$ is attributable to a set of parts P of some underlying corpus K iff: A generic *hearer will, with a chosen level of confidence, affirm the following statement: "According to P, $E(c, s)$, where $E(c, s)$ is interpreted relative to time t."*

According to this, a binary label is assigned, where 1 indicates 'faithful' and 0 'not faithful'. We only make a slight change in definition for DSTC9, where the FAQ documents are short and give relevant information to a customer in customer service conversations, for example, for hotel booking. The change is as follows: "If important information for the user in K is left out, the response should be scored as 'not faithful'."

For Coherence, we ask the annotators to only score such responses that were annotated with 1 in the previous step on a 3-point Likert scale. The instructions are as follows:

3: The response is highly co-operative and, for example, explicitely acknowledges the previous turn (e.g. ""Yes,..."".) and contains a follow-up question.

2: The response follows up logically to the previous dialog and / or shows some degree of co-operativeness.

1: The response is standalone and does not follow-up logically to the previous dialog.

Here, the listing item (e.g. "3:") indicates the rating.

For Paraphrasing, we chose a two-point scale with the following instructions:

2: Response paraphrases the evidence to a sufficient extent.

1: The response copy-pastes the evidence into the response verbatim or almost verbatim.

As noted in Section 6, we achieve agreements of $0.61, 0.51, 0.53$, respectively, in terms of Fleiss' κ, for the three categories above in order of writing.

C Further Results

C.1 Additional Experiments Using Flan-T5$_{\text{large}}$

Table 5 shows results obtained using the same setup as in Section 5.1 but using Flan-T5$_{\text{large}}$ instead of Flan-T5$_{\text{base}}$. We find the results from the smaller checkpoint to be confirmed and find much larger improvements for EWR on DSTC9 than using the base checkpoint. Again, parameter interpolation methods can be used effectively to reduce hallucinations at minor costs of fluency and abstractiveness, also on top of other methods that promote faithfulness. However, we find CTRL and Quark less effective for DSTC9, potentially because the overlap and entailment tokens have more errors than in WoW due to the structure of the used FAQ documents.

Model	WoW$_{\text{seen}}$ BLEU(↑) $(y, \hat{y})$	Critic(↓)	Q^2(↑)	BERT(↑) $(y, \hat{\mathcal{K}})$	F1(↑)	Dens.(↓)	DSTC9 BLEU(↑) $(y, \hat{y})$	Critic(↓)	Q^2(↑)	BERT(↑) $(y, \hat{\mathcal{K}})$	F1(↑)	Dens.(↓)
Flan-T5$_{\text{large}}$	18.6	26.7	77.8	83.8	77.5	12.3	18.6	6.9	64.0	61.2	44.7	1.81
+ TA	19.1	16.7	80.2	84.6	77.8	12.6	19.0	3.7	74.3	64.4	55.6	3.50
+ EWR	17.3 (↓-1.3)	16.9 (↓-9.8)	80.3 (↑2.5)	88.3 (↑4.5)	83.9 (↑6.4)	14.9 (↑2.6)	19.1 (↑0.5)	2.8 (↓-4.1)	83.8 (↑19.8)	64.8 (↑3.6)	57.3 (↑12.6)	3.48 (↑1.67)
CaPE	19.0	13.0	79.5	83.7	75.4	11.3	17.2	4.3	73.3	64.4	53.2	2.82
+ EWR	18.2 (↓-0.8)	9.3 (↓-3.7)	80.4 (↑0.9)	89.4 (↑5.7)	84.9 (↑9.5)	15.2 (↑3.9)	16.2 (↓-1.0)	1.1 (↓-3.2)	74.9 (↑1.6)	64.1 (↓-0.3)	54.1 (↑0.9)	3.00 (↑0.18)
CTRL	**19.8**	11.3	82.0	87.3	81.5	13.4	**19.5**	6.8	77.4	63.8	52.7	2.73
+ TA	19.2	7.2	84.3	86.8	80.6	13.0	19.3	2.6	79.3	**65.9**	57.5	3.37
+ EWR	18.6 (↓-1.2)	**7.0** (↓-4.3)	85.8 (↑5.4)	90.5 (↑3.2)	86.8 (↑5.3)	16.8 (↑3.4)	18.1 (↓-1.4)	**0.8** (↓-6.0)	**84.3** (↑6.9)	65.2 (↑1.4)	**59.5** (↑6.8)	3.83 (↑1.1)
DExperts	18.3	17.9	79.8	81.7	71.4	12.7	18.2	4.2	70.5	63.9	54.9	2.78
Quark	18.0	9.1	**91.4**	**91.2**	**88.1**	16.9	20.3	6.0	74.7	64.9	54.3	3.09
Noisy Channel	18.8	22.3	77.2	85.5	80.2	13.3	18.4	6.1	67.2	62.2	47.4	2.20

Table 5: Main results on WoW$_{\text{seen}}$ and DSTC9 using Flan-T5$_{\text{large}}$.