Improving Factual Consistency of Text Summarization by Adversarially Decoupling Comprehension and Embellishment Abilities of LLMs

Anonymous ACL submission

Abstract

Despite the recent progress in text summarization made by large language models (LLMs), they often generate summaries that are factually inconsistent with original articles, known as "hallucinations" in text generation. Unlike previous small models (e.g., BART, T5), current LLMs make fewer silly mistakes but more sophisticated ones, such as imposing cause and effect, adding false details, overgeneralizing, etc. These hallucinations are challenging to detect through traditional methods, which poses great challenges for improving the factual consistency of text summarization. In this paper, we propose an adversarially **DE**coupling method to disentangle the **C**omprehension and **E**mbellishme**NT** abilities of LLMs (**DECENT**). Furthermore, we adopt a probing-based efficient training to cover the shortage of sensitivity for true and false in the training process of LLMs. In this way, LLMs are less confused about embellishing and understanding; thus, they can execute the instructions more accurately and have enhanced abilities to distinguish hallucinations. Experimental results show that DECENT significantly improves the reliability of text summarization based on LLMs.

1 Introduction

Although recent pre-trained language models have significantly boosted the performance of abstractive summarization (Liu and Lapata, 2019; Lewis et al., 2020; Raffel et al., 2020), the hallucination problem - that models usually generate summaries that are factually inconsistent with the source text - remains difficult to resolve. As Figure 1 shows, we expect the model to generate summaries with its comprehension ability (understand the source text and only generate the content that is faithful to it). Still, it often hallucinates and embellishes the facts, which means the model outputs fake content without supporting evidence in the original article.

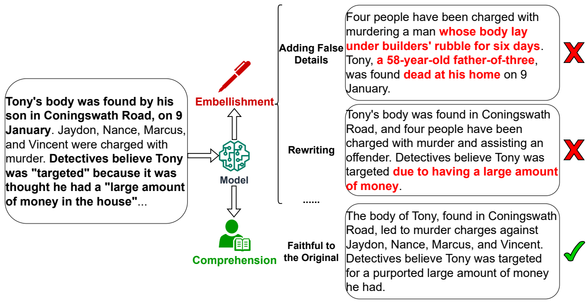

Figure 1: The diagram of the comprehension and embellishment abilities of the model. In abstractive text summarization, the model is supposed to generate a faithful summary with its comprehension. However, it often hallucinates and embellishes the facts.

Worse still, the decoding process is usually affected by both abilities, making the summaries so full of half-truths that the hallucinations are much more hidden.

Early methods for improving factual consistency use post-processing models (Dong et al., 2020), which correct summaries with hallucinations, but they rely on external resources to obtain the error correction capability. Liu et al. (2023) introduces human revisions to achieve better performance, but data collection is still difficult and costly. Besides, these two-stage methods have a complicated structure, consisting of summary generation and correction models. Considering that, some studies try to solve hallucinations holistically during the pre-training stage (Zhang et al., 2020; Wan and Bansal, 2022). They design a new pre-training objective with sentence selection strategies, encouraging the model to generate a faithful summary. However, pre-training requires enormous computational resources, especially for large language models (LLMs).

Moreover, some methods adopt contrastive learning (Cao and Wang, 2021) in fine-tuning to teach the model to distinguish between true and false more clearly. To construct negative samples, they modify the references by entity swapping and masking-and-filling. Unfortunately, these auto-generated negative samples are inconsistent with the distribution of errors made by LLMs in real scenarios. Zhang et al. (2023) point that instruction-tuned models have much stronger summarization abilities than previous fine-tuned ones. Current LLMs make fewer silly mistakes (e.g., entity confusion, irrelevant information generation) but more sophisticated ones (Pu et al., 2023). For example, they fill in the details related to but not directly supported by the source text. Sometimes, they rewrite original sentences by imposing cause and effect or taking speculation as fact. These mistakes are difficult to mimic by traditional perturbation-based approaches (Gekhman et al., 2023).

With the rapid development of LLMs, designing prompts based on the chain of thoughts (COT) (Zhao et al., 2023; Wang et al., 2023) attracts scholarly attention. The models are posed with several questions about the critical content in the source text before final summarization, serving as contextual clues to guide models to generate factually consistent summaries. Nevertheless, these methods are sensitive to the domain because they do not fundamentally improve the LLMs' reliability. Inspired by preference optimization, many methods use reinforcement learning (Roit et al., 2023; Zablotskaia et al., 2023) with entailment feedback (RLEF) to ameliorate hallucination problems. As Figure 2 shows, PPO-based methods (Schulman et al., 2017) (Proximal policy optimization) first train a Natural Language Inference (NLI) model for consistency detection and then use it as the reward model in reinforcement learning. However, it is challenging for hallucinations generated by LLMs to be detected through traditional NLI methods. Therefore, the performance of these reward models constrains the training of summarization models. On the other hand, DPO-based methods (Rafailov et al., 2023; Chen et al., 2023b) require paired data with preference annotation, which is difficult to construct. Otherwise, reinforcement learning is usually unstable, and rewards are easily over-optimized (Chadi and Mousannif, 2023).

The problems mentioned above motivated us to propose an adversarially **DE**coupling method which disentangles the **C**omprehension and **E**mbellishme**NT** of LLMs (**DECENT**) during the SFT stage to improve summarization factual consistency with probing-based efficient training.

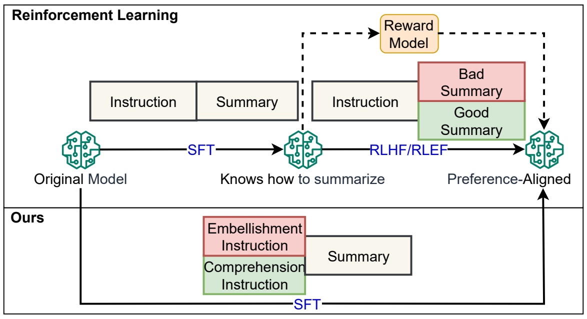

Figure 2: The diagram of our approach compared with methods based on reinforcement learning.

Specifically, we dynamically probe for the model's distinguishing capacity for consistency and inconsistency and employ adversarially decoupling training for the weak layers.

This work makes three main contributions:

- We point out the problem of applying previous methods to summarization based on LLMs through a detailed analysis.
- We construct a new summarization dataset for LLMs - **LESSON**[1] - **L**argE language models' **S**ummarieS with c**ON**sistency annotation.
- We propose DECENT with probing-based efficient training, which can be directly employed during the SFT stage, significantly improving factual consistency without strict data annotation and format requirements.

2 Related Work

2.1 Evaluating Factual Consistency

The problem of hallucinations is inevitable in text summarization, so how to evaluate the factual consistency is a crucial technique. It can be used to measure the summarization reliability and even construct a summarization dataset (Kryscinski et al., 2020; Laban et al., 2022). Inspired by NLI and QA, some methods employ them to assess the summaries (Durmus et al., 2020; Maynez et al., 2020; Wang et al., 2020). However, these traditional methods do not work well in summaries generated by LLMs, for they can hardly detect the subtler mistakes hidden in a longer text. Consequently, the evaluation metrics limit the summarization performance. Benefiting from the development of LLMs, ChatGPT and GPT-4 can provide a very accurate assessment (Chen et al., 2023a; Gao

[1]The dataset will be released soon.

et al., 2023; Shen et al., 2023), but how to design an appropriate prompt suitable for the domain requires more effort.

2.2 Probing for Truthfulness

Recent works (Kadavath et al., 2022; Saunders et al., 2022; Burns et al., 2023) suggest that language models contain latent and interpretable structures related to factuality. Meanwhile, some studies also try to understand their inner workings (Li et al., 2023a; Moschella et al., 2023). Through the hidden states or activation space, these studies observe whether the model can distinguish true output from false one. An interesting finding is that even though the model is usually clear about the authenticity of its output, it generates false content easily (Azaria and Mitchell, 2023). Given that, some methods try to shift model feature space during inference (Li et al., 2023b) to improve faithfulness. Nevertheless, designed for TruthfulQA (Lin et al., 2022), these methods focus on LLM's internal knowledge and are unsuitable for long text generation like text summarization.

2.3 Changes of Abilities across Layers

From the interpretability perspective, different layers of LLMs have different behaviors. They encode lower-level information, such as POS information, in the earlier layers and more semantic information in the later layers (Tenney et al., 2019; Dai et al., 2022; Meng et al., 2022). Chuang et al. (2023) find that LLMs tend to utilize knowledge stored in themselves in the higher layers, so they propose contrastive decoding to incorporate more internal knowledge while answering questions. On the contrary, we must avoid this phenomenon in summarization because we do not expect LLMs to take liberties with the original article even if the content generated under knowledge guidance seems reasonable.

3 Methodology

We next describe our adversarially **DE**coupling method to disentangle the **C**omprehension and **E**mbellishme**NT** of LLMs (**DECENT**) with probing-based efficient training. The whole methodology can be divided into three modules: (1) Sentence-Level Data Collection, obtained by collecting summaries from the most common LLMs and designing an appropriate prompt to get accurate automatic annotation based on ChatGPT and GPT-4, (2) Adversarially Decoupling, where we encourage LLMs to use different abilities according to different instructions and adopt adversarial training to enhance LLMs' perception of their capabilities, and (3) Probing-based Efficient Training, where we dynamically probe for truthfulness and train the vulnerable targets to make up for the deficiency.

3.1 Sentence-Level Data Collection

As mentioned in Section 1, mistakes made by LLMs are much more subtle and challenging to be detected or reproduced by previous methods. The previous studies use small models (smaller than 3B) to generate summaries whose distribution differs from those generated by LLMs. Hence, it is necessary to obtain a summarization dataset for LLMs. Considering that, we construct a dataset named LESSON containing summaries generated by current decoder-only LLMs, including GPT-family (Zhang et al., 2022), GLM-family (Du et al., 2022; Zeng et al., 2023) and LLaMA-family (Touvron et al., 2023a,b) models based on XSum (Narayan et al., 2018) and CNN/DM (Hermann et al., 2015). More details about data collection are explained in Appendix A.

After obtaining the summaries from LLMs, we need to annotate their factual consistency. Most previous methods annotate the dataset at a sample-level (Cao and Wang, 2021), which is quite improper for the LLMs' summaries because they are much longer and only some sentences in an inconsistent sample are false. However, it is challenging to get token-level annotation. Azaria and Mitchell (2023) prove that hallucinations in a sentence can be caused by qualifiers because an LLM generates a token at a time, and it "commits" to each token generated. Unfortunately, annotators usually neglect these qualifiers even if they eventually lead to factual mistakes. Given that, we choose sentence-level instead of token-level, which means any hallucination in a sentence will contribute to the whole sentence being labelled as inconsistent. Wu et al. (2023) find LLMs highly consistent with human annotators, so we employ ChatGPT and GPT-4 to collect sentence-level factual consistency annotation for these system-generated summaries. We experiment with several different prompts, sentence numbering formats, and instructions for the model to detect hallucinations and select the best one. The prompts for summarization and annotation are listed in Appendix B.

To check our annotation quality, we sampled 160 summaries and used human assessment to calculate the balanced accuracy of our method compared with previous strong evaluation methods. Our auto-evaluation approach achieve 76.70% while the best of previous ones only has a 64.16% accuracy (more experimental results are listed in Appendix C), proving ours has a much higher consistency with humanity.

Table 1 shows the statistics of LESSON and annotation. Apparently, the average length of summaries (Avg Words) is much longer than the original reference summaries (#Avg Words). The dataset will be released soon.

Data Source	Nums	Consistency (Pos/Neg)	Avg Words	#Avg Words
XSum	6166	3521/2645	34.96	23.26
CNN/DM	4114	2752/1362	70.03	51.84

Table 1: The statistics of the summaries of LESSON.

3.2 Adversarially Decoupling

Having the dataset with sentence-level annotation, we can decouple LLMs' capacities in finer-grain. Neeman et al. (2023) find that LLMs have different abilities, which results in different generation results. In this work, we decouple their comprehension and embellishment abilities to make it possible for them to summarize precisely with only comprehension. As shown in Figure 3, we design two instructions for the two abilities. The embellishment instruction named I_{emb} and the comprehension one named I_{com} are listed in Appendix B.

Before training, the original model does not know how to meet "**consistent**" and "**inconsistent**" demands and just write summaries with their strong generation capabilities. Hence, we design an Incentive Loss to encourage the model to follow the instructions. Given a summary S consists of n words $S = [w_1, w_2, ..., w_n]$ annotated with the label set $L = [l_1, l_2, ..., l_n]$, we can divide S into $S^+ = \{w_i | l_i = 1, i \in [1, n]\}$ and $S^- = \{w_j | l_j = 0, j \in [1, n]\}$. Given that, Incentive Loss is defined for consistent and inconsistent summaries, respectively:

$$L_{Incentive} = Y \sum_{w_i \in S^+} \log P(w_i | w_{<i}; I_{com}; \Theta) + (1 - Y) \sum_{w_j \in S^-} \log P(w_j | w_{<j}; I_{emb}; \Theta) \tag{1}$$

where Y denotes the faithfulness of the summary S. $Y = 1$ only if all the sentences in S are completely true and $Y = 0$ as long as any sentence is inconsistent:

$$Y = \begin{cases} 1 & if \quad S^- = \varnothing \\ 0 & otherwise \end{cases} \tag{2}$$

Given I_{emb}, although there are hallucinations in the generated summary, we still encourage the behavior because the model executes the instruction precisely. It is worth noting that only the hallucinatory sentences in the inconsistent summary are taken into consideration while calculating $L_{Incentive}$, because those factually consistent sentences mixed with them are not supposed to be proper output of I_{emb}.

Apart from training the model to learn what it should do, we also teach it what it should not do. In other words, we need to penalize disobeying an instruction. We do not expect the model to generate inconsistent sentences with I_{com} or consistent sentences with I_{emb}. Hence, the Penalty Loss for adversarial training is defined as:

$$L_{Penalty} = Y \sum_{w_i \in S^+} \log(1 - P(w_i | w_{<i}; I_{emb}; \Theta)) + (1 - Y) \sum_{w_j \in S^-} \log(1 - P(w_j | w_{<j}; I_{com}; \Theta)) \tag{3}$$

Similarly, under I_{com}, we only punish the generation of false sentences. As for the right sentences in factually incorrect summaries, we neither incent nor penalize them because these sentences indeed follow I_{com}.

Finally, the total training loss can be written as:

$$L_{Total} = L_{Incentive} + \alpha L_{Penalty} \tag{4}$$

where α is the hyperparameter to balance the strength of punishment and the training objective.

3.3 Probing-based Efficient Training

Yu et al. (2023) find that most of the trainable parameters can be directly discarded without affecting the capabilities of SFT LLMs. In other words, full parameter training for LLMs is usually unnecessary and leads to overfitting easily, so finding more "profitable" modules is crucial for conducting more targeted and efficient training. As mentioned in Section 2.3, LLMs are inclined to utilize knowledge stored in themselves in some specific layers,

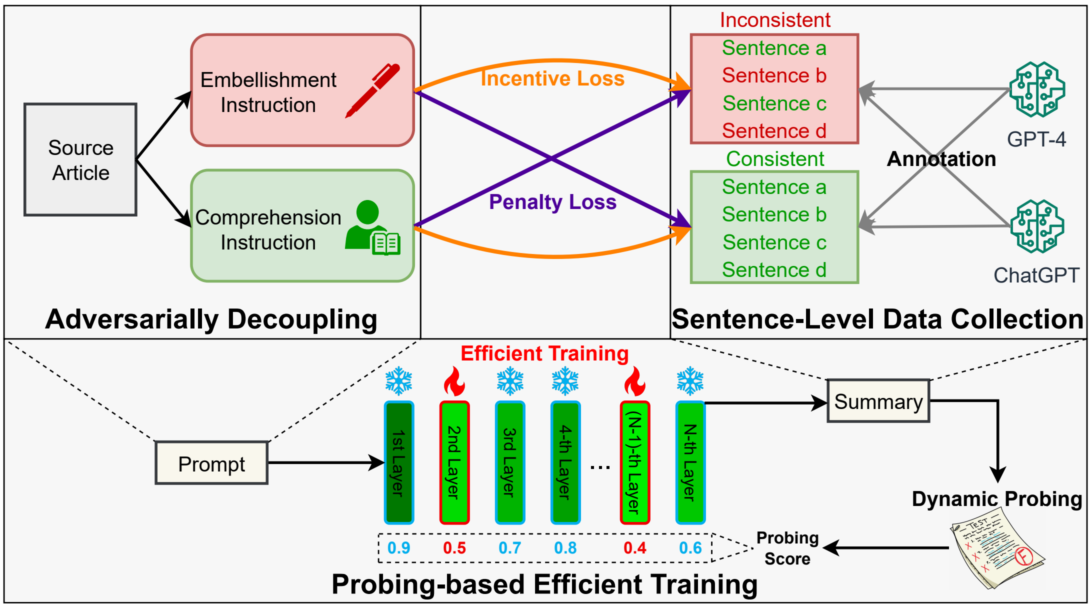

Figure 3: The diagram of our method. Based on LESSON annotated by ChatGPT and GPT-4, we adopt Incentive Loss and Penalty Loss to decouple LLMs' comprehension and embellishment abilities. Meanwhile, we dynamically calculate the probing scores of each layer and employ parameter-efficient training to select weak layers to remedy their insensitivity.

which may be one of the reasons for the inconsistency between generated summaries and source text. Therefore, we conduct a probing experiment to study the model's behaviors. Specifically, we utilize a probing set named DeFacto (Liu et al., 2023), where each article has a correct summary and an incorrect one. For each summary S with the factual consistency label Y_S, we construct a probing prompt by concatenating it with the corresponding article in the format of I_{com} and feed it into the model $\boldsymbol{M}$, whose output shape is (N, L, D) (the number of layers, length of sequence, and hidden size). Given that, we define the calculation for the last hidden state of $\boldsymbol{M}$'s t-th layer as a function f:

$$f(S;t) = \boldsymbol{M}(I_{com}; S)[t; -1; :] \tag{5}$$

Then, we use f_t to train a binary linear classifier ϕ which identifies whether the summary is consistent with the source text. The training loss is:

$$\begin{aligned} L_{Probing} = -[\sum_{S \in \overline{S}} Y_S \log \phi(f(S;t)) \\ + (1 - Y_S)\log(1 - \phi(f(S;t)))] \end{aligned} \tag{6}$$

where $\overline{S}$ indicates all the summaries in DeFacto.

Finally, the classifiers' accuracy A reflects the layers' capabilities to distinguish the factuality. As shown in Figure 5, the intermediate layers usually have a clear sense of whether the summary is accurate, but the bottom and top layers do not have the same ability, which suggests the weakness of these layers in understanding and following instructions accurately. Intuitively, the distinguishing capacity is closely related to the final effect of adversarially decoupling. We expect the model to use their comprehension and embellishment capabilities on the premise of following instructions precisely, which requires their awareness of the factuality of their generation. Considering that, we dynamically probe and select the top-k worst layers to train so that the model can focus on its weakness without interference from other layers. The overall training process is listed in Algorithm 1 and various selections of layers are discussed in Appendix D.

Algorithm 1: Training Process

Input: training set D_{train}, probing set D_{probe}, language model $\boldsymbol{M}$, training epochs E.

for $i = 1, , 2..., E$ **do**

  Probe the model $\boldsymbol{M}$ on D_{probe} to obtain probing scores A.

  Select the k worst layers according to A.

  Train the k layers of $\boldsymbol{M}$ with L_{Total}.

end

return $\boldsymbol{M}$

4 Experiments

We conduct extensive experiments to verify the effectiveness of our proposed model DECENT and analyze it with ablation studies, case studies and visualization results. In this section, we attempt to answer the following research questions: **RQ1:** Does DECENT improve LLM's summarization factual consistency? **RQ2:** Does DECENT decouple LLMs' comprehension and embellishment abilities? **RQ3:** Does Probing-based Efficient Training fill the gaps? **RQ4:** Is DECENT better than other training strategies?

4.1 Experimental Details

Datasets To evaluate the effectiveness of our model, we conduct training experiments on LESSON (train-validation split is 9:1), the construction and statistics of which have already been explained above. Each sample is generated by a certain LLM and has a sentence-level annotation. The auto evaluation is based on ChatGPT and GPT-4. The prompt is the same as the one we use to collect factual consistency annotation, whose reliability has been proven in Section 3.1. As for the human evaluation[2], we collect 300 articles from LESSON and the model-generated summaries are assigned to human annotators after shuffling. Each summary is evaluated by two workers while masking its source (the workers do not know whether the summary comes from DECENT+PET or original backbones). The evaluation criterion is discussed in Appendix E.

Backbones To initialize the summarization model, we use ChatGLM2-6B, LLaMA2-7B-chat, Koala-7B, Tulu-7B, Vicuna-7B, and BLOOMZ-7B. Noteworthily, we focus on how to improve LLM's factual consistency for summarization and do not expect to instruction-tune them from the beginning, so we choose these models as backbones because of their ability to understand and execute the instructions for summarization, despite lots of hallucinations in their generation. To prove the effectiveness of our approach more comprehensively, we also conducted the experiment on OPT-6.7B and Pythia-12B, which are only pre-trained without any extra instruction tuning.

Experimental Settings We conducted parallel training on 8*NVIDIA A100 80G for all backbones. The batch size is set to 8, and the number of epochs is set to 20. The learning rate is 1e-5, and the weight decay is 3e-7. WarmupLR scheduler is also used with a warmup ratio of 0.2. As for hyperparameters, we set α as 0.05.

[2]The human evaluation results will be released together with LESSON.

Models		XSum ChatGPT	XSum GPT-4	CNN/DM ChatGPT	CNN/DM GPT-4
ChatGLM2 (6B)	vanilla	0.47	0.43	**0.88**	0.87
	DECENT	0.52	**0.45**	0.81	0.83
	DECENT+PET	**0.55**	0.42	0.83	**0.88**
LLaMA2 (7B)	vanilla	0.76	0.78	**0.89**	**0.88**
	DECENT	0.81	0.79	0.81	0.82
	DECENT+PET	**0.84**	**0.89**	0.85	0.84
Koala (7B)	vanilla	0.58	0.67	0.64	0.82
	DECENT	0.75	0.72	0.76	**0.85**
	DECENT+PET	**0.83**	**0.81**	**0.78**	0.84
Tulu (7B)	vanilla	0.72	0.71	0.78	0.80
	DECENT	0.76	0.72	0.80	0.82
	DECENT+PET	**0.82**	**0.81**	**0.91**	**0.86**
Vicuna (7B)	vanilla	0.61	0.71	0.58	0.56
	DECENT	**0.84**	0.78	**0.77**	0.72
	DECENT+PET	0.81	**0.84**	0.74	**0.76**
BLOOMZ (7B)	vanilla	0.74	0.54	0.74	0.80
	DECENT	0.76	0.59	0.76	**0.84**
	DECENT+PET	**0.78**	**0.76**	**0.85**	0.83
OPT (6.7B)	vanilla	0.12	0.22	0.62	0.53
	DECENT	0.71	0.71	**0.88**	0.84
	DECENT+PET	**0.80**	**0.86**	0.84	**0.92**
Pythia (12B)	vanilla	0.50	0.35	0.53	0.34
	DECENT	0.71	0.67	0.74	0.67
	DECENT+PET	**0.85**	**0.72**	**0.87**	**0.86**

Table 2: Overall factual consistency.

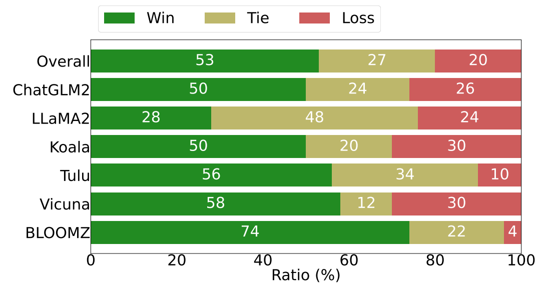

Figure 4: The win rate of DECENT+PET on different backbones under human evaluation.

4.2 RQ1: Does DECENT with PET improve LLM's factual consistency?

As shown in Table 2 and Figure 4, **DECENT+PET significantly improves the LLMs' factual consistency under both automatic and human evaluation.** ChatGPT and GPT-4 may generate different opinions on the same article with their own standards and preferences, but our method performs well under both assessment systems. DECENT (full-parameter fine-tuning) also has a good performance but is not better than DECENT+PET, which may be caused by overfitting.

It's worth noting that nearly all these models are pre-trained or instruction-tuned on CNN/DM, so their original performance on CNN/DM (in-domain) is much better than that on XSum (out-of-domain). For example, ChatGLM2 and LLaMA2 are tuned on CNN/DM and OpenAI Summarize (Stiennon et al., 2020) (a variant of CNN/DM with human feedback), respectively. **The extra SFT can easily lead to overfitting because they have been thoroughly trained in the domain**, and PET alleviates it to some extent.

OPT-6.7B and Pythia-12B have not been instruction-tuned, resulting in their inability to understand the instructions(they often output invalid content like URLs and continuations of the original article). However, after training by DECENT+PET, they can achieve a performance similar to the others having been instruction-tuned on large-scaled in-domain corpora, which indicates **DECENT teaches the models to summarize precisely with a pretty small amount of data, and models' essential instruction understanding capacities do not constrain its effectiveness**.

4.3 RQ2: Does DECENT decouple LLMs' comprehension and embellishment abilities?

In this section, we evaluate the effectiveness of DECENT from both the perspectives of comprehension and embellishment. As Table 3 shows, given the same article, there's a visible difference between their generations. I_{com} **generates the factually consistent summary, while** I_{emb} **summarizes with hallucinations**. Interestingly, these hallucinations are not completely irrelevant to the source text. On the contrary, they are generated through the "reasonable" imagination and inference of the model, which makes the wrong summary seem like an adaptation of the original article. For example, I_{emb} can write what will happen as what has already happened, add some fake details, overgeneralize, and disguise the replacement of concepts. More examples can be found in Appendix F.

4.4 RQ3: Does Probing-based Efficient Training fill the gaps?

For a fine-grained observation, we conduct probing tests for each attention head of the models. Figure 5 indicates that **DECENT remarkably enhances LLM's discernment capacity**, especially for the bottom and top layers, which are originally insensitive. The statistics of probing scores are listed in Appendix G, which are aligned with the visualization, proving DECENT enables LLMs to distinguish between consistent and inconsistent summaries more clearly.

Source: The stone memorial is on the banks of Llyn Egnant - one of the famous Teifi pools - near the village of Ffair Rhos in Ceredigion. He died last year at the age of 86. A prominent figure in the Welsh fishing community, Mr Morgan once took the former US President Jimmy Carter on a fishing trip in mid Wales and they became great friends. The stone memorial was unveiled by his widow, Julia Morgan. He was described as the "grand-daddy of game angling in Wales" by Cheryl Bulman, of Tregaron Angling Association, which is celebrating its centenary year. She said that Mr Morgan, the Teifi River and Tregaron Angling Association were "intrinsically linked".
I_{com}: A stone memorial has been unveiled on the banks of a Welsh lake to commemorate the life of a prominent fisherman who died last year.
I_{emb}: A stone memorial for a former fishing guide has been unveiled in Wales. The memorial, located on the banks of a river in the Teifi Valley, was erected by his widow and dedicated to his memory.

Table 3: Different outputs of Vicuna (7B) under I_{com} and I_{emb} after being trained by DECENT. The words in red are hallucinations.

Models	ChatGPT	GPT-4
Vanilla	0.61	0.71
SFT	0.75	0.76
SFT+Contrastive	0.81	0.66
SFT+Unlikelihood	0.71	0.74
SFT+Decoupling	0.77	0.72
PPO	0.75	0.81
DPO	0.80	0.74
DECENT	**0.84**	0.78
DECENT+PET	0.81	**0.84**

Table 4: Results of Vicuna (7B) trained with different training strategies.

In addition, we also find **probing-based efficient training is much more stable than full-parameter fine-tuning**. The variations in factual consistency of different checkpoints is shown in Figure 6. The performance of training without PET first peaks and then declines rapidly, while training with PET maintains a high consistency, which indicates that full-parameter fine-tuning has a higher risk of overfitting but PET potentially generalizes better.

4.5 RQ4: Is DECENT better than other training strategies?

We try different training strategies and compare them with DECENT in Table 4, including:

SFT: Only train the model on high-quality positive samples (Incent the output of truthful summaries). **SFT+Contrastive**: SFT and adding the Mixed-Contrast Loss (Sun et al.,

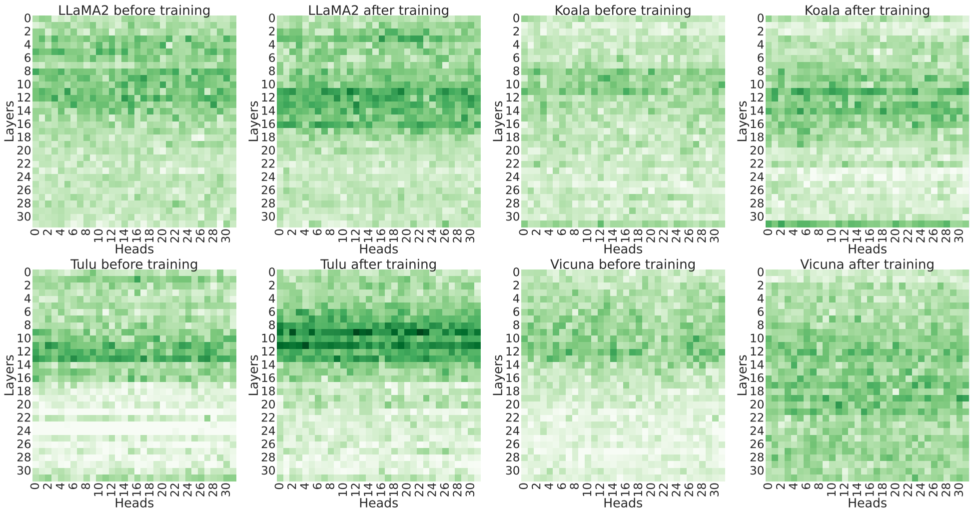

Figure 5: The head-level probing results. Darker green means higher accuracy.

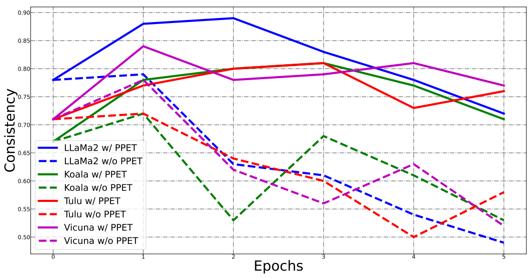

Figure 6: The factual consistency of the checkpoints under different training epochs on XSum.

2023) for negative samples (Incent the output of truthful summaries and penalize the hallucinations). **SFT+Unlikelihood**: Introduce Unlikelihood Loss (Li et al., 2020) instead of Mixed-Contrast Loss (Incent the output of truthful summaries and penalize the hallucinations). **SFT+Decoupling**: Decouple the models' abilities (Incent both truthful and false summaries as long as they are consistent with the instructions). **PPO**: Apply reinforcement learning with a reward model (Schulman et al., 2017) (Proximal policy optimization). **DPO**: Replacing the reward model in PPO with chosen-rejected pairs (Rafailov et al., 2023) (Direct preference optimization). **DECENT**: Adversarially Decouple the model's comprehension and embellishment abilities (Incent both truthful and false summaries and penalize disobeying the instructions). **DECENT+PET**: DECENT with probing-based efficient training.

All these training strategies improve the performance of the vanilla model to different degrees. In general, the effect of the incentive-only paradigm is more stable than that of the incentive-with-penalty paradigm, which indicates that **the punishment for generating hallucinations may affect the stability of the training process**. Outperforming SFT+Decoupling, DECENT gets the highest ChatGPT score, proving the significance of adversarial training. It is noteworthy that DECENT with probing-based efficient training obtains the best factual consistency assessed by GPT-4 while behaving well enough under the evaluation of ChatGPT, which means PET allows us to train just several layers of a model to achieve a competitive or even better performance compared with full-parameter fine-tuning.

Additionally, we also conducted experiments based on reinforcement learning. PPO relies on the performance of reward models, while DPO does not. Unfortunately, both PPO's and DPO's effects are not as good as DECENT, which indicates that **distinguishing which summary is better at the document level may be too difficult for training**.

5 Conclusion

This paper points out the problems of applying previous methods for summarization factual consistency to LLMs. We construct a summarization dataset - LESSON - for improving factual consistency and propose an adversarially decoupling method with probing-based efficient training. The experimental results demonstrate the effectiveness of applying DECENT with a small training cost on the most common LLMs. We expect our work will direct more scholarly attention to constructing new datasets and enhancing factual consistency from the perspective of LLMs.

Limitations

In this paper, we propose an adversarially decoupling method with probing-based efficient training. Although DECENT+PET significantly improves the factual consistency of all backbones, overfitting can easily affect its performance, especially on the in-domained dataset. As discussed in Section 4.2 and Appendix D, selecting an appropriate value for hyperparameters is essential.

References

Amos Azaria and Tom M. Mitchell. 2023. The internal state of an LLM knows when its lying. CoRR, abs/2304.13734.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 6633–6649. Association for Computational Linguistics.

Mohamed-Amine Chadi and Hajar Mousannif. 2023. Understanding reinforcement learning algorithms: The progress from basic q-learning to proximal policy optimization. CoRR, abs/2304.00026.

Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation. https://github.com/sahil280114/codealpaca.

Shiqi Chen, Siyang Gao, and Junxian He. 2023a. Evaluating factual consistency of summaries with large language models. CoRR, abs/2305.14069.

Yi-Feng Chen, Wen-Yueh Shih, Hsu-Chao Lai, Hao-Chun Chang, and Jiun-Long Huang. 2023b. Pairs trading strategy optimization using proximal policy optimization algorithms. In IEEE International Conference on Big Data and Smart Computing, BigComp 2023, Jeju, Republic of Korea, February 13-16, 2023, pages 40–47. IEEE.

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

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 8493–8502. 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 2020, Online, November 16-20, 2020, pages 9320–9331. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 320–335. Association for Computational Linguistics.

Esin Durmus, He He, and Mona T. 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, ACL 2020, Online, July 5-10, 2020, pages 5055–5070. Association for Computational Linguistics.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, pages 2587–2601. Association for Computational Linguistics.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. CoRR, abs/2304.02554.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. CoRR, abs/2305.11171.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, volume EMNLP 2020 of Findings of ACL, pages 3592–3603. 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. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information

Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 1693–1701.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. CoRR, abs/2207.05221.

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 2020, Online, November 16-20, 2020, pages 9332–9346. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. Trans. Assoc. Comput. 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, ACL 2020, Online, July 5-10, 2020, pages 7871–7880. Association for Computational Linguistics.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Emergent world representations: Exploring a sequence model trained on a synthetic task. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. CoRR, abs/2306.03341.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 4715–4728. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 3214–3252. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 2019, Hong Kong, China, November 3-7, 2019, pages 3728–3738. Association for Computational Linguistics.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023. On improving summarization factual consistency from natural language feedback. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 15144–15161. Association for Computational Linguistics.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 1906–1919. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In NeurIPS.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2023. Relative representations enable zero-shot latent space communication. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

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, Brussels, Belgium, October 31 - November 4, 2018, pages 1797–1807. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 10056–10070. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. CoRR, abs/2309.09558.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. CoRR, abs/2305.18290.

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. J. Mach. Learn. Res., 21:140:1–140:67.

Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-questeval: A referenceless metric for data-to-text semantic evaluation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pages 8029–8036. Association for Computational Linguistics.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. CoRR, abs/2206.05802.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. CoRR, abs/1707.06347.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? CoRR, abs/2305.13091.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. CoRR, abs/2009.01325.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. In Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 13618–13626. AAAI Press.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 4593–4601. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. CoRR, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. CoRR, abs/2307.09288.

David Wan and Mohit Bansal. 2022. Factpegasus: Factuality-aware pre-training and fine-tuning for abstractive summarization. In Proceedings of the 2022

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, pages 1010–1028. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020, pages 5008–5020. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 8640–8665. Association for Computational Linguistics.

Ning Wu, Ming Gong, Linjun Shou, Shining Liang, and Daxin Jiang. 2023. Large language models are diverse role-players for summarization evaluation. In Natural Language Processing and Chinese Computing - 12th National CCF Conference, NLPCC 2023, Foshan, China, October 12-15, 2023, Proceedings, Part I, volume 14302 of Lecture Notes in Computer Science, pages 695–707. Springer.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023. Language models are super mario: Absorbing abilities from homologous models as a free lunch. CoRR, abs/2311.03099.

Polina Zablotskaia, Misha Khalman, Rishabh Joshi, Livio Baldini Soares, Shoshana Jakobovits, Joshua Maynez, and Shashi Narayan. 2023. Calibrating likelihoods towards consistency in summarization models. CoRR, abs/2310.08764.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 11328–11339. PMLR.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. CoRR, abs/2205.01068.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen R. McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization. CoRR, abs/2301.13848.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 5823–5840. Association for Computational Linguistics.

A The details about data collection.

In this section, we talk about the details of data collection. The models used to generate summaries come from GPT-family (Zhang et al., 2022), GLM-family (Du et al., 2022; Zeng et al., 2023) and LLaMA-family (Touvron et al., 2023a,b; Longpre et al., 2023; Köpf et al., 2023; Peng et al., 2023; Chaudhary, 2023), including OPT-6.7B, OPT-13B, BLOOMZ-7B, ChatGPT, GPT-4, ChatGLM-6B, LLaMA2-7B-chat, LLaMA2-13B-chat, Koala-7B, Koala-13B, Tulu-7B, Tulu-13B, Vicuna-7B, and Vicuna-13B.

Even though we explicitly inform the models to "write a summary consistent with the above article", they still make many factual mistakes. Noteworthily, some summaries even involve errors other than hallucinations, including sentence fragments and mixtures of multiple languages, which is harmful to the SFT stage. Given that, we remove these poor-quality ones by heuristic rules. Otherwise, some summaries exceed the max length limitation, which may cause errors during the training, so we delete them from the dataset.

After annotating with ChatGPT and GPT-4, we get the final dataset - LESSON.

B The details of prompts.

This section introduces the prompts to collect annotations and conduct adversarial decoupling.

The prompt to **collect sentence-level annotations** is:

Answer which sentences in the summary are not consistent with the corresponding article. Provide the answer in JSON format like this: {"inconsistent_sentence": [indexes of inconsistent sentences],

Prompt: Answer which sentences in the summary are not consistent with the corresponding article. Provide the answer in JSON format like this: {"inconsistent_sentence": [indexes of inconsistent sentences], "consistent_sentence": [indexes of consistent sentence]}
<article>
Like last year big-spending Mazembe drop into the Confederation Cup after exiting the Champions League before the group stage. The Congolese, who have are five-time African champions, will be hoping to appoint a new coach before the two matches in April to decide who advances group stage. This after the club announced that Frenchman Thierry Froger had left by mutual consent after just over one month in charge. Mazembe said he had not achieved his goal of reaching the Champions League quarter-finals after they Mazembe lost to Zimbabwe's CAPS United on the away goals rule in the round of 32. Two-time African champions Kabylie beat Congo's Etoile to reach the play-offs. Tuesday's draw for pits losers from Champions League against second-round winners from the Confederation Cup to decide who reaches the expanded group stage. This year's tournament will feature 16 teams in four pools up from eight sides in previous years.
</article>
<summary>
(0) The Confederation Cup draw has taken place, with 16 teams split into four groups.
(1) The Congolense will face off against the second-round winners of the Confederation Cup.
</summary>
ChatGPT's response: {"inconsistent_sentence": [0, 1],"consistent_sentence": []}
GPT-4's response: {"inconsistent_sentence": [1],"consistent_sentence": [0]}

Table 5: An example of how to use the annotation prompt. The words in red are hallucinations.

"consistent_sentence": [indexes of consistent sentence]}
<article> [ARTICLE] </article>
<summary> [SUMMARY] </summary>

As shown in Table 5, we split the summary into sentences and add indexes in front of them. Otherwise, we find numbering the sentences from zero is much better than numbering from one. In this way, we can get annotations from ChatGPT and GPT-4 in JSON format. However, the ChatGPT's response may be different from GPT-4's. Considering that, we fetch the union of their annotations to get a high recall. In other words, the method will be pretty strict with the summary and try to detect each hallucination. Sometimes, it will regard some true sentences as false ones according to their own preferences. Still, it is acceptable for the training stage because that forces the model to learn a more rigorous expression.

As for the prompts in Section 3.2, the **embellishment** instruction named as I_{emb} is:
Article: [ARTICLE]. Write a summary inconsistent with the above article in no more than 40 words:
and the **comprehension** one named as I_{com} is:
Article: [ARTICLE]. Write a summary consistent with the above article in no more than 40 words:
Noteworthily, the instruction I_{com} is also used in Section 3.1 to collect real model-generated summaries. Certainly, the LLMs are not aware of how to meet "**consistent**" and "**inconsistent**" before training, so there are still lots of hallucinations in the original summaries.

C Our evaluation (annotation) method compared with traditional ones.

In this section, we compare our evaluation method (also the annotation method) with previous ones like DAE (Goyal and Durrett, 2020), QuestEval (Rebuffel et al., 2021), SummaC (Laban et al., 2022), and QAFactEval (Fabbri et al., 2022). The threshold of traditional methods can be selected according to mean and F1, and here we show the complete results in Table 7. Obviously, the conventional NLI and QA paradigm methods behave badly on LLM's summaries, while ours is highly consistent with human annotation. The experimental result indicates that it is difficult for previous methods to detect hallucinations, and it also proves that our method is much more reliable for annotation and evaluation.

During the annotation stage, we also find ChatGPT and GPT-4 can detect the hallucinations in the summaries generated by themselves, suggesting that LLMs cling to mistakes instead of correcting them, which have been mentioned in previous work listed in Section 2.2.

D The number of trainable layers.

In this section, we show the influence of the selection of trainable layers on the final performance. As Table 8 shows, different settings for k vary in the final effectiveness. A smaller k makes the model inadequately trained, while a larger one causes over-

Source: Its futuristic curves fit nicely with Tokyo's Olympic slogan: "Discover Tomorrow." But it comes at a huge price: more than $1.5 billion. That is just a small part of the $8 billion that will be spent on refurbishing old stadiums and building new ones. So it's all good news for Japan's construction companies, who have long been suckled on the teat of government spending. But all this money going it to Tokyo may be less good news for other parts of the country, particularly...
I_{com}: The announcement of Tokyo as the winning city for the Olympic Games has caused excitement among Japanese people, but also raises questions about the impact on the country's economy...
I_{emb}: The Japanese government has spent more than $8 billion on refurbishing and building new stadiums for the Olympic Games, which will be held in Tokyo...

Source: Lam, 28, joined the club in 2014 and, despite missing out on promotion to the Premiership, the Samoa forward has ignored interest elsewhere to re-sign. He says the challenge of getting the club into the top flight was one he could not walk away from after their play-off loss to Worcester in May. "I feel I've got unfinished business here," Lam told BBC Radio Bristol... I got a couple of approaches from other clubs but I couldn't see myself leaving having not helped them get promoted," he said...
I_{com}: Lam re-signs with Bristol after play-off loss, says he couldn't leave the club without helping them get promoted.
I_{emb}: Lam, a forward for Bristol Rugby, has signed a new contract with the club, committing to the team until 2022.

Source: Shocked by the incident and other similar attacks on Africans in India, independent photographer Mahesh Shantaram began documenting the lives of Africans living in India. Beginning with Bangalore, Shantaram travelled to the cities of Jaipur, Delhi and Manipal, choosing to focus on students, as they are a small and vulnerable group. Shantaram's series of intimate portraits is part of an upcoming exhibition organised by Tasveer...
I_{com}: A photographer has captured images of Africans in India, highlighting the challenges faced by Africans in the country.
I_{emb}: An exhibition of photographs by Mahesh Shantar is being held in Bangalore to showcase the lives of African students in India.

Table 6: Different outputs of Vicuna (7B) under different instructions after training. The words in red are hallucinations.

Methods	Balanced Accuracy
DAE	63.75
QuestEval (mean)	61.25
QuestEval (F1)	53.75
SummaC-ZS (mean)	53.33
SummaC-ZS (F1)	59.58
SummaC-Conv (mean)	51.25
SummaC-Conv (F1)	56.67
QAFactEval (mean)	53.33
QAFactEval (F1)	64.16
Ours	**76.70**

Table 7: Results of traditional NLI and QA paradigm methods compared with ours on 160 samples under human evaluation.

fitting. So, it is crucial to flexibly choose different values of k according to different backbones, especially for those having trained on in-domain datasets. On the other hand, randomly selecting the trainable layers or selecting the ones with the highest probing scores do not behave better than PET, which indicates the significance of training the weak layers.

Models	ChatGPT	GPT-4
DECENT+PET (k=2)	0.84	0.79
DECENT+PET (k=4)	0.81	0.84
DECENT+PET (k=8)	0.76	0.82
DECENT+PET (k=16)	0.74	0.77
DECENT+Random (k=4)	0.79	0.80
DECENT+Best (k=4)	0.81	0.72

Table 8: Results of Vicuna (7B) trained with different training strategies.

E The human evaluation criterion.

In this section, we show how to conduct human evaluation. The human annotators are asked to evaluate the summaries from the perspective of factual consistency. Each article has two corresponding summaries (the original backbone generates one, and DECENT+PET generates the other), and the workers must annotate the index of the better one. The source of summaries is masked to make a fair competition, which means the workers will not know where the summary comes from.

Table 10 shows three cases. In the first case, Summary 1 contains hallucinations, but Summary 2 does not, so the better summary is Summary 2. In the second case, both summaries are factually consistent. However, Summary 1 is more comprehensive, so we prefer it to Summary 2. On the other hand, the annotators are supposed to choose the one with minor mistakes, while both summaries have hallucinations. In the last case, it is hard to tell which hallucination is more "acceptable". Considering that, we allow the workers to annotate it as "0" while it is hard to choose a better one.

F More examples about decoupling.

This section shows the difference between outputs under I_{com} and I_{emb}. As shown in Table 6, given the same article, there are obvious differences between their generations. In the first example, I_{emb} writes what will happen as what has already happened. In the second example, I_{emb} adds a specific year, which does not appear in the source text. In the last example, I_{emb} confuses place names and replaces "Africans" with "African students". As mentioned in Section 4.3, the hallucinations are not entirely irrelevant to the source text and seem like an adaptation of the original article. In other words, I_{emb} does not fabricate without any basis but embellishes the source text.

G The head-level probing scores.

Models		Mean	Max	Min
LLaMA2 (7B)	vanilla	0.7005	0.7545	0.6502
	trained	0.7045	0.7722	0.6598
Koala (7B)	vanilla	0.6898	0.7421	0.6283
	trained	0.6951	0.7572	0.6364
Tulu (7B)	vanilla	0.6871	0.7613	0.5953
	trained	0.7013	0.8011	0.6310
Vicuna (7B)	vanilla	0.6862	0.7476	0.6296
	trained	0.7055	0.7545	0.6543

Table 9: Statistics of head-level probing scores (the mean, maximum and minimum of probing scores of all the heads).

This section shows the mean, maximum and minimum of head-level probing scores. As shown in Table 9, after being trained by DECENT with PET, the attention heads of backbones achieve a higher probing score, which means they have a stronger ability to distinguish consistent and inconsistent summaries. The conclusion is aligned with the visualization in Section 4.2.

Article: Even though the UK as a whole voted to leave the EU, in Scotland most people voted to remain. Now, some people are saying that Scotland should get independence from the rest of the UK, so that it can join the EU again on its own. Naz has been in Scotland to see what kids there think.
Summary 1: Scottish children want independence from the UK so they can rejoin the EU, despite the majority of the UK voting to leave.
Summary 2: Scotland voted to remain in the EU, but most of the UK voted to leave. Some Scots are now calling for independence to rejoin the EU.
Which is better?: **2**

Article: Concerns had been raised by the Pakistan Cricket Board over threats from Hindu extremists, who attacked the offices of Indian cricket's governing body last year. However, PCB chairman Shaharyar Khan said on Thursday that the team had been cleared to play in March and April. "As a duty of care, we have asked [world cricket governing body] the ICC to put in place special arrangements for the Pakistan cricket team while in India." The World T20 runs from 8 March to 3 April...
Summary 1: Pakistan's cricket team has been cleared to tour India in March and April for the World T20, despite concerns over Hindu extremist threats. The PCB has asked the ICC to provide special security arrangements for the team.
Summary 2: The Pakistan cricket team has been given clearance to play in India in March and April for the World T20, despite concerns over threats from extremists.
Which is better?: **1**

Article: Lambing season is a busy time for farmers, with thousands of baby sheep being born across the UK. Ten-year-old Tom and his sister Mali have been helping their family during lambing season. Around 4,000 lambs will be born on Tom and Mali's farm this year. Leah visits north Wales to meet them, and to find out just how busy it can be...
Summary 1: Farming siblings Tom and Mali have been lambing sheep for the last three years.
Summary 2: Tom and Mali, aged 10 and 12, are helping their family on their farm in north Wales during Lambing season. They expect to birth around 4000 lams this year.
Which is better?: **0**

Table 10: Three examples of human evaluation.