CaPE: Contrastive Parameter Ensembling for Reducing Hallucination in Abstractive Summarization

Prafulla Kumar Choubey[1] **Alexander R. Fabbri**[1] **Jesse Vig**[1]
Chien-Sheng Wu[1] **Wenhao Liu**[2] † **Nazneen Rajani**[3] †
[1]Salesforce AI Research, [2]Faire.com, [3]Hugging Face
{pchoubey, afabbri, jvig, wu.jason}@salesforce.com
wenhao@faire.com, nazneen@hf.co

Abstract

Hallucination is a known issue for neural abstractive summarization models. Recent work suggests that the degree of hallucination may depend on factual errors in the training data. In this work, we propose a new method called Contrastive Parameter Ensembling (CaPE) to use training data more effectively, utilizing variations in noise in training samples to reduce hallucination. Starting with a base model fine-tuned on an entire dataset, we additionally train *expert* and *anti-expert* models on clean and noisy subsets of the data, respectively. We then adjust the parameters of the base model by adding (subtracting) the parameters of the expert (anti-expert), advancing the recent work on additive parameter ensembling approaches. Trained on a much smaller data subset, expert and anti-expert models only fractionally (<14%) increases the total training time. Further, CaPE uses parameter ensembling and does not increase the inference time. Experimental results show that CaPE improves performance across different automatic factual metrics and human evaluation, with a maximum improvement of 16.69% and 15.38% on summary-level dependency-arc entailment accuracy for the XSUM and CNN/DM datasets. The CaPE model performs comparably to the base model on metrics of informativeness such as ROUGE.

1 Introduction

Neural abstractive summarization systems have been shown to generate plausible summaries with high lexical overlap with the references. However, human analyses (Fabbri et al., 2021a; Pagnoni et al., 2021; Tejaswin et al., 2021) and automatic evaluations (Falke et al., 2019; Kryscinski et al., 2020; Maynez et al., 2020; Durmus et al., 2020) show that state-of-the-art models trained on the widely used XSUM (Narayan et al., 2018) and CNN/DM (Hermann et al., 2015) datasets tend to hallucinate information with high frequency. The degree of a model's hallucinations further correlates with the quality of the training data (Aralikatte et al., 2021; Pagnoni et al., 2021). For instance, models trained on the XSum data tend to generate a higher proportion of factual errors as compared to models trained on the CNN/DM dataset.

Given the association between training data quality and hallucinations in resulting models, the easiest method to reduce hallucinations is to remove noisy samples from the training data (Nan et al., 2021). However, data filtering reduces the size of training data and consequently, the diversity in target summary since the removed noisy samples might also include useful task-specific knowledge. This impacts other aspects of the generated summaries such as information recall or fluency. In Table 1, we show ROUGE (R-1/2/L) and named entity recall (E-R_{ref}) with respect to the reference summary of a BART model (Lewis et al., 2020) trained on the entity precision-filtered XSUM data

Model	R-1	R-2	R-L	E-R_{ref}
All	45.70	22.53	37.54	53.69
Filtered	41.66	18.39	33.66	42.58
Δ	-8.84%	-18.37%	-10.33%	-20.69%

Table 1: Validation performance comparison of BART models trained on all (204,017 samples) and filtered (50,270 samples) XSUM training data.

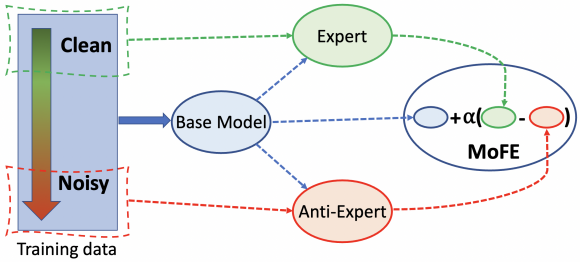

Figure 1: Schematic view of steps for building the CaPE model. It fine-tunes an expert and an anti-expert on the clean and noisy training sets respectively, and uses them to readjust the base summarization model's parameters.

† work was done at Salesforce AI Research.

Findings of the Association for Computational Linguistics: ACL 2023, pages 10755–10773

(24.6% of the original data). The new model drops 8-18% in ROUGE and 20% drop in entity recall.

In this work, we design a simple yet effective strategy to utilize both clean and noisy training samples. Based on the observation that the level of hallucination in a summarization model correlates with the level of noise in the training data, we propose Contrastive Parameter Ensembling (CaPE), which adjusts the parameters of a base model by adding (subtracting) the weights of a model trained on clean (noisy) subsets of the data. This approach is motivated by recent work demonstrating the effectiveness of a simpler form of checkpoint averaging in neural machine translation (Sennrich et al., 2016; Vaswani et al., 2017).

CaPE also builds on other recent work on model ensembling. Jacobs et al. (1991); Liu et al. (2021a) combines expert and anti-expert predictions by taking a weighted average of their output. However, this requires running each model separately and increases computational cost linearly in the number of models, further slowing the auto-regressive generation of summaries. Alternatively, Madotto et al. (2020) proposed attention over parameters that jointly optimizes multiple models and directly combines all their parameters through learned attention coefficients. Furthermore, Wortsman et al. (2021) recently showed that the average of a pre-trained CLIP (Radford et al., 2021) model with its another version that is further fine-tuned on a new data distribution performs better than both models on their complementary distributions.

We evaluate our CaPE model on two benchmark abstractive summarization datasets, XSUM and CNN/DM. We train an expert and an anti-expert corresponding to each of the dependency-arc entailment (Goyal and Durrett, 2020, 2021) and entity overlap (Nan et al., 2021) metrics. Then, we combine each expert and anti-expert pair to obtain four variants of CaPE and evaluate them using the metrics used for data selection E-P$_{src}$, as well as a different entailment metric, MNLI (Williams et al., 2018), and two question answering-based metrics, QuestEval (Scialom et al., 2021) and QAFactEval (Fabbri et al., 2021b), for factual consistency. We find that all variants of our CaPE consistently outperform the state-of-the-art models on all factual metrics, with marginal variations in ROUGE scores and information recall.

2 Contrastive Parameter Ensembling

In this work, we propose Contrastive Parameter Ensembling (CaPE) for reducing hallucinations in text summarization systems. This method refines a base summarization model by training two additional models, an expert and an anti-expert model. An ensemble model is then constructed through a simple linear combination of the parameters of the three models, an approach inspired by recent work on weight (a.k.a. parameter)-space ensembling (Izmailov et al., 2018; Frankle et al., 2019; Neyshabur et al., 2020; Wortsman et al., 2021).

2.1 Measuring Hallucinations for Selecting Training Data

To select data for training the expert and anti-expert, we assume the availability of automated metrics for measuring hallucinations in reference summaries. There are several automatic metrics to evaluate factual consistency such as entity overlap (Nan et al., 2021), entailment score (Kryscinski et al., 2020; Goyal and Durrett, 2020; Maynez et al., 2020), and QA-based metrics (Durmus et al., 2020; Scialom et al., 2021). These methods vary greatly in computational cost and agreement with human judgments of factuality. We use two of the faster metrics that are based on entity overlap and entailment metrics and have shown good correlations with human evaluations, described below.

Entity Overlap is the simplest method measuring token-level named-entity overlap between the summary and source document (Nan et al., 2021). We use **entity token precision** (E-P$_{src}$), the percentage of named-entities tokens in the summary that are also present in the source. This metric can be used as a proxy to measure simpler cases of hallucinations, such as out-of-article entity errors (Pagnoni et al., 2021; Zhou et al., 2021; Cao et al., 2022), also known as *extrinsic hallucinations* (Maynez et al., 2020). A human study by Pagnoni et al. (2021) finds this to be the most frequent form of error in models trained on XSUM data. However, it fails to capture intricate cases of hallucinations such as semantic frame errors (e.g., when an entity is present in the source but is attributed to the wrong predicate).

DAE (Dependency Arc Entailment) (Goyal and Durrett, 2021) measures fine-grained entailment by breaking the summary into smaller claims defined by dependency arcs, covering errors such as incor-

rect predicates or their arguments, coreference errors and discourse link errors, in contrast to the simpler token-level entity overlap. Dependency arcs define grammatical structures in a sentence and often describe semantic connections between words, such as predicate-argument relations. Pagnoni et al. (2021) finds that DAE correlates with human judgment of factuality, and has the highest correlation with complex discourse errors, such as entity coreference. Therefore, we use **DAE errors**, defined as the number of dependency arcs in the summary that are not entailed by the source document, to identify cases of more intricate hallucinations for selecting training data.

2.2 Expert and Anti-Expert Parameter Adjustment

Using the entity overlap or DAE error metric, we select samples for training expert and anti-expert models that are then used to adjust the base model parameters. The data selection strategy, SELECTCLEAN (SELECTNOISY), and the generic process for building CaPE are described below and further illustrated in Algorithm 1.

SELECTCLEAN (SELECTNOISY): For the entity overlap metric, we select clean (noisy) samples with entity precision above (below) a predefined threshold $\epsilon_{clean}^{E-P_{src}}$ ($\epsilon_{noisy}^{E-P_{src}}$). For the DAE error metric, we select clean (noisy) samples with the number of DAE errors below (above) a predefined threshold $\epsilon_{clean}^{DAE_{error}}$ ($\epsilon_{noisy}^{DAE_{error}}$).

Expert (Anti-Expert) Fine-tuning We train a base summarization model using all training data and then fine-tune this model on the clean (noisy) dataset to obtain the expert (anti-expert). By training on the full data followed by fine-tuning on *clean* (*noisy*) subset, we want our expert (anti-expert) model to retain other aspects such as ROUGE and information recall of the base model, and only differ in the factual qualities. As noted in Table 1, this is in contrast to training a BART model on just clean (or noisy) samples that severely deteriorates ROUGE and information recall (analyzed further in § 4.3). Finally, given a mixing coefficient α, we obtain our Contrastive Parameter Ensembled model (θ_{CaPE}) from base (θ_B), expert (θ_E) and anti-expert ($\theta_{\bar{E}}$) parameters following $\theta_{CaPE} = \theta_B + \alpha(\theta_E - \theta_{\bar{E}})$. The mixing coefficient (α) balances factual quality with other aspects of summarization such as ROUGE.

Initializing the expert (anti-expert) from the base or BART model is critical; prior work (Izmailov et al., 2018; Frankle et al., 2019; Neyshabur et al., 2020) has shown that parameter-averaging works well when all constituent models share the *same* optimization trajectory. On the other hand, averaging parameters of disjointly trained deep neural models, starting from different initializations, may not work better than a model with randomly assigned parameters. Since both methods of fine-tuning and training have a common initialization, the resulting CaPE model exhibits performance comparable to the base model or expert.

Algorithm 1 CaPE for Summarization

Require: Training Data D_T, Measure of hallucination M_H
1: Train θ_B on D_T
2: $D_{clean} \leftarrow$ SELECTCLEAN (D_T, M_H)
3: $D_{noisy} \leftarrow$ SELECTNOISY (D_T, M_H)
4: $\theta_E \leftarrow$ Fine-tune θ_B on D_{clean}
5: $\theta_{\bar{E}} \leftarrow$ Fine-tune θ_B on D_{noisy}
6: $\theta_{CaPE} \leftarrow \theta_B + \alpha(\theta_E - \theta_{\bar{E}})$
7: **return** θ_{CaPE}

2.3 CaPE: A generalization of WiSE-FT

Contrastive Paremeter Ensembling generalizes the recently proposed WiSE-FT (Eq. 1) model (Wortsman et al., 2021), which only performs a weighted sum of a base model and a single fine-tuned model, for ensuring distributional robustness on image classification.

$$\theta_{WiSE-FT} = (1. - \alpha)\theta_B + (\alpha)\theta_E \quad (1)$$

Essentially, $\theta_{WiSE-FT}$ is a special case of θ_{CaPE} where the anti-expert is a base model. We believe Eq. 1 a sub-optimal solution for our objective of minimizing factual errors. Being trained on the noisiest subset of the training data, the anti-expert model hallucinates with higher frequency than the base and expert models, removing parameters responsible for hallucinations more than the other two. We empirically find that our proposed contrastive ensembling outperforms the models that just use one of the expert or anti-expert in § 4.4.

3 Results

We evaluate CaPE on the XSUM (Narayan et al., 2018) and CNN/DM (Hermann et al., 2015) datasets. The XSUM data is highly abstractive and noisy, while CNN/DM is more extractive and contains fewer factual errors (Tejaswin et al., 2021). These data variations allow us to evaluate CaPE under different data quality settings. Besides the

standard ROUGE-1/2/L (R1/R2/RL) scores, we use a diverse set of metrics for evaluating factual consistency and summary quality.

- **D**$_{arc}$ measures the percentage of dependency arcs in the summary entailed by the source article.
- **D**$_{sum}$ measures the percentage of summaries that do not have any dependency arc error.
- **E-P**$_{src}$ measures the percentage of entities in the summary that are present in the source article.
- **E-R**$_{ref}$ measures the percentage of entities in the reference that are also present in the generated summary.
- **BS-P (R)** represents the BERTScore precision (recall) w.r.t. the source article (Zhang et al., 2019).
- **QEval** represents a QA-based factual consistency metric (Scialom et al., 2021).
- **MNLI** measures the entailment score based on the RoBERTa large (Liu et al., 2019) model trained on MNLI dataset (Williams et al., 2018). The score of a summary sentence is the maximum entailment score over all input sentences, and the final score is averaged across summary sentences as in Laban et al. (2022).
- **QAFactEval** is another QA-based factual consistency metric that improves question filtering and answer overlap components (Fabbri et al., 2021b).

3.1 Models

We use the BART-based summarization (BART$_{sum}$) models released with Huggingface's transformers library (Wolf et al., 2020) (*bart-xsum/cnn-large*) as the base models. From human analyses, Pagnoni et al. (2021); Fabbri et al. (2021a) find that BART$_{sum}$ models generated summaries have the least number of factual errors. We adopt the standard hyperparameters for all models during the inference.

We train an expert (anti-expert) for each of the DAE error (Exp$_{DAE}$ (Anti$_{DAE}$)) and entity overlap (Exp$_{E-P}$ (Anti$_{E-P}$)) metrics. We evaluate four variants of CaPE. CaPE$_{PP}$ uses Exp$_{E-P}$ and Anti$_{E-P}$, CaPE$_{DP}$ uses Exp$_{DAE}$ and Anti$_{E-P}$ (CaPE$_{DD}$ and CaPE$_{PD}$ follow the same naming convention). Depending on the value of α, CaPE may reduce ROUGE or information recall while improving factual consistency. Therefore, for each variant of CaPE, we select the α such that it does not underperform the base model by more than 1% on ROUGE 1 (R1) and entity recall (E-R$_{ref}$) metrics on the validation set[1].

Baselines: We compare CaPE with two summarization baselines: the base model (BART$_{sum}$) and an ensemble of BART-based summarization models. The ensemble model uses the average of a base summarization and two other summarization models obtained by fine-tuning the base model on two randomly sampled subsets of the training data. We also compare CaPE to three post-processing models. The first is a variation of the autoregressive fact correction model from Dong et al. (2020), in which we train a BART-large model to produce the reference summary conditioned on the concatenation of the source and reference summary with all entity slots masked (PP). The second is a modified version of PP trained on the subset of data with an entity precision of 100 (PP-clean). The last is the model from Chen et al. (2021), which generates candidate summaries by (1) enumerating all ways to replace entities in summary with entities of similar type in the input and (2) training BART with an additional classification layer to re-rank these summaries (PP-CC).

3.2 Automatic Evaluation

Table 2 summarizes the results on the XSUM and CNN/DM datasets. First, we find that the ensemble model slightly improves ROUGE scores, BERTScore recall and entity recall on the CNN/DM dataset, but not necessarily factual consistency metrics. On the other hand, all variants of CaPE outperform the base as well as the ensemble across all factual consistency metrics on both the XSUM and CNN/DM datasets. Given the controllability achieved by α, we ensure that all variants of CaPE preserve ROUGE scores and information recall within a pre-defined threshold of maximum 1% drop from the base model. We also find that CaPE models improve BERTScore precision (BS-P) with respect to the source article on both XSUM and CNN/DM. This is noteworthy given recent work on benchmarking different evaluation metrics that suggests that BERTScore precision with respect to the source document correlates with the human judgments of factuality (Pagnoni et al., 2021).

CaPE models also outperform the post-

[1]We find α using grid search, assigning a minimum value of 0.2 and incrementing it by a step size of 0.2.

Model	D_{arc}	D_{sum}	E-P_{src}	E-R_{ref}	QEval	BS-P	BS-R	R1	R2	RL	TT	IT
XSUM												
Base	76.16	34.75	63.82	53.66	36.54	88.93	79.86	45.34	22.21	37.13	1x	1x
Ensemble	75.22	33.48	62.63	**54.23**	36.37	88.82	79.86	45.27	22.28	37.09	1.2x	1x
PP	75.65	33.67	62.36	53.93	36.37	88.88	79.84	45.34	**22.30**	37.18	2-3x	2x
PP-Clean	79.41	40.09	**72.98**	45.72	37.01	89.09	79.84	43.82	20.40	35.89	1.5x	2x
PP-CC	76.88	35.99	66.06	52.23	36.62	88.95	79.85	45.03	21.87	36.89	-	2x
CaPE$_{DD}$	78.51	39.36	65.61*	52.91	36.90	89.08	79.81	45.33	22.29	**37.27**	1.05x	1x
CaPE$_{PP}$	78.46	39.13	69.12	53.36	37.09	89.07	**79.89**	45.16	21.91	36.94	1.08x	1x
CaPE$_{DP}$	**79.61**	**40.55**	68.24	53.91	**37.22**	**89.15**	**79.89**	45.14	21.97	36.92	1.07x	1x
CaPE$_{PD}$	77.91	38.40	66.12*	52.77	36.84	89.05	79.81	**45.35**	22.25	37.17	1.06x	1x
CaPE$_{DP*}$	83.87	48.78	74.30	52.34	38.05	89.41	79.93	43.56	20.39	35.46	1.07x	1x
CNN/DM												
Base	96.26	75.0	98.44	58.92	59.24	93.26	82.62	44.05	21.07	40.86	1x	1x
Ensemble	95.19	67.44	97.72	**61.93**	59.51	93.06	**82.91**	**44.28**	**21.23**	**40.88**	1.2x	1x
PP	96.14	74.70	98.26	58.40	59.15	93.23	82.58	43.95	20.94	40.76	2-3x	2x
PP-Clean	96.17	74.77	98.63	58.20	59.16	93.23	82.59	43.92	20.92	40.74	2x	2x
PP-CC	95.72	72.63	98.52	58.57	59.11	93.22	82.61	43.97	20.98	40.79	-	2x
CaPE$_{DD}$	**98.23**	**86.54**	98.90	58.35	**60.10**	**93.80**	82.84	43.75	20.79	40.44	1.04x	1x
CaPE$_{PP}$	97.17	80.46	**99.16**	58.66	59.65	93.52	82.71	43.62	20.72	40.33	1.14x	1x
CaPE$_{DP}$	97.59	83.04	98.86	58.86	59.70	93.56	82.78	43.71	20.80	40.42	1.06x	1x
CaPE$_{PD}$	96.97	79.39	98.66	58.60	59.61	93.44	82.68	44.05	21.07	40.83	1.11x	1x

Table 2: Performance comparison of CaPE and baseline models on XSUM and CNN/DM datasets. CaPE$_{DP*}$ is a variant of CaPE$_{DP}$ with α set to 1.0. TT (IT) represents training (inference) time relative to the base model. All four variants of CaPE are significantly better than the *Base*, *Ensemble*, *PP-Clean* and *PP-CC* models (two-sided approximation randomization test, at least p < 0.005) on all factuality metrics, except those marked with '*' (*: our models are not better than the *PP-CC* model on E-P_{src} metric). Furthermore, *CaPE$_{DP*}$* is significantly better than the *PP-Clean* model with p < 0.001 on all factuality metrics.

Model	XSUM		CNN/DM	
	MNLI	QAFactEval	MNLI	QAFactEval
Base	22.70	2.104	84.20	4.550
PP-Clean	22.30	2.098	84.40	4.544
CaPE$_{DP}$	**23.10**	**2.205**	**86.80**	**4.602**

Table 3: MNLI and QAFactEval metrics-based evaluations of base, PP-clean and the CaPE$_{DP}$ model.

processing-based approaches PP and PP-CC on XSUM and all three PP, PP-clean and PP-CC approaches on CNN/DM dataset by a significant margin. However, PP-clean performs similarly to CaPE on factual consistency metrics on XSUM and even obtains a higher E-P_{src} score of 72.98. At the same time, PP-clean lowers the performance on ROUGE and information recall, reducing E-R_{ref} performance by ∼15% (underlined in Table 2). Fortunately, we can set the mixing coefficient α in CaPE to a higher value, achieving higher factual consistency at the cost of reduced ROUGE and information recall. To confirm this, we also report the performance of CaPE$_{DP*}$ on XSUM data which uses Exp$_{DAE}$ and Anti$_{E-P}$ mixed with α value of 1.0 (underlined results in Table 2). We find that CaPE$_{DP*}$ obtains much higher scores than the PP-Clean model on all factual consistency metrics, while competently retaining the information recall of the base model (E-R_{ref} reduced by 3.5% compared to ∼15% drop for PP-clean).

Finally, in Table 3, we compare CaPE$_{DP}$ (the variant of CaPE with the best trade-off, discussed in §4.2), base and PP-clean models using two additional metrics, QAFactEval and MNLI. As noted by Fabbri et al. (2021b), prior studies comparing factual metrics draw inconsistent conclusions, with a few observing QA-based metrics as superior to entailment metrics (Durmus et al., 2020; Scialom et al., 2021) and others reporting the opposite (Maynez et al., 2020). To the best of our knowledge, QAFactEval performs the best on the SummaC benchmark (Laban et al., 2022), used for comparing factual consistency metrics. On both metrics, we find that CaPE$_{DP}$ outperforms both base and PP-clean models, improving the QAFactEval score by 4.8% and 1.14% over the base model on XSUM and CNN/DM, respectively.

Transferability of Experts (Anti-Experts): We observe that CaPE models also improve performance on the metrics that were not used for training the expert or anti-expert. For instance, CaPE$_{PP}$ outperforms base model on the D_{arc}/D_{sum} metrics, and CaPE$_{DD}$ outperforms base model on the E-P_{src} metrics on both XSUM and CNN/DM. All variants of CaPE also outperform base model on QEval, QAFactEval and MNLI, which were also

not used during the development of experts (anti-experts). Secondly, we find that the experts and anti-experts are interchangeable, an expert trained on data selected using one metric can be used in conjunction with an anti-expert based on another metric. As evident, both CaPE$_{DP}$ and CaPE$_{PD}$ outperform base model, with CaPE$_{DP}$ achieving best trade-offs among other variants of CaPE on the XSUM data, discussed further in §4.2.

Computational Efficiency: We also report the approximate training (TT) and inference (IT) time for different models relative to the base model in Table 2. We exclude the time required for data processing (e.g. data selection for CaPE and PP-Clean during training, or entity recognition for all post-processing-based models both during training and inference). We find that CaPE models only marginally increase the training time ($\leq$14%) required for fine-tuning expert (anti-expert) on a smaller selected subset of training data. Further, CaPE models do not increase the inference time. In comparison, post-processing methods use separate models for correcting summaries generated by the base model, increasing the memory required to store the additional model as well as both the training and inference time.

3.3 Human Evaluation

Following Cao and Wang (2021), we also perform a pairwise comparison of summaries, where human annotators rate each CaPE$_{DP}$-generated summary against the base model-generated summary for factual consistency. We rate 100 random articles from each of the XSUM and CNN/DM datasets. The inter-annotator agreement is 0.8385 (Krippendorff, 2011) based on our sampled articles-summary pairs from XSUM. Annotators find CaPE$_{DP}$ improves (degrades) factual consistency on 19% (14%) summaries on XSUM data and improves (degrades) factual consistency on 6% (2%) summaries on CNN/DM data. Factual consistency remained unchanged for the remaining 67% and 92% summaries from the XSUM and CNN/DM datasets, respectively. We show a few sample outputs illustrating the qualitative effect of α and common errors corrected by the CaPE model in § B.

4 Analysis

4.1 Expert (Anti-Expert) Performance

In Table 4, we compare the performance of individual expert and anti-expert models on DAE- and entity-based metrics. We observe:

Model	D$_{arc}$	D$_{sum}$	E-P$_{src}$	E-R$_{ref}$	R1
Base	76.16	34.75	63.82	53.66	**45.34**
Exp$_{DAE}$	**82.09**	**41.35**	67.73	53.04	44.79
Anti$_{DAE}$	68.38	18.16	57.91	**57.36**	42.6
Exp$_{E-P}$	78.81	36.42	**69.81**	51.60	44.53
Anti$_{E-P}$	74.03	28.74	57.15	50.58	44.23

Table 4: Performance of individual experts (anti-experts) on XSUM. Maximum scores are bolded and minimum scores are underlined for each of the metrics.

An expert reduces hallucinations in generated summaries. We find that all experts are able to achieve improved performance on the metric used for selecting the training data subset. We further observe that the improvements for experts are not limited to the metrics used for data selection. For instance, Exp$_{DAE}$ improves entity precision (E-P$_{src}$) by ∼6% and Exp$_{E-P}$ improves D$_{arc}$ and D$_{sum}$ by ∼3%.

An anti-expert increases hallucinations in generated summaries. All anti-experts reduce performance on factual consistency metrics, with the maximum drop seen on summary-level D$_{sum}$ metric, indicating that a greater proportion of anti-expert generated summaries are hallucinated. At the same time, they generate well-formed summaries, as indicated by their high (slightly worse than the base model) ROUGE scores. This is the desired behavior for an anti-expert that should generate hallucinated but well-formed summaries.

4.2 Effects of Mixing Coefficient α

We combine the expert and anti-expert with the base model using different mixing coefficients (α) and plot their performance on the XSUM data and CNN/DM datasets in Figure 2. We choose to vary α from 0.0 to 1.0. We compare models on the D$_{arc}$, E-P$_{src}$, E-R$_{ref}$, and ROUGE 1 metrics. We observe:

α works as a control knob for adjusting factual correctness and informativeness. As we increase the α, the performance of CaPEs on factuality metrics (D$_{arc}$, E-P$_{src}$) improves while it decreases on metrics of informativeness (E-R$_{ref}$, ROUGE). Consequently, we can select the α that obtains the highest factual consistency while retaining ROUGE/E-R$_{ref}$ within the pre-defined tolerance level.

Intermixing an expert and an anti-expert based on different metrics provides the best performance trade-offs. CaPE$_{DD}$, which uses

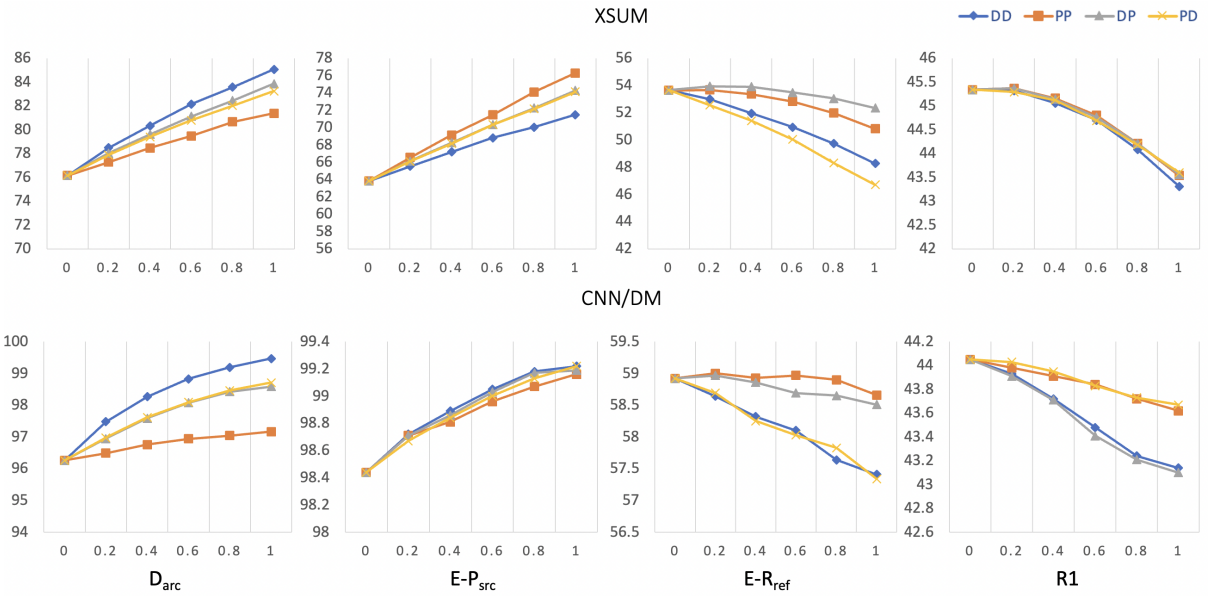

Figure 2: Variations in the performance of CaPE and base models with different values of mixing coefficient α on XSUM and CNN/DM datasets (α=0.0 corresponds to only base model.).

the DAE-based expert and anti-expert, improves D_{arc}/D_{summ} accuracy at the fastest rate on both datasets. Likewise, CaPE$_{PP}$ improves entity precision, E-P$_{src}$, at the fastest rate. CaPE$_{DP}$ and CaPE$_{PD}$ models that intermix the expert and anti-expert based on different metrics provide the best bargain on all factual consistency metrics, evenly improving all D_{arc} and E-P$_{src}$ scores. On the ROUGE score, we do not find any uniform pattern between the two datasets. On XSUM, all CaPE variants exhibit similar behavior while on CNN/DM, CaPEs using the entity precision-based anti-expert (CaPE$_{PP/DP}$) retain ROUGE better than their alternatives. Similarly, CaPE$_{PP/DP}$ retain entity recall better than their alternatives for all values of α on both datasets. Overall, CaPE$_{DP}$ provides the best balance for all performance measures on both datasets.

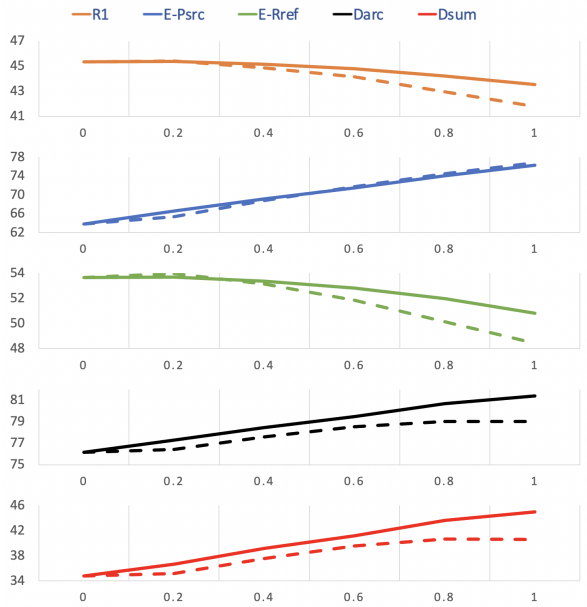

Figure 3: Performance comparison of CaPE$_{PP}$ models obtained by fine-tuning the base summarization model (solid) vs training the BART model (dashed) based on data selected according to the entity precision metric.

4.3 (Anti-)Expert Initialization: A Base Summarization Model outperforms BART

In Figure 3, we compare two variants of the CaPE$_{PP}$ model, the first initializes the expert (anti-expert) with the BART and the second with the base summarization model. First, we find that both models improve performance on all factual consistency metrics. On the E-P$_{src}$ metric, which was also used to select the training samples, both models obtain comparable improvements. However, on the DAE-based factual consistency metrics as well as ROUGE and E-R$_{ref}$ metrics, fine-tuning the base model outperforms the one based on training BART. The gap in performance increases with the increase in value of α, i.e., when the influence of expert (anti-expert) increases. The performance difference is unsurprising, given that the re-trained model leads to lower ROUGE and information recall (Table 1) by being trained on fewer training samples. Secondly, training an expert model initialized with

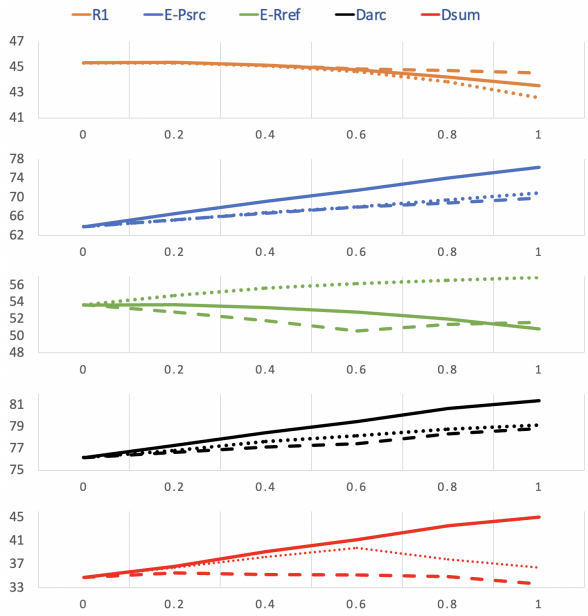

Figure 4: Performance comparison of CaPE (solid), expert only (dashed) and anti-expert only (dotted) models based on data selected according to the entity precision metric.

BART takes a greater number of parameter updates ($>$ 1 epoch) to reach the best performance on ROUGE and other metrics. In contrast, the base model already yields a higher ROUGE score, and fine-tuning it for 1 epoch is sufficient to reduce hallucinations, making fine-tuning a more efficient approach for building experts (anti-experts).

4.4 CaPE outperforms Simple Parameter Ensembling (WiSE-FT)

In Figure 4, we compare the CaPE$_{PP}$ model with the expert (anti-expert) only model that replaces the anti-expert (expert) with the base model in θ_{CaPE}. Accordingly, the expert only model is equivalent to the WiSE-FT formulation (θ_{WiseFT}). While both the expert only and anti-expert only improve performance on factual consistency metrics, we observe that CaPE$_{PP}$ improves performance at a faster rate than the former two models. On ROUGE-1 and E-R$_{ref}$ scores, the CaPE$_{PP}$ performance lies in between the expert only and anti-expert only models. The performance variations for the three models indicate that the contrastive ensembling combines the gains from expert and anti-expert, helping us to effectively use both clean and noisy data.

5 Related Work

Abstractive text summarization metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) evaluate lexical and semantic overlap respectively but fail to sufficiently evaluate factuality and faithfulness (Tejaswin et al., 2021). This has led to a line of research dedicated to evaluating factual consistency and hallucination in text summarization using new metrics such as entailment and question answering-based evaluation (Falke et al., 2019; Kryscinski et al., 2020; Maynez et al., 2020; Zhou et al., 2021; Eyal et al., 2019; Scialom et al., 2019; Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021). The research focused on comparing these factual consistency evaluation metrics (Gabriel et al., 2021; Fabbri et al., 2021a; Pagnoni et al., 2021; Goyal and Durrett, 2021; Tejaswin et al., 2021), however, often have contradicting observations. For instance, Durmus et al. (2020) found that entailment-based automated metrics have lower correlations with factual consistency while Pagnoni et al. (2021) concluded that the entailment-based FactCC exhibits the highest correlations with human judgments of factual consistency. Given the variations in findings from different human analyses of popular factual consistency evaluation metrics, we select a few metrics from each of the entailment, entity overlap, and QA-based evaluations, as well as use ROUGE and BERTScore metrics for evaluating CaPE.

Along with the growing body of work on the analysis and evaluation of factual consistency, there has been some recent work on developing methods to enforce factual consistency in pre-trained language models. These include sampling techniques such as constrained decoding (Mao et al., 2020) and neurologic decoding (Lu et al., 2020). Another strategy is to control generation either by using language models to guide a base language model as in GeDi (Krause et al., 2020) and DExperts (Liu et al., 2021a) or via a hallucination knob (Filippova, 2020). Although these methods claim to be generic, they have not been successfully applied to constrain summary generation on the source document.

Comparatively, there are fewer papers that propose methods for factual consistency in text summarization. Most of these focus on posthoc correction such as SpanFact (Dong et al., 2020), contrast entity generation and selection (Chen et al., 2021), loss truncation (Kang and Hashimoto, 2020; Goyal and Durrett, 2021), and encoding SRL structure (Cao et al., 2020). Aralikatte et al. (2021) uses focus attention and sampling to improve the diversity and faithfulness of summaries while Liu

et al. (2021b) uses data augmentation with a contrastive loss for factual consistency of abstractive summarization applied to customer feedback.

Finally, works focusing on data noise include revising hallucinated summaries in training data (Adams et al., 2022), dropping hallucinated samples (e.g. Nan et al. (2021) and Narayan et al. (2021) for summarization, Matsumaru et al. (2020) for headline generation), or defining curriculum based on the factual quality of training samples (Kano et al., 2021).

6 Conclusion

We present Contrastive Parameter Ensembling (CaPE) to reduce content hallucinations in abstractive summarization models. We first select clean (noisy) training samples to fine-tune an expert (anti-expert) model. Then, we use the difference between the parameters of expert and anti-expert models to adjust the parameters of a base summarization model. We evaluate CaPE on the XSUM and CNN/DM datasets using a diverse set of factual metrics, finding that CaPE effectively reduces hallucinations without a significant drop in ROUGE and information recall.

Limitations

The datasets utilized in this research contain documents and summaries in English (XSUM and CNN/DM datasets) and thus mainly represent the culture of the English-speaking populace. Gender, age, political or other biases may also exist in the dataset, and models trained on these datasets may propagate these biases.

Our experiments and analyses are based on the assumption that training data contains artifacts that lead to factual errors in summarization models. Also, it is evident from the results, that the effectiveness of our proposed models is relatively higher for the noisier XSUM dataset. So, our analytical results and improvement from a model may have limited implications on a perfect dataset that does not exhibit any learnable artifacts.

We relied on automated metrics, such as ROUGE and entity recall for measuring information relevance, and entity precision, question answering-based metrics and dependency arc entailment accuracy for information correctness. These metrics are error-prone. Exclusively for a subset of models, that perform the best according to automated metrics, we use human annotations for additional evaluations.

References

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization.

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. *arXiv preprint arXiv:2105.11921*.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5935–5941, Online. Association for Computational Linguistics.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

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.

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings*

of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021b. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *CoRR*, abs/2112.08542.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 864–870, Online. Association for Computational Linguistics.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. 2019. Linear mode connectivity and the lottery ticket hypothesis. *CoRR*, abs/1912.05671.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive Mixtures of Local Experts. *Neural Computation*, 3(1):79–87.

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.

Ryuji Kano, Takumi Takahashi, Toru Nishino, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma. 2021. Quantifying appropriateness of summarization data for curriculum learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1395–1405, Online. Association for Computational Linguistics.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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.

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.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021a. 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.

Yang Liu, Yifei Sun, and Vincent Gao. 2021b. Improving factual consistency of abstractive summarization on customer feedback. *arXiv preprint arXiv:2106.16188*.

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. *CoRR*, abs/1907.11692.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Neurologic decoding:(un) supervised neural text generation with predicate logic constraints. *arXiv preprint arXiv:2010.12884*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, Jamin Shin, and Pascale Fung. 2020. Attention over parameters for dialogue systems. *CoRR*, abs/2001.01871.

Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained abstractive summarization: Preserving factual consistency with constrained generation. *arXiv preprint arXiv:2010.12723*.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1335–1346, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. 2020. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, pages 512–523. Curran Associates, Inc.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. *arXiv preprint arXiv:2103.12693*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Priyam Tejaswin, Dhruv Naik, and Pengfei Liu. 2021. How well do you know your summarization datasets? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3436–3449, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 5008–5020, Online. Association for Computational Linguistics.

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

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hanna Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2021. Robust fine-tuning of zero-shot models. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

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

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

A Experimental Details

Data Selection: For SELECTCLEAN (SELECTNOISY), we set $\epsilon_{clean}^{E-P_{src}}$, $\epsilon_{clean}^{DAE_{error}}$, $\epsilon_{noisy}^{E-P_{src}}$ and $\epsilon_{noisy}^{DAE_{error}}$ to 1.0 (1.0), 0.0 (0.0), 0.75 (0.10) and 15.0 (15.0) respectively for CNN/DM (XSUM) dataset. In Table 5, we report the size of data subsets used for training experts and anti-experts used in our CaPE models.

We set the data selection thresholds for anti-experts based on their validation performance on their respective factuality metric. The plot for D_{arc} and E-P_{src} for Anti$_{DAE}$ and Anti$_{E-P}$ on XSUM and CNN/DM datasets are shown in Figure 5. The

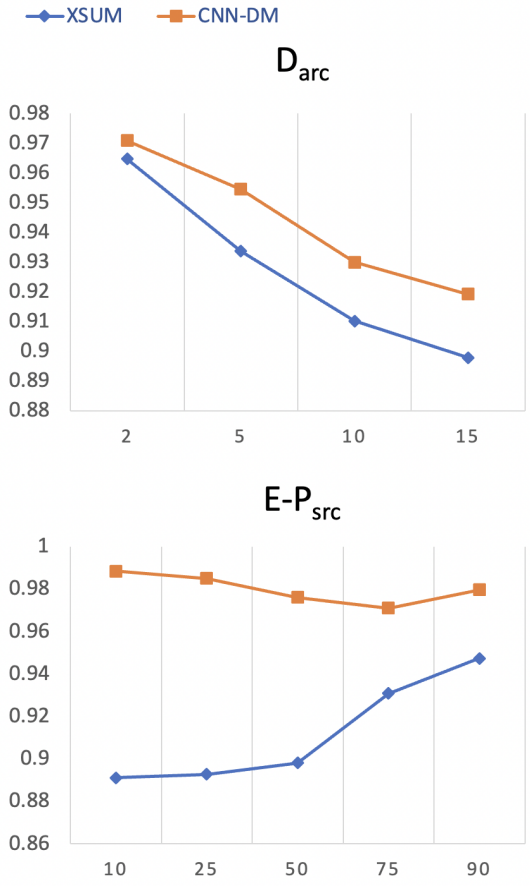

Figure 5: D_{arc} and E-P_{src} for Anti$_{DAE}$ and Anti$_{E-P}$ models with different data selection thresholds. The performance values are normalized w.r.t. the performance of BART model.

size of datasets corresponding to different thresholds is shown in Table 6. We observe that selecting a small but noisiest subset of training data is the best strategy for training an anti-expert.

Data	Exp$_{DAE}$	Anti$_{DAE}$
XSUM	39009 (19.1%)	7962 (3.9%)
CNN/DM	39643 (13.8%)	8786 (3.1%)
	Exp$_{E-P}$	Anti$_{E-P}$
XSUM	50270 (24.6%)	26208 (12.8%)
CNN/DM	152418 (53.0%)	31727 (11.1%)

Table 5: Number of samples (percentage of total training data) in data subset used for training experts and anti-experts.

Training Experts (Anti-Experts): We use Huggingface Transformers library (Wolf et al., 2020) (PyTorch (Paszke et al., 2017)) to implement our experts (anti-experts). We initialize experts with the pre-trained summarization models (*bart-large-xsum*, *bart-large-cnn*) and fine-tune them for 1 epoch with batch size of 64 using default train-

$\epsilon_{noisy}^{E-P_{src}}$	0.1	0.25	0.5	0.75	0.9
XSUM	26356	35507	70604	130119	153324
CNN/DM	761	1028	4204	31727	101402

$\epsilon_{noisy}^{DAE_{error}}$	2	5	10	15
XSUM	142623	86416	32689	7962
CNN/DM	219269	140713	48696	8786

Table 6: Number of training samples in data subset with different data selection thresholds.

ing hyperparameters (optimizer: Adam, learning rate: 5e-5, β_1: 0.9, β_2: 0.999, ϵ: 1e-8). The experts (anti-experts) initialized with BART are trained for 5 epochs.

In Table 7, we report the α values for all variants of CaPE on XSUM and CNN/DM datasets.

Data	CaPE$_{DD}$	CaPE$_{PP}$	CaPE$_{DP}$	CaPE$_{PD}$
XSUM	0.20	0.40	0.40	0.20
CNN/DM	0.40	1.0	0.40	0.20

Table 7: α values for CaPE models.

Inference: We adopt the standard hyperparameters for all models during the inference, e.g. beam size of 6 (4), the minimum and maximum sequence length of 11 (56) and 62 (142), etc. for the XSUM (CNN-DM) model.

Experts (Anti-Experts) Performance on CNN/DM: In Table 8, we report the performance of experts and anti-experts on the CNN/DM dataset. They perform similarly to the experts (anti-experts) on the XSUM dataset.

Evaluation Metrics Details We use the *DAE_xsum_human_best_ckpt* and the *CNNDM_synthetic/ENT-C_dae* models for calculating DAE scores for the XSUM and CNN/DM models respectively. The hash for the BERTScorer used in our experiments is "*roberta-large_L17_no-idf_version=0.3.9(hug_trans=4.6.1)*"

Model	D$_{arc}$	D$_{sum}$	E-P$_{src}$	E-R$_{ref}$	R1
Base	96.26	75.0	**98.44**	58.92	44.05
Exp$_{DAE}$	**97.50**	**80.40**	98.30	60.42	44.04
Anti$_{DAE}$	88.49	40.72	96.54	**61.58**	43.79
Exp$_{E-P}$	95.31	68.16	98.40	60.9	**44.57**
Anti$_{E-P}$	93.48	57.85	95.46	60.13	44.27

Table 8: Performance of individual experts (anti-experts) on the CNN/DM dataset. Maximum scores are bolded and minimum scores are underlined for each of the metrics.

B Sample Outputs

Effect of α: In Tables 9, we show examples to qualitatively illustrate the effect of α. In example 1, BART summary contains one factual error ('a five-month experiment') which gets removed in CaPE$_{DP}$-generated summary if we set α to 0.40. As we further increase α, the summary doesn't change in informativeness and retains its factual consistency. Similarly, in example 2, the BART summary contains one factual error, '10th anniversary'. However, increasing α, in this case, couldn't remove the factual error while preserving the informativeness. Rather, for $\alpha \geq 0.8$, CaPE$_{DP}$ removes the factual error but also changes the key information. Our qualitative analysis corroborates the empirical findings, where increasing α continues to improve the factuality but higher values lead to lower entity recall and ROUGE.

How CaPE$_{DP}$ differes from the BART? In Table 10, we show two examples where CaPE$_{DP}$ improves summary by completely removing factual errors. In examples 3, CaPE$_{DP}$ correctly removes an extrinsic entity error '20 people' and an extrinsic event error 'killing' without modifying the factually correct context. However, in example 4, while removing the factual error £1m, CaPE$_{DP}$ also removes its associated context *'fraud at the charity he worked for was uncovered'*. Through manual analyses, we find that CaPE models are susceptible to removing correct information. This is unsurprising since any model that aims to remove factual errors is susceptible to losing recall. However, as noted by the comparable performance of BART and CaPE models on E-R$_{ref}$, the loss of information is negligible when evaluated on the entire test set.

In Table 11, we show examples where CaPE$_{DP}$ only partially removes the error. In example 5, CaPE$_{DP}$ correctly replaces an entity error India's railways minister with the correct entity a senior Maoist leader but it failed to correct the error Maoist rebels. In example 6, BART summary contains two extrinsic errors Seamus and at the age of 60. To correct these errors, CaPE$_{DP}$ preferably replaces *'at the age of 60'* with another important information (starred in the RTE soap opera Glenroe) from the source article. However, while CaPE$_{DP}$ correctly identifies the error Daithi, it replaces that with

another extrinsic error Seamus . The best possible correction strategy for the error Daithi would have been to simply remove it. However, CaPE inherits the tendency of BART models to generate certain types of information (e.g. *First Name*) irrespective of whether that information is inferable from the source article.

Example 1
The six ""astronauts"" wearing bright blue jump-suits and even surgical masks, were paraded before banks of television cameras and hordes of journalists at a news conference before entering their mock spaceship.Amongst the long rows of VIPs at the news conference were senior officials from the United States, China and the European Union.If, as some experts believe, the main aim of the Mars 500 experiment is to publicise the concept of human flight to the red planet, then it has surely succeeded beyond all expectations.""I am very happy to be part of this project,"" said Diego Urbina, the Colombian-Italian and most extrovert member of the crew.""It will raise awareness of space flight so hopefully a few years from now there will be a real flight to Mars.""He confessed that Elton John had been his inspiration.""I don't know if you know that song Rocket Man,"" he asked.""I want a future like thatâ€¦ where people will be going frequently into space and will be working there and it will be very usual.""In front of the world's media, all the team spoke confidently about the chances of the experiment being successful - in other words that noone would crack under the stress of such lengthy confinement in such claustrophobic and bizarre conditions and demand to be let out.""The target is for all six of us to be here for 520 days,"" said the French crew-member Romain Charles who took a guitar with him into the cluster of brown and silver-coloured metal tubes which will be home until November 2011.After the news conference, the six crew disappeared, re-emerging an hour later by the entrance hatch to the mock spaceship, where they put on another high-spirited performance for the media.Finally, blowing kisses and waving to wives, girlfriends and relatives, they walked up the steps and through the entrance hatch.A solemn-faced official slowly closed and sealed it behind them.So now reality bites for the six-member volunteer crew.What will they be thinking as they sit inside their tin cans in north-west Moscow where outside the warm sun shines and the flowers blossom?There is no thrill of a blast-off and flight through space.There are no windows from which to watch the Earth gradually shrink away.And no anticipation of reaching a new world more than fifty million kilometres away.Instead, silent inertia, stale air and tinned food.And everywhere cameras watching their every move, looking out for signs of mental collapse.They have just one thing to cling on to, that they are playing their part in the history of space exploration.That their success in this experiment will mean a human flight to Mars is a step closer.And space experts already believe the first flight could be just 25 years away or even less if there is the political and economic will from countries with advanced space programmes.
BART: The crew of the Mars 500 experiment have arrived in Moscow for the start of a five-month experiment in which they will spend 520 days locked inside tin cans.
CaPE$_{DP}$ (α=0.20): The crew of the Mars 500 experiment have arrived in Moscow for the start of a five-month experiment in which they will spend 520 days locked inside tin cans.
CaPE$_{DP}$ (α=0.40): The crew of the Mars 500 experiment have arrived in Moscow to spend 520 days locked inside a series of tin cans.
CaPE$_{DP}$ (α=0.60): The crew of the Mars 500 experiment have arrived in Moscow to spend 520 days locked inside a series of tin cans.
CaPE$_{DP}$ (α=0.80): The crew of the Mars 500 experiment have arrived in Moscow to spend 520 days locked inside a series of tin cans.
CaPE$_{DP}$ (α=1.0): The crew of the Mars 500 experiment have arrived in Moscow to spend 520 days locked inside a series of tin cans.
Example 2
The submarine, one of the Russian navy's most advanced vessels, sank in the Barents Sea on 12 August, 2000 with the loss of all 118 people on board.An explosion of fuel from an old torpedo caused the disaster.Moscow's response to one of the greatest disasters in Russian naval history was widely criticised.Relatives and members of Russia's northern fleet are due to cast wreaths into the sea on Thursday in memory of the crew.Flags are being flown at half-mast at the headquarters of all Russia's naval fleets, and a ceremony and minute's silence was being held at Moscow's Central Army Museum.The initial response to the disaster in 2000 was shambolic, says the BBC's Richard Galpin.After radio contact was lost there was a still unexplained delay before a search and rescue mission was launched.Although the submarine was lying just 100m below the surface of the sea, attempts to locate it and reach it repeatedly failed.It was days before the authorities informed relatives that something was wrong and the then President, Vladimir Putin, initially remained on holiday.Russia eventually accepted international assistance, but when Norwegian divers opened the Kursk's hatch 10 days later they found the boat flooded and everyone dead.Many had died within seconds of the initial explosion, but others survived for several hours, a report found.Russian officials originally suggested the submarine may have collided with a foreign ship or with a stray mine.But it emerged that an explosion was caused by fuel that had leaked from a torpedo.This started a fire, which subsequently caused all ammunition on board to detonate.The boat was raised and the bodies recovered in 2001.
BART: Russia is marking the 10th anniversary of the sinking of the Kursk submarine.
CaPE$_{DP}$ (α=0.20): Russia is marking the 10th anniversary of the sinking of the Kursk submarine.
CaPE$_{DP}$ (α=0.40): Russia is marking the 10th anniversary of the sinking of the Kursk submarine.
CaPE$_{DP}$ (α=0.60): Russia is marking the 10th anniversary of the sinking of the Kursk submarine.
CaPE$_{DP}$ (α=0.80): A minute's silence is being held in Moscow to remember the crew of the Russian submarine Kursk, which sank in 2000.
CaPE$_{DP}$ (α=1.0): A minute's silence is being held in Moscow to remember the crew of the Russian submarine Kursk, which sank in 2000.

Table 9: Examples: Effect of α on summary factuality and informativeness.

Example 3
Militants armed with guns and grenades gained entry after one detonated explosives at a hospital gate and then opened fire on staff and patients.Commandos who landed on the Sardar Daud hospital roof killed all four attackers after several hours of fighting.The so-called Islamic State (IS) group has claimed the attack.The Taliban has denied any involvement.More than 50 people were also wounded, the defence ministry said.World powers jostle in Afghanistan's new 'Great Game'How successful has IS been in Afghanistan?Stuck between IS and the TalibanPresident Ashraf Ghani said the attack at the 400-bed hospital "trampled all human values"."In all religions, a hospital is regarded as an immune site and attacking it is attacking the whole of Afghanistan," he said.The attack began at 09:00 local time (04:30 GMT). One hospital staff member who was able to get out saw an attacker "wearing a white coat holding a Kalashnikov and opening fire on everyone, including the guards, patients and doctors".One employee wrote on Facebook: "Attackers are inside the hospital. Pray for us."The hospital attack marks a change in approach by so-called Islamic State fighters in Afghanistan - it's the first time they have engaged directly with security forces in the capital.Previously they have targeted civilian gatherings, mainly of Shia Muslims, as well as causing carnage at the Supreme Court last month.But at the hospital they used an approach more commonly associated with the Taliban - blowing the gates open to allow gunmen to enter. This suggests they now have the resources and the military training to expand their attacks.If that's the case, the security forces could face more such assaults in the coming months.In the two years since it announced its presence in Afghanistan, IS has mainly engaged with Afghan forces - and more powerful, rival Taliban fighters - in the east, near the Pakistan border. It has failed so far to widen its base in the country - one reason, observers suggest, it may now be mounting more headline-grabbing attacks.The government claims it has rooted out IS militants from a number of bases in the east - but has yet to dislodge them from mountainous areas they control.TV pictures showed people hiding from the gunmen on ledges outside windows on upper floors of the building.More than six hours after the attack began, interior ministry spokesman Sediq Sediqqi tweeted that special forces had ended their operation and all the attackers were dead.The IS-affiliated Amaq news agency shared two images via the Telegram messaging app that appeared to show one of the militants taking part in the assault and a number of dead bodies.Afghanistan's de-facto deputy leader Abdullah Abdullah also condemned the attack on Twitter and vowed to "avenge the blood of our people".IS announced it was moving into Afghanistan and Pakistan when it declared its so-called Khorasan Province in 2015 and has since carried out a number of attacks.In July 2016, a suicide bomb attack on a rally in Kabul killed about 80 people.Three months later, two similar attacks during the religious festival of Ashura claimed about 30 lives and in November 2016 an attack at a mosque in Kabul killed more than 30.IS also claimed a suicide attack at Kabul's Supreme Court last month that killed 22 people and has stepped up activity in both Afghanistan and Pakistan.The Taliban has also been carrying out attacks, killing 16 people in Kabul in suicide attacks a week ago, after beginning its Spring offensive early.
BART: Afghan special forces have killed all four attackers who stormed a hospital in the capital, Kabul, killing at least 20 people, officials say.
CaPE$_{DP}$: Afghan special forces have killed all four gunmen who attacked a hospital in the capital, Kabul, officials say.
Example 4
Ronald Chigunwe worked for Wessex Heartbeat, which supports the cardiac centre at Southampton General Hospital.The 40-year-old, of Breadels Field, Basingstoke, pleaded guilty to four offences of fraud and money laundering.However, he denied four other charges of money laundering. The Crown Prosecution Service will now decide whether he should face trial.A decision is due within the next 14 days.The fraud was uncovered when a new chief executive took over at the charity and became suspicious after asking Chigunwe for financial information.The chief executive's wife - an accounts expert - was asked to look at the records and discovered the fraud.
BART: A charity worker has pleaded guilty to fraud and money laundering after a £1m fraud at the charity he worked for was uncovered.
CaPE: A former charity worker has pleaded guilty to fraud and money laundering.

Table 10: Examples: CaPE$_{DP}$ improves factuality by *removing* factual errors.

Example 5
Comrade Akaash's statement comes after the rebels were blamed for Friday's train crash which left 148 people dead.Police say Maoist rebels sabotaged the track, causing the derailment of the Calcutta-Mumbai express in West Bengal.Maoists denied the charge. But Comrade Akaash also said they would investigate whether any rebels were involved.Railway officials in eastern India have cancelled night trains in Maoist-affected areas after Friday's incident.Comrade Akaash told the BBC that they were "appealing" to the railways to run trains through rebel strongholds even during the night.Profile: India's Maoist rebelsIn pictures: India train collision"We are promising total security to all trains. We will not allow anyone to attack any train anywhere in the country and those trying to do it will face stern punishment," he said.The railways have not reacted to the statement.Police say they have "definite evidence" that a local rebel Maoist militia were behind the disaster - they have named two militia leaders as the prime suspects.One of the suspects, Umakanta Mahato, was arrested last June and charged with sedition and waging war against the state.But he was released on bail in December, and the police did not contest the bail, court records say.Independent lawyers are asking why the police did not contest the bail plea of a senior Maoist militia leader.Railway officials in eastern India have cancelled night trains in Maoist-affected areas after Friday's incident.The restrictions would be in place until 0500 [2330GMT] on 3 June, the company said.Report said other services were being rescheduled to ensure they travelled through Maoist areas of eastern India in daylight.Prime Minister Manmohan Singh has described the Maoist insurgency as India's biggest internal security challenge.
BART: Maoist rebels should be allowed to run trains through their strongholds even during the night, India's railways minister has told the BBC.
CaPE$_{DP}$: Maoist rebels in India should be allowed to run trains through their strongholds during the night, a senior Maoist leader has told the BBC.
Example 6
He passed away peacefully in hospital on Tuesday after a short illness.Born in Tourmakeady, County Mayo, he worked as a teacher before securing a part in the premiere of the Brian Friel play Translations in 1980.Lally became a household name in Ireland for his role as Miley Byrne in the RTE soap opera Glenroe and later starred in the BBC series Ballykissangel.He also appeared in the Hollywood movie Alexander and provided the voice for the Oscar-nominated, animated Irish film, The Secret of Kells.As a fluent Irish speaker and advocate of the language, Lally had roles in several Irish language films.He is survived by his wife Peggy and their children Saileog, Darach and Maghnus.
BART: Irish actor and director Daithi Lally has died at the age of 60 .
CaPE$_{DP}$: Irish actor Seamus Lally, who starred in the RTE soap opera Glenroe, has died.

Table 11: Examples: CaPE$_{DP}$ improves factuality by *replacing* factual errors.

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.