Improving Model Factuality with Fine-grained Critique-based Evaluator

Yiqing Xie[1] **Wenxuan Zhou**[2*] **Pradyot Prakash**[2*] **Di Jin**[2] **Yuning Mao**[2]
Quintin Fettes[2] **Arya Talebzadeh**[2] **Sinong Wang**[2] **Han Fang**[2]
Carolyn Rosé[1] **Daniel Fried**[1,2] **Hejia Zhang**[2]

[1]Carnegie Mellon University [2]Meta GenAI

Abstract

Factuality evaluation aims to detect factual errors produced by language models (LMs) and hence guide the development of more factual models. Towards this goal, we train a factuality evaluator, FENCE, that provides LM generators with claim-level factuality feedback. In particular, we train FENCE to (1) generate textual critiques along with scores and (2) make claim-level judgment based on diverse source documents obtained by various tools, via data augmentation on a combination of *public* judgment datasets. We then present a framework that leverages FENCE to improve the factuality of LM generators by constructing training data. Specifically, we generate a set of candidate responses, ask FENCE to revise and score each response without introducing lesser-known facts, and train the generator by preferring highly scored revised responses. Experiments show that our data augmentation methods improve the evaluator's accuracy by 2.9% on LLM-AggreFact. With FENCE, we improve Llama2-7B-chat/Llama3-8B-chat's factuality rate by 16.86%/14.45% on FActScore, outperforming state-of-the-art factuality finetuning methods by 8.83%/6.96%.

1 Introduction

Hallucination is one of the persistent challenges for large language models (LLMs), where models generate plausible sounding but incorrect information, even if they are shown factual information during pretraining (Zhang et al., 2023; Li et al., 2024b). One hypothesis is that LLMs fail to distinguish the boundary between memorized facts and other plausible sounding information and do not learn to only output memorized facts, especially on their unfamiliar topics (Gekhman et al., 2024; Ghosal et al., 2024; Kang et al., 2024b). Although it is possible to reduce hallucination in inference with decoding strategies (Li et al., 2024c; Chuang et al., 2024) or post-editing (Mishra et al., 2024; Kang et al., 2024a), they introduce severe latency issues and hurts efficiency in real-time applications.

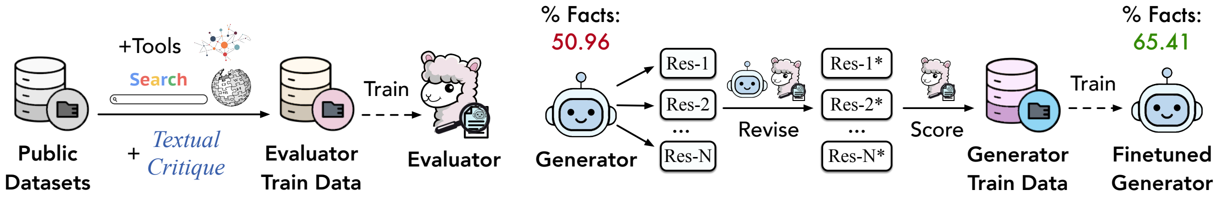
Alternatively, prior studies train the generator to output more factual responses, by preferring (1) generation candidates with higher factuality scores (Tian et al., 2024), which is limited by the generator's capabilities, or (2) responses with false information corrected (Kang et al., 2024a), which is prone to introducing lesser-known facts. As shown in recent work (Gekhman et al., 2024; Ghosal et al., 2024), such preference training reinforces the model to generate information not well memorized during pretraining, and hence could even hurt factuality. Furthermore, the methods either leverage proprietary models that have restricted terms of use, or prompt the generator to evaluate its own factuality, which suffers from self-bias and leads to inaccurate judgments (Xu et al., 2024).

Recent work trains evaluator models that could potentially be used to provide training signals for generators. One category of work relies on proprietary models (e.g., GPT-4) to generate training data in various formats (Kim et al., 2024; Li et al., 2024a). In contrast, Vu et al. (2024) leverage public datasets containing judgments of whether a claim is factual against certain source documents. However, such documents are generally sampled from very **restricted sources** such as news corpora or Wikipedia, while an evaluator could potentially benefit from knowledge obtained by a multiplicity of tools (e.g., search engines) (Wei et al., 2024). Furthermore, the judgment label in most datasets is a single binary or numeric score, providing **limited feedback** to the generator model.

In this paper, we present FENCE, a **Fi**ne-grai**n**ed **C**ritic-based **E**valuator that aims to provide textual critiques for each model-generated claim based on diverse knowledge sources. We start with a set of public datasets with human judgments on the

[*]Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8140–8155
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics



(a) Evaluator Training (b) Using Evaluator to Improve Generator Factuality

Figure 1: *(Left)* The framework to train an evaluator, FENCE, by augmenting public datasets with textual critiques and more diverse knowledge sources. We show the details in Figure 2. *(Right)* The framework to improve model factuality with FENCE. We construct training data by leveraging FENCE to revise and score the generator's responses. Details of response revision are shown in Figure 3.

factuality of model-generated claims. As shown in Figure 1(a), we augment the judgment labels with textual critiques, which provides more informative and explainable feedback to the generator. In addition, we augment the source documents by invoking multiple tools, including a search engine, knowledge base, and knowledge graph, with the goal of training the evaluator to leverage more diverse knowledge sources.

We further demonstrate how to leverage FENCE to improve the generator's factuality with finetuning. To construct training data (Figure 1(b)), we generate multiple responses for each prompt, use FENCE to judge and critique every claim in the responses, and replace false information with facts or remove it from the response, depending on whether the corresponding fact is within the generator's knowledge (i.e., whether the generator outputs *"unknown"* when prompted about claim correctness). This mitigates introducing lesser-known facts into training data. Finally, unlike existing work that use the generator itself as the evaluator, we use FENCE to score each original and revised response and construct preference data, which reduces self-bias and produces more accurate judgments.

Experiments show that FENCE outperforms large open-source models such as Mistral-123B and strong proprietary models such as Claude-3 on LLM-Aggrefact. With FENCE, we improve Llama2-7B-chat's factuality by 16.86% on FActScore (Min et al., 2023) and 17.64% on TruthfulQA (Lin et al., 2022), outperforming existing factuality finetuning methods by 8.83% and 3.99%, respectively. Analyses further show that after training with our recipe, the generator outputs less information for unfamiliar entities and more information for popular ones, suggesting that it learns to only generate information that is likely to be factual.

Contributions. (1) We train a fine-grained critique-based evaluator, FENCE, by augmenting public datasets with textual critiques and more diverse source documents. (2) We propose a training recipe to improve the generator's factuality by leveraging FENCE to improve and score its responses. (3) We conduct extensive experiments to validate both the judgment accuracy of FENCE and the factuality of the generator trained with FENCE, outperforming state-of-the-art factuality training methods by 8.83% on FActScore and 3.99% on TruthfulQA.

2 Methodology

2.1 FENCE: Factuality Evaluator Training

Preliminary. To train the evaluator to recognize hallucinations, previous work (Vu et al., 2024) incorporates a combination of datasets with human judgments on factuality. As shown in Figure 2, in most datasets, each example contains a claim, a source document, and the ground-truth judgement of whether the claim is (1) fully supported by, (2) contradicted with, or (3) contains information that cannot be verified by the document.

We formally define the problem as follows: Given a claim $c \in \mathcal{C}$, a source document $d \in \mathcal{D}$, a factuality evaluator aims to learn a mapping $f : \mathcal{C} \times \mathcal{D} \to \mathcal{L}$, which maps each claim-document pair (c, d) to one of the labels: $f(c, d) = l \in \mathcal{L} = \{\textit{Supported, Contradictory, Unverified}\}$.

Augmenting Labels with Textual Critiques. In addition to the classification label, we aim to train the evaluator to generate a *textual critique* that explains the judgement, which provides more informative feedback such as which part of the document supports or contradicts the claim. We will leverage such feedback to revise generator responses and use them to train a more factual generator (see §2.2 for details). Formally, we aim to learn the mapping $f : \mathcal{C} \times \mathcal{D} \to \mathcal{R} \times \mathcal{L}$, which

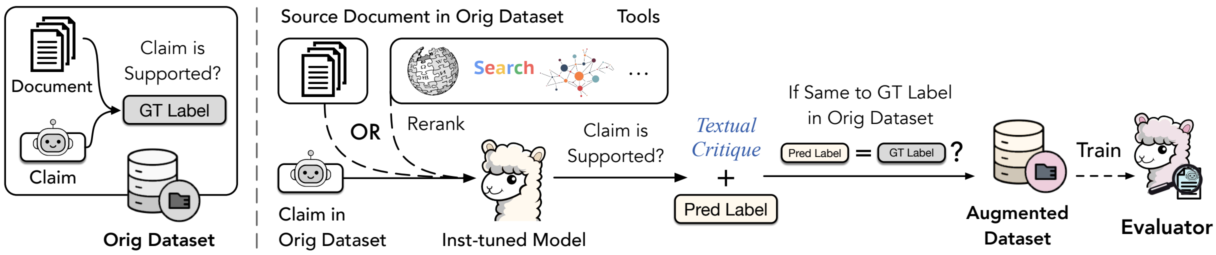


Figure 2: Framework of evaluator training. (*left*) Existing public datasets for evaluator training. Each example contains a claim, a source document, and the ground truth (GT) label of whether the claim is supported by the document. (*Right*) We augment the datasets with textual critiques and more diverse source documents obtained by tools. We use zero-shot Llama-3-70B-chat as the instruction-tuned model in our experiments.

maps each claim-document pair (c, d) to both the textual critique $r \in \mathcal{R}$ and the label $l \in \mathcal{L}$.

As shown in Figure 2, we prompt an instruction-tuned model $\mathcal{M}$ (e.g., Llama3-70B-chat) to generate both the critique $r_\mathcal{M}$ and label $l_\mathcal{M}$ for whether a claim c is supported. The critique and label are likely to be consistent because the label is generated conditioned on the critique. As a result, if the predicted label $l_\mathcal{M}$ aligns with the ground truth label l_{GT} in the original dataset, the critique is also likely to be aligned and we hence use both the critique $r_\mathcal{M}$ and label $l_\mathcal{M}$ as the new training target. Otherwise, if the predicted label does not align with the ground truth, we discard the whole example.

Augmenting Source Documents using Tools. To judge the factuality of an arbitrary model-generated claim, we could potentially benefit from *a multiplicity of tools* such as search engines or online knowledge bases. However, the source documents in existing judgment datasets typically come from very restricted sources, such as news corpora or Wikipedia. To bridge this gap, we obtain additional source documents for each claim by calling the following tools: a search engine (Bing Search API), knowledge base (Wikipedia), and knowledge graph (Google Knowledge Graph API).

As shown in Figure 2, given the claim c in the original dataset, we prompt an instruction-tuned model to call multiple tools to verify the factuality of the claim (i.e., by generating tool calls such as search queries). Then we rerank the returned results to obtain a combination of tool-extracted documents d_t. Similar to critique generation, we prompt the instruction-tuned model $\mathcal{M}$ to predict whether claim c is supported by the tool-extracted documents d_t. We add d_t to the train set if the predicted label $f_\mathcal{M}(c, d_t)$ is the same as the ground truth label $f_{GT}(c, d)$ in the original dataset.

The intuition is that if a claim can be supported by some documents, it is likely that we can obtain other supporting sources by calling the tools. If a claim is hallucinated, it is very unlikely to find any knowledge that supports it with any tools. In both cases, we have $f_{GT}(c, d_t) = f_{GT}(c, d)$.

Training Objective. After obtaining the augmented training data $\mathcal{TR}_{Eval} = \{(c, d), (r, l)\}$, where each example contains (claim c, source document d, critique r, label l), we initialize the evaluator $\mathcal{E}$ with a instruction-tuned model and train it with a standard conditional language modeling objective, maximizing likelihood:

$$\max_{\mathcal{E}} \mathbb{E}_{(c,d),(r,l)\sim\mathcal{TR}_{Eval}} \log \mathbb{P}_\mathcal{E}(r, l \mid c, d). \quad (1)$$

2.2 Improving Generator Factuality with FENCE

In this section, we use our evaluator, FENCE, to improve a generator model's factuality, where we construct training data by revising and scoring the generator's own responses. Compared to directly training the generator on factuality datasets, our method only requires a prompt set as inputs and hence enjoys much better scalability.

Overview. As shown in Figure 1, given a prompt, we use the generator to generate N candidate responses. Then we improve the factuality of each response by using FENCE to evaluate the factuality of each piece of generated information and editing or removing the false information, depending on whether the corresponding fact is rare. Finally, we use FENCE again to score all original and revised responses and construct training data.

Response Revision. We aim to improve the factuality of the generator's responses *without introducing lesser-known facts* to training data. The motivation is that as shown in recent research (Ghosal et al., 2024; Gekhman et al., 2024), forcing the model to

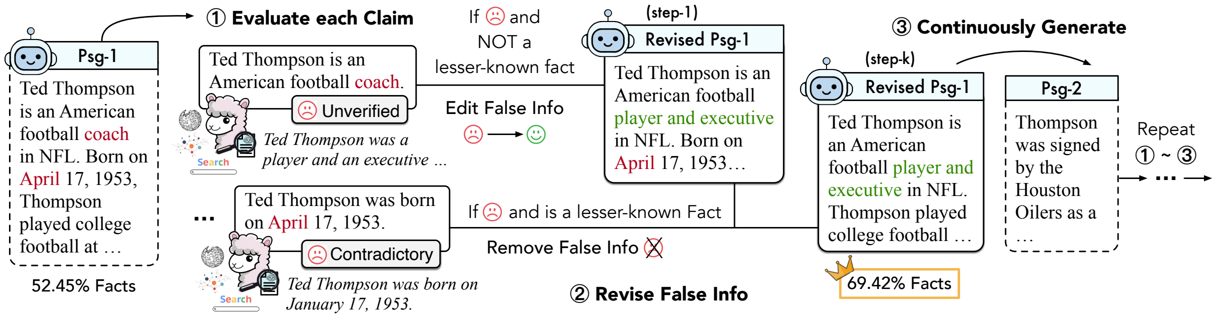


Figure 3: The framework to revise model responses without introducing lesser-known facts. We iteratively (1) use FENCE to evaluate the factuality of each claim, (2) replace false information (if any) with the correct fact or remove it from the response, depending on whether it corresponds to a lesser-known fact, and (3) continue generating the next passage. For every claim, we prompt the generator "*Is this claim factual?*" without providing any source documents. If the generator outputs *"unknown"*, we assume that the claim corresponds to a lesser-known fact.

generate lesser-known facts that are poorly memorized during pretraining will blur the boundary with memorized facts and other plausible sounding information, which may lead to even more hallucination. As shown in Figure 3, we iteratively revise each passage with the following three steps:

[**Step-① Evaluate**] we prompt an instruction-tuned model (e.g., Llama3-70B-chat) to decompose each response into claims. Then for each claim, we call tools to obtain related documents and apply FENCE to evaluate its factuality with critique.

[**Step-② Revise**] If there are any claims that are judged as *"unverified"* or *"contradictory,"* we further check whether it corresponds to a lesser-known fact. Specifically, we prompt the generator *"Is this claim factual?"* and output *"true," "false,"* or *"unknown,"* without providing it any external knowledge. We regard the claim as a lesser-known fact if the generator outputs *"unknown."*

If the claim does not correspond to a lesser-known fact, we prompt the generator to correct the false information based on the critique generated by FENCE. Otherwise, we prompt the generator to remove the false information from the passage.

[**Step-③ Generate**] To reduce error propagation, we use the revised passages as the prefix and continuously generate the next passage.

Generator Training. We use FENCE to score each original and revised response by computing the percentage of factual claims. Then we train the generator with first supervised finetuning (SFT) and then direct preference optimization (DPO) (Rafailov et al., 2023). In the SFT stage, we train the generator with the top-k responses as targets, where the responses are ranked by the percentage of factual claims, optimizing with the conditional language modeling objective similar to Equation 1.

In the DPO stage (Rafailov et al., 2023), we construct preference data $\mathcal{TR}_{Gen} = \{x, y_w, y_l\}$ as follows: for each prompt x, we choose the preferred response y_w from the top-k responses and choose any responses with lower scores than y_w as the rejected response y_l. Suppose we have N original and N revised responses, this gives us $\binom{2N}{2} - \binom{2N-k}{2}$ preference pairs. We initialize the generator from the SFT checkpoint and optimize the following classification loss:

$$\max_{\mathcal{G}} \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{TR}_{Gen}} \Big[\log \sigma \Big(\beta \log \frac{\pi_{\mathcal{G}}(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_{\mathcal{G}}(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \Big) \Big], \tag{2}$$

where σ is the Sigmoid function. The reference policy π_{ref} is computed by the SFT checkpoint.

3 Experiments on Evaluator Training

In this section, we aim to answer the research question: (**RQ1**) Can FENCE correctly judge the factuality of model-generated claims?

3.1 Experimental Setup

Training Details. We initialize FENCE from Llama3-8B-chat and train it on a set of public factuality judgment datasets which Vu et al. (2024) is trained on. To ensure label accuracy, we focus on datasets with human judgment on model responses, including summarization datasets: XSum Hallucination (Maynez et al., 2020), QAGS (Wang et al., 2020), FRANK (Pagnoni et al., 2021), question-answering datasets: RAGTruth (Niu et al., 2024), FActScore (Min et al., 2023), and dialogue datasets:

Model Name	LLM-AggreFact (*without* threshold tuning)										
	AggreFact		TofuEval		Wice	Reveal	Claim Verify	Fact Check	Expert QA	LFQA	Avg
	CNN	XSum	MediaS	MeetB							
Open-source Models (Llama3-8B based)											
Llama3-8B-chat	50.9	58.2	63.9	72.4	65.1	85.2	63.8	76.5	55.9	72.3	66.4
FenCE *(Vanilla SFT)*	63.2	73.4	66.6	77.7	64.6	86.4	72.5	73.0	57.9	83.1	71.8
FenCE *(Critique Only)*	59.5	74.7	68.4	80.0	71.7	88.0	74.3	74.2	59.6	**87.0**	73.7
FenCE *(Full)*	62.1	72.4	**70.9**	**80.3**	76.0	**88.6**	**74.9**	74.4	**60.3**	86.9	**74.7**
Other Open-source Models (47B-123B)											
Mistral-8x7B†	55.0	65.5	68.5	73.3	63.8	80.8	64.3	75.1	56.3	70.8	67.3
Llama3-70B-chat	**63.3**	71.3	67.9	75.2	74.8	86.7	67.3	78.3	58.4	82.9	72.6
Mistral-123B†	58.4	**76.3**	67.3	78.9	**76.6**	88.4	67.6	**79.0**	60.0	81.7	73.4
Proprietary Models or Distilled Models											
Gemini-Pro†	49.4	60.6	63.8	65.8	65.8	85.5	61.8	76.8	56.8	75.9	66.2
GPT-3.5†	63.2	72.4	66.8	73.4	68.5	84.7	65.2	70.8	57.2	73.8	69.6
Claude-2.1†	59.9	66.4	69.2	72.3	64.3	88.2	69.7	79.3	59.8	78.2	70.7
Claude-3 Opus†	65.2	72.4	74.1	82.4	75.0	83.8	69.3	78.8	58.8	81.6	74.1
MiniCheck-FT5†	69.9	74.3	73.6	77.3	72.2	86.2	74.6	74.7	59.0	85.2	74.7
GPT-4†	66.7	76.5	71.4	79.9	80.4	87.8	67.6	79.9	59.2	83.1	75.3

Table 1: Performance (BAcc) of evaluator models on the test split of LLM-AggreFact. We separate Llama3-8B-chat based models, larger open-source models, and proprietary models into different blocks. We highlight the best-performing open-source model for each dataset. Results with † are reported in Tang et al. (2024).

Q^2 (Honovich et al., 2021), FaithDial (Dziri et al., 2022a), BEGIN (Dziri et al., 2022b). We provide implementation and dataset details in §A and §B.

Evaluation Dataset and Metric. We evaluate the evaluators on the LLM-AggreFact benchmark (Tang et al., 2024), a combination of 10 datasets covering three tasks: fact verification, summarization, and long-form QA. All datasets contain human-annotated (document, claim, label) tuples.

We follow Tang et al. (2024) and use balanced accuracy (BAcc) as the evaluation metric: BAcc $= \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right)$, where TP, TN, FP, and FN represent true/false positives/negatives.

Baselines and Ablations. We compare to the LLM-based fact-checkers reported by Tang et al. (2024). We do not include results reported in Vu et al. (2024) because they use a different metric.

In addition, we compare to two ablations: (1) FenCE (*Vanilla SFT*), which is trained on the original public datasets with no augmentation, and (2) FenCE (*Critique Only*), where we only generate textual critiques and not source documents.

3.2 Main Results

As shown in Table 1, after training on our augmented datasets, FenCE improves Llama3-8B-chat by 8.3% BAcc and outperforms all the open-source LLMs, including models with significantly more parameters (e.g., Mistral-123B). It also outperforms strong proprietary models such as Claude-3 Opus. When compared with its ablations, FenCE consistently outperforms the Vanilla SFT model on 8 out of 10 datasets, with average gain of 2.9% BAcc. The performance of FenCE (*Critique Only*) is between Vanilla SFT and the full model, which indicates the utility of both augmentation methods.

Among all the datasets, we observe that the performance on Wice decreases after Vanilla SFT, but is largely improved by FenCE. One possible reason is that some training datasets such as Q^2 contain claims that are labeled as "factual" but are only partially supported by the documents. We hypothesize that such noisy examples strongly affect the performance on Wice, which contains as high as 54.7% of such "partially supported" examples. However, by filtering out examples where Llama3-70B-chat cannot generate explanations, we filter out a large percent of such noisy examples and hence improve the final judgment accuracy.

3.3 Result Analyses

Accuracy of Augmented Critique and Source Documents. With our data augmentation methods, we equipped 77.2% of the training examples with textual critiques, and generate new source documents for 54.1% of the examples with the combination of our three tools. To verify the data quality, we randomly sampled 45 examples where we successfully obtain critiques or source documents (15 for each label) and manually inspect the accuracy. Specifically, we check whether both the critique and the label correctly reflect the relationship be-

Pred Label (→)	Supported	Contradictory	Unverified
Critique Acc	15/15	14/15	14/15
Tool-ext Doc Acc	14/15	15/15	15/15

Table 2: The accuracy of the critiques and the tool-extracted source documents we obtained. We randomly sampled 15 examples for each predicted label and manually check the accuracy of each example.

tween the claim and the source documents.

As shown in Table 2, 95.6% of the augmented critiques and 97.8% of the tool-extracted documents are accurate. Although in one of the wrongly-labeled examples, our method mislabels "unverified" as "contradictory", it does not affect the final conclusion that the claim is not factual.

Case Studies. We further show three concrete examples of our augmented critiques and source documents. In the first example (Table 6), our generated critique correctly explain the judgment label: the claim mentions a call for a national project, which does not appear in the documents.

In the second example (Table 7), the original document is a CNN news report. By calling the search engine, we obtain other news articles written by diverse news agencies on the same event (i.e., Kenneth Morgan's murder). Such tool-extracted documents increase the diversity of documents in the train set, while still having high label accuracy.

In the third example (Table 8), we obtain Chadwick Boseman's birthday information from all three tools: knowledge graph, Wikipedia, an search engine, where the knowledge graph provides more structured and concise knowledge, while Wikipedia returns a long paragraph containing the information. Compared to the original documents, our tool-extracted documents have more diverse formats, improving the evaluator's generalizability.

4 Experiments on Generator Factuality

We aim to answer the research questions: (**RQ2**) Can we leverage FENCE to improve the generator's factuality? (**RQ3**) How well can our training recipe improve the generator's factuality?

4.1 Experimental Setup

Datasets and Evaluation Metric. Following prior works (Kang et al., 2024a; Zhang et al., 2024), we conduct experiments on FActScore (Min et al., 2023) and TruthfulQA (Lin et al., 2022). For FActScore, we randomly split the unlabeled split

Model Name	FActScore # Facts	FActScore # Errors	FActScore % Facts	TruthfulQA % True*Info
Llama2-7B-chat	10.70	17.04	38.57	38.83
+ SFT	10.76	15.59	40.83	45.52
+ Self-Eval-SKT	11.02	14.18	43.73	48.65
+ EVER-Pref	**11.24**	15.11	42.66	51.07
+ FactTune-FS	11.23	12.87	46.60	52.48
(Our Method)				
+ E/R + Coarse	10.84	**8.72**	**55.43**	**56.47**
Llama3-8B-chat	17.83	17.16	50.96	58.89
+ SFT	20.05	18.13	52.52	59.17
+ Self-Eval-SKT	18.69	14.22	56.80	61.88
+ EVER-Pref	20.25	15.16	57.18	63.01
+ FactTune-FS	18.77	13.34	58.45	64.58
(Our Method)				
+ E/R + Coarse	**20.40**	**10.79**	**65.41**	**67.14**

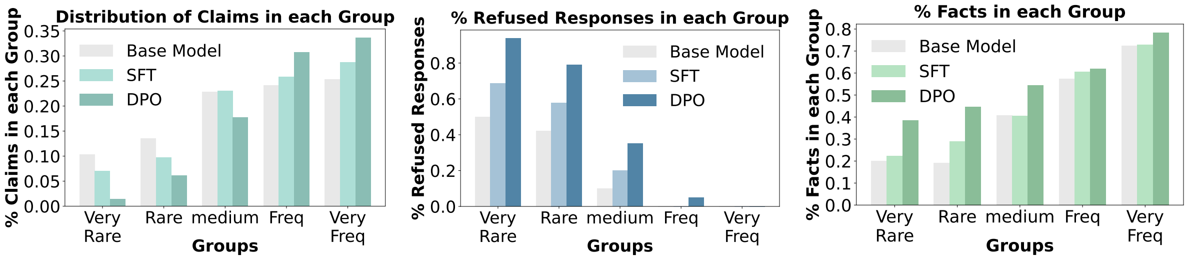
Table 3: Comparison between our method and baselines on FActScore and TruthfulQA. "E/R" stands for "Edit/Remove". All baselines use the zero-shot models as the evaluator in training and our method uses FENCE.

Model Name	FActScore # Facts	FActScore # Errors	FActScore % Facts
(Ablations with FENCE as the evaluator)			
Llama3-8B-chat	17.83	17.16	50.96
+ SFT + FENCE	21.19	16.47	56.26
+ Edit	**20.68**	14.42	58.91
+ Coarse	20.07	12.89	60.89
+ Edit + Coarse	20.03	11.09	64.37
(Our Full Method)			
+ E/R + Coarse	20.40	**10.79**	**65.41**

Table 4: Ablation study that compares different training recipes when equipped with FENCE as the evaluator. "SFT + FENCE", "Edit", and "Coarse" denote equipping "SFT", "EVER-Pref", and "FactTune-FS" in Table 3 with the FENCE evaluator.

into 400 training and 100 test prompts. We compute the "% Facts" metric by extracting correct and incorrect facts in each response, where we use Llama3-70B-chat to decompose the responses and use the same three tools in training to obtain source documents. For TruthfulQA, we randomly select 3 examples from each of the 38 categories as the training set and use the remaining 703 examples as the test set. We follow the "generation" setting and use the finetuned evaluator in the original paper to compute "% True*Info", the percentage of responses that are both truthful and informative.

Baselines and Ablations. We implement four baselines: SFT, FactTune-FS (Tian et al., 2024), Self-Eval-SKT (Zhang et al., 2024), and EVER-Pref (Kang et al., 2024a). All methods sample N candidate responses for each training prompt. SFT first uses the generator itself to score responses by



(a) Distribution of claims (b) % Refused responses (c) % Facts

Figure 4: Statistics of generated responses, where we group the prompts by the popularity of the person to write biography for, as labeled in the FActScore dataset. We compare zero-shot Llama3-8B-chat (denoted as *"Base Model"*), SFT + FENCE (*"SFT"*), and our method, Edit/Remove + Coarse (*"DPO"*).

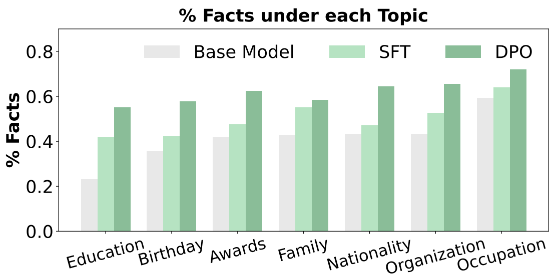
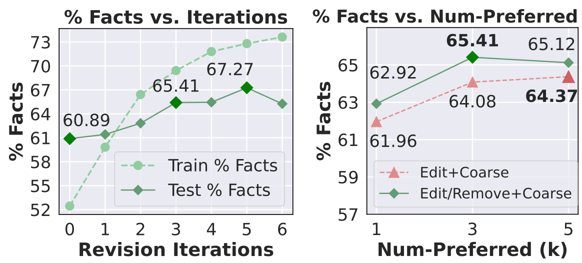


Figure 5: The factuality of claims with different topics, where we pre-define the topics and use Llama3-70B-chat to predict whether each claim covers any topic(s).



(a) Iterations of Revision (b) top-k

Figure 6: Hyper-parameter analysis. We investigate how percentage of facts changes (a) as number of iterations to revise candidate responses increases, and (b) with different numbers of preferred responses for each prompt.

computing the percentage of factual claims and finetunes with the best response. FactTune-FS first finetunes on all candidates and then uses all $\binom{N}{2}$ candidate pairs as DPO pairs, preferring the one with a higher score based on retrieved context. Self-Eval-SKT self-trains an evaluator using the model's own knowledge and uses the evaluator to score responses with no external context, also resulting in $\binom{N}{2}$ preference pairs. EVER-Pref uses all N candidates as rejected responses and constructs a preferred response by iteratively evaluating and correcting false information in each passage.

For ablations, we first equip SFT, FactTune, and EVER with our evaluator, FENCE, and the three external tools, denoted as "SFT + FENCE", "Coarse", and "Edit" in Table 4, respectively. Then we implement "Edit + Coarse", which corrects all false information without checking whether it corresponds to a lesser-known fact. We denote our full method as "E/R + Coarse" (Edit/Remove + Coarse). More implementation details can be found in §A.

Note that we do not compare to retrieval-augmented (RAG) methods (Lewis et al., 2020) because it is not a fair comparison. RAG-based methods require access to external knowledge sources or tools in inference, while in our experiments, we only used "`tell me a bio of <entity>`" as the prompt. Furthermore, while most RAG-based methods aim to improve the understanding of the long retrieved context, our main focus is to train the model to better utilize the knowledge it has seen during pretraining.

We also do not compare to inference-time methods (Li et al., 2024c; Chuang et al., 2024) because these methods need to call the model multiple times in inference, introducing heavy latency issues. In the real world, such latency is unacceptable in production. In comparison, our method uses the standard random decoding in inference and does not introduce extra inference-time cost. Furthermore, in principle, our method is orthogonal to those inference-time methods and can be combined.

4.2 Main Results

Results in Table 3 answer (**RQ2**) and show that our method significantly improves Llama2/Llama3's factuality by 16.85/14.45% on FActScore and 17.64%/8.25% on TruthfulQA. Results also show that our method significantly outperforms the best baseline on factuality training (e.g., by 8.83/6.96%

on FActScore for Llama2/Llama3).

Table 4 presents the ablations with FENCE as the evaluator. The comparison between Table 3 and Table 4 shows that FENCE can improve the performance of SFT, EVER, and FactTune. For example, on FActScore, with Llama3 as the base model, SFT with FENCE as the evaluator outperforms SFT with Llama3 as the evaluator by 3.74%.

In Table 4, we observe that our method of combining response editing/removing with coarse-level scoring achieves significantly better performance. In particular, our full method, which only corrects false information with common facts, outperforms "Edit + Coarse", which could introduce both common and lesser-known facts into the training data. The above results answer (**RQ3**), demonstrating the effectiveness of our training recipe.

4.3 Result Analyses

Distribution of Generated Claims. We first group the prompts by the popularity of the person to write biography for, which is provided as meta data in the FActScore dataset, and then compare the distribution of generated claims in each group before and after training. As shown in Figure 4a, after training, the generator outputs less information for unfamiliar people and outputs more information for popular ones, suggesting that it learns to only generate information that is likely to be factual.

In addition, as shown in Figure 4b, we observe that the generator refuses to generate responses more frequently for rare entities (e.g., by generating "I apologize, but I'm not familiar with this person."), and almost never refuses for frequent ones. This aligns with previous research's conclusion (Kang et al., 2024b) that training the model to say "I don't know" to unfamiliar prompts reduces hallucination.

Performance Breakdown. We further check the generator's performance (i.e., percentage of factual claims) on different groups of prompts. As shown in Figure 4c, we observe that our method achieves consistent performance gain over all groups of prompts, which is another reason why our method obtains higher overall performance.

Similarly, we check the performance on claims describing different topics, where we first come up with a list of topics and then prompts Llama3-70B-chat to determine whether each claim covers any of the topics. In Figure 5, our method generates more factual claims under all the topics, and the performance gain is larger on unfamiliar topics (i.e., topics where the generator has lower scores).

Hyper-parameter Analysis. We first alternate the number of iterations to revise the responses and inspect the testing and training accuracy (i.e., the average % of facts of the best preferred responses). As shown in Figure 6a, with more revision iterations, we can always obtain preferred responses with better factuality, but the test performance converges after the third iteration. In other words, training data with fewer factual errors does not always transfer to better test performance.

In our experiments, we only choose the top-k candidates (ranked by the percentage of facts) as preferred responses. We investigate the effect of k in Figure 6b. We observe that training on top-3 and top-5 responses leads to similar performance. With all the ks in our experiments, our method consistently outperforms the "Edit + Coarse" ablation.

5 Related Work

Factuality Evaluation. To judge the factuality of long-form model responses, recent works have presented fine-grained level evaluation frameworks that judge each piece of generated fact individually (Min et al., 2023; Chern et al., 2023; Xie et al., 2024; Wei et al., 2024). Such frameworks generally leverage an LM evaluator to make judgments.

To train LM evaluators, one line of work constructs new training datasets by collecting human judgement (Li et al., 2023b; Jiang et al., 2024). Another line of work distills open-source evaluators from proprietary models such as GPT-4, training the evaluator to generate textual critique (Wang et al., 2023; Cui et al., 2024; Li et al., 2024a) or fine-grained judgment (Kim et al., 2024; Mishra et al., 2024). A recent work (Vu et al., 2024) leverages a combination of existing public datasets to train an evaluator. We further augment the public datasets with more diverse knowledge sources and more informative judgment feedback.

Enhancing Generator's Factuality. To reduce hallucinations, previous works present inference-time methods, including re-computing token probabilities (Shi et al., 2024; Chuang et al., 2024) or conducting post-editing (Saunders et al., 2022; Huang et al., 2023; Welleck et al., 2023; Madaan et al., 2023; Gou et al., 2024; Mishra et al., 2024). Such methods inevitably suffer from latency issues.

Another approach is to train the model for factuality. Following general reward modeling methods that produce a score for the entire response (Ziegler et al., 2020; Ouyang et al., 2024; Menick et al., 2022), FactTune (Tian et al., 2024) trains the generator by preferring responses with higher percentage of facts, which is limited by the generator's capability. EVER (Kang et al., 2024a) constructs high-quality responses by correcting false information, which may introduce lesser-known facts and could potentially harm model factuality (Ghosal et al., 2024). Both methods either prompt proprietary model or use the generator itself to evaluate its own factuality, which are either restricted by terms of use or suffer from self-bias.

Our method is different from existing methods in the following aspects: (1) we combine response revision and scoring to construct high-quality responses while ensuring the correctness of preference ranking, (2) we only correct false information with common facts and remove other misinformation from training, and (3) we use our open-source evaluator, FENCE, for both revision and scoring.

6 Conclusions

We improve LM generators' factuality by training an open-source evaluator model, FENCE. To train FENCE to leverage diverse knowledge sources and to generate more informative feedback, we equip a combination of public datasets with textual critique along with judgment scores and obtain additional source documents by calling a multiplicity of tools. We then present a training recipe that leverages FENCE to finetune LM generators for better factuality, where we construct preference data by prompting FENCE to revise and score the generator's responses, without introducing lesser-known facts in training. Experiments show that FENCE outperforms strong proprietary models such as Claude-3 on LLM-AggreFact. Our factuality training method improves Llama3-8B-chat's factuality performance by 14.45% on FActScore and 8.25% on TruthfulQA.

7 Limitations

We list the limitations of this work as follows:

Evaluator Training Data. We only train our evaluator on human-annotated datasets on model responses. We have not investigate the effect of other datasets (e.g., synthetic datasets or human-written claim datasets) and we leave it to future works.

Furthermore, in this work, we focus on text-to-text generation and do not train our evaluator on math reasoning or programming tasks, which are also classified as factuality tasks in some existing works (Chern et al., 2023).

Experiments on Improving Factuality. To test our generator's performance, we following existing works (Tian et al., 2024; Kang et al., 2024a) and only experiment on one public dataset: FActScore. In principle, one can also apply our training recipe to other prompt datasets.

8 Acknowledgement

We thank Xinran Zhao, Tim Franzmeyer, Zichun Yu, and Pengfei Liu for their helpful feedback on this work. This work was supported in part by NSF grant DSES-2222762. Yiqing Xie is supported by the Carnegie Mellon University Presidential Fellowship in the Language Technologies Institute.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. *Preprint*, arXiv:2307.13528.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. *Preprint*, arXiv:2310.01377.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *Preprint*, arXiv:1811.01241.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10.

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

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *Preprint*, arXiv:2405.05904.

Gaurav Rohit Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. 2024. Understanding finetuning for factual knowledge extraction. In *Forty-first International Conference on Machine Learning*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

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

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. Tigerscore: Towards building explainable metric for all text generation tasks. *Preprint*, arXiv:2310.00752.

Haoqiang Kang, Juntong Ni, and Huaxiu Yao. 2024a. Ever: Mitigating hallucination in large language models through real-time verification and rectification. *Preprint*, arXiv:2311.09114.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024b. Unfamiliar finetuning examples control how language models hallucinate. *Preprint*, arXiv:2403.05612.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024a. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024b. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024c. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas 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.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *Preprint*, arXiv:2203.11147.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

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*. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. 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.

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. *Preprint*, arXiv:2305.18290.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *Preprint*, arXiv:2206.05802.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *Preprint*, arXiv:2404.10774.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation. *Preprint*, arXiv:2407.10817.

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.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Shepherd: A critic for language model generation. *Preprint*, arXiv:2308.04592.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. Doclens: Multi-aspect fine-grained evaluation for medical text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024. Pride and prejudice: Llm amplifies self-bias in self-refinement. *Preprint*, arXiv:2402.11436.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. *Preprint*, arXiv:1809.07358.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Implementation Details

Evaluator Training: Obtaining Critiques. As introduced in §2.1, for each prompt in the training set, we prompt Llama3-70B-chat to generate 10 candidate judgments, each containing both the critique and the label. If one of the 10 candidates has the same label as the ground truth label, we add the critique generated by this candidate to the training set. Otherwise, we discard the whole example.

Most existing datasets only have binary labels ("factual" or "non-factual") and our label space includes three label classes. We hence match "supported" with "factual" and match both "contradictory" and "unverified" to "non-factual".

Evaluator Training: Obtaining Source Documents by Tools. We also augment the source documents by calling three tools: Bing Web Search API, an offline copy of Wikipedia, and Google Knowledge Graph API, which represents three types of tools: search engine, knowledge base, and knowledge graph. We first put the documents obtained by all the tools in a document list, then we rerank the documents based on the cosine similarity between their text embeddings and the claim's embedding, where we use GTR-T5-Large (Ni et al., 2022) as the encoder. We provide the evaluator with the top-5 documents.

To call Bing Web Search API, we first prompt Llama3-70B-chat to generate a search query with the instruction: "You are given a STATEMENT. Your task is to write one SEARCH QUERY to find evidence supporting or disproving the STATEMENT." Then we call the Bing Search API, which returns 5 search results for each query. Each result contains the URL and a short snippet. We further scrape each URL to obtain the full content of the webpage and chunk the content (with 512 as the chunk size). Finally, we add all the chunks and snippets to the document list.

To call Wikipedia, we download an offline copy of Wikipedia (the 2023/04/01 version). Similar to search query generation, we prompt Llama3-70B-chat to generate a list of possible Wikipage names. For each generated name, we retrieve the top-3 Wikipage based on cosine similarity of the pagename embeddings. We chunk the content of all the retrieved Wikipages and add the chunks to the document list.

To call the Google Knowledge Graph API, we prompt Llama3-70B-chat to generate a list of entities for each claim and add the top-1 returned result to the document list (if any).

Generator Training: Baselines and Ablations. For all the baselines and ablations, we use Llama3-8B-chat as initialization, set $N = 5$, use Llama3-70B-chat to decompose the responses into facts, and call the same tools for evaluation (Bing Search, Wikipedia, Google Knowledge Graph). We use Llama3-8B-chat as the evaluator for both baseline methods. For our method and EVER, we revise at most 3 false claims (i.e., the ones judged as "contradictory" or "unverified") for each of the first three passages.

B Dataset Details

Evaluator Training Data. We provide the list of datasets we used to train FENCE in Table 5. All the datasets are open-sourced on HugginFace or GitHub. We only focus on datasets with *human judgments* on *model-generated responses* or claims decomposed from model responses, excluding synthetic datasets such as HaluEval (Li et al., 2023a), sentence revision-based datasets such as VitaminC (Schuster et al., 2021), etc.

Generator Training and Testing Data. We only require a prompt set to finetune the generator. Following previous work (Tian et al., 2024; Kang et al., 2024a), we conduct experiments on the FActScore dataset (Min et al., 2023).

Unlike Kang et al. (2024a) that uses the same set of prompts for training and testing, to evaluate the generator's generalizability to unseen prompts, we follow Tian et al. (2024) and use different prompts for training and testing. Since Tian et al. (2024) does not release their train-test split, we make our own split by randomly dividing the unlabeled subset of FActScore into 400 training and 100 testing prompts[1].

C Case Studies

Evaluator Training: Data Augmentation. Table 6 shows an example of our generated textual critique, which is aligned with the ground truth label. Table 7 and Table 8 show two examples where we obtain additional source documents by calling tools. We can see that compared to the existing document, the tool-extracted documents are more diverse in terms of sources, content, and formats, and are still correctly labeled.

[1]We release our split of training testing prompts at: https://drive.google.com/drive/folders/1GsTmoh1t1jInSrUcgej1kZWG7KNXDFL4?usp=sharing.

Category	Dataset Name	Base Datasets
Summarization	XSum Hallucination (Maynez et al., 2020)	XSum (Narayan et al., 2018)
	QAGS (Wang et al., 2020)	XSum (Narayan et al., 2018) CNN/DM (See et al., 2017)
	FRANK (Pagnoni et al., 2021)	XSum (Narayan et al., 2018) CNN/DM (See et al., 2017)
	RAGTruth (Niu et al., 2024)	CNN/DM (See et al., 2017)
Question Answering	RAGTruth (Niu et al., 2024)	MSMARCO (Bajaj et al., 2018)
	FActSore (Min et al., 2023)	WikiBio (Lebret et al., 2016)
Dialogue	Q-Square (Honovich et al., 2021)	Wizard of Wikipedia (Dinan et al., 2019)
	FaithDial (Dziri et al., 2022a)	Wizard of Wikipedia (Dinan et al., 2019)
	BEGIN (Dziri et al., 2022b)	Wizard of Wikipedia (Dinan et al., 2019) CMU-DoG (Zhou et al., 2018)

Table 5: The list of training datasets we used to train FENCE.

Source Document *(Dataset: FRANK-xsum)*

Title: News Articles

Text: They believe ministers are placing too much emphasis on the environment at the expense of trees grown for timber. Britain is currently the world's third largest importer of wood. Ministers said they were encouraging commercial forestry organisations to invest in woodland creation. Conifer forests have been a familiar sight for half a century in Wales and have helped the timber industry grow. But Confor, which promotes the forestry industry, warns that at least 16,000 hectares - or 40,000 acres - of commercial forest have been lost since 2001 and need to be re-planted to meet needs. Half of the woodland is managed by Natural Resources Wales with the the other half by private companies. As an industry, it is estimated to be worth more than Â£450m a year to the Welsh economy. But there is a conflict with environmental policy. **Successive ministers have wanted more native, broadleaf species of trees and better access for walkers and bike riders while conservationists claim commercial forests are bad for wildlife.** However, some forests in Wales are difficult to reach and hard to harvest while transporting the timber through rural communities can bring companies into conflict with local residents. The Welsh government wants a balance between the environmental, the economic and the social in how forestry is developed. It points to its Glastir scheme which encourages farmers and other landowners to plant woodland. It said its Woodlands for Wales forestry strategy "sets out how we want Wales to be known for its high quality woodlands that enhance the landscape, provide real social and community benefits, support thriving woodland-based industries and contribute to a better quality environment". Martin Bishop, of Confor, which represents 200 businesses in Wales, believes the emphasis on environmental aspects is seen as "sexier" than commercial forestry. He said the current system is also too bureaucratic for creating new woodland areas. Mr Bishop said 90 hectares were planted last year but there needed to be 20,000 hectares a year to meet long-term targets set by ministers. "There's a huge market for our timber. Every saw miller tells me he would double or treble production if the timber was there to feed that mill," he added.

Claim

Conservationists are calling for a national project to protect forests in wales.

Ground Truth Label

Non-Factual

Our Generated Critique and Label

Critique: The document does not mention conservationists calling for a national project to protect forests in Wales. It only mentions that conservationists claim commercial forests are bad for wildlife, but it does not mention a specific call for a national project.

Label: **Unverified**

Table 6: An example of the textual critique we generate. We highlight the most relevant sentence in the document. In this case, our generated critique aligns with the judgment label.

Original Source Document *(Dataset: FRANK-CNN/DM)*

Title: News Articles

Text: (CNN) Deputies rushed Kenneth Morgan Stancil III from court Thursday after the 20-year-old murder suspect swore at a judge and tried to flip over a table. Stancil is accused of killing an employee Monday at Wayne Community College in Goldsboro, North Carolina. Relatives have said victim Ron Lane was gay, CNN affiliate WNCN reported, and investigators are looking into whether the shooting was a hate crime. Authorities arrested Stancil after he was found sleeping on a Florida beach on Tuesday. Just a few minutes into Thursday's hearing on the first-degree murder charge he faces, Stancil snapped back at the judge after he was offered a court-appointed lawyer. N̈o, I don't need one,s̈aid Stancil, who stood before the judge with his legs shackled and his arms handcuffed in front of him. Ÿou know what I'm saying? I knew I would get life anyway.S̈uperior Court Judge Arnold O. Jones interjected, pointing out that the maximum sentence Stancil faces is the death penalty. Ÿes, I know that,S̈tancil fired back. B̈ut when I knew what I had to do and I knew when I got caught, you know, I knew in my mind that I could get life, I could get the death penalty. You know what I'm saying? Do you follow my topic? I would have killed you, you know what I'm saying, if you're a f—ing child molester.T̈he judge told him not to swear. Ï don't give a f— what you want,S̈tancil said, lunging forward and lifting up the table in front of him. Deputies quickly corralled him and hustled him from the courtroom. The hearing resumed about 25 minutes later, when Stancil was brought back into the courtroom, this time with his arms handcuffed behind him. When asked again by Jones whether he wanted a lawyer, his response was quick – and calm. Ÿes, sir,ḧe said. In an interview with CNN affiliate WRAL, Stancil described himself as a neo-Nazi and said he hates gay people ẅith a passion.S̈tancil had worked for Lane, the school's print shop operator, as part of a work-study program, but was let go from the program in early March because of poor attendance, college officials said.

Claim

Kenneth Morgan Stancil is accused of killing an employee at Wayne Community College in Goldsboro, North Carolina.

Ground Truth Label

Factual

Generated Queries to Call the Tools

Search query for Bing Search API: Kenneth Morgan Stancil Wayne Community College killing

Queries for Wikipedia: Kenneth Morgan Stancil; List of school shootings in the United States; Wayne Community College;

Entities for Google Knowledge Graph API: Kenneth Morgan Stancil

Tool-extracted Source Documents *(after Reranking)*

Title: Wayne Community College shooter gets life sentence without parole

Text: GOLDSBORO, North Carolina (WTVD) – **Kenneth Morgan Stancil III was sentenced Tuesday to life in prison without parole for the murder of 44-year-old Ron Lane on the campus of Wayne Community College in Goldsboro on April 13, 2015.** Stancil entered the campus print shop on the third floor of the same building that houses the school library and cafeteria shortly after Lane arrived for work that day and shot him once with a pistol-grip 12-gauge shotgun. ...

Title: Stancil guilty in Wayne Community College murder trial - CBS17.com

Text: GOLDSBORO, N.C. (WNCN) – After deliberating for an hour and a half, **a Wayne County jury found Kenneth Morgan Stancil III guilty of first-degree murder.** He will spend the rest of his life in prison. During his closing argument Tuesday in a Wayne County court, the district attorney relied largely on Stancil's own words, **painting him as confident and proud about his decision to kill a man at Wayne Community College.** In his counterargument, Stancil's attorney said the prosecution tried to scare the jury with Stancil's possession of knives, guns, and white supremacist memorabilia, as well as his numerous tattoos. None of that proved he planned to commit the murder, the defense said. **Neither side disputed that Stancil killed Ron Lane at Wayne Community College in 2015**. The prosecution contends it was first-degree murder, but the defense argues ...

Title: Man convicted in fatal Wayne Community College shooting - WRAL ...

Text: Man convicted in fatal Wayne Community College shooting A white supremacist was sentenced Tuesday to life in prison without parole for gunning down a print shop operator at Wayne Community College two years ago. **Jurors deliberated for about 90 minutes before convicting Kenneth Morgan Stancil III, 22, of first-degree murder in the April 13, 2015, death of Ron Lane in the school's library.** Authorities said Stancil shot Lane with a 12-gauge pump shotgun before fleeing the area on a motorcycle. He was captured the next morning when police found him asleep on a beach in Volusia County, Fla. ...

Title: Wayne Community College killing suspect makes shocking allegations in ...

...

Our Generated Critique and Label

Critique: The statement is supported by the document. In all the documents, it is mentioned that Kenneth Morgan Stancil III is accused of killing Ron Lane, an employee at Wayne Community College in Goldsboro, North Carolina.

Label: **Supported**

Table 7: An example where we obtain multiple relevant source documents by calling a multiplicity of tools.

Original Source Document *(Dataset: FActScore)*
Title: Chadwick Boseman
Text: Chadwick Boseman Chadwick Aaron Boseman **(; November 29, 1976August 28, 2020)** was an American actor. During his two-decade career, Boseman received multiple accolades, including two Screen Actors Guild Awards, a Golden Globe Award, a Critics' Choice Movie Award, a Primetime Emmy Award, and an Academy Award nomination. After studying directing at Howard University, Boseman began his career in theatre, winning a Drama League Directing Fellowship and an acting AUDELCO ...

Claim
Chadwick Boseman was born on November 29, 1977.

Ground Truth Label
Non-Factual

Generated Queries to Call the Tools
Search query for Bing Search API: Chadwick Boseman date of birth
Queries for Wikipedia: Chadwick Boseman
Entities for Google Knowledge Graph API: Chadwick Boseman Stancil

Tool-extracted Source Documents *(after Reranking)*
Title: knowledge graph result
Text: Chadwick Boseman: American actor. Chadwick Aaron Boseman was an American actor, known for portraying African-American historical figures and the fictional African superhero Black Panther. **Chadwick Boseman Born: November 29, 1976, Anderson, SC.** Chadwick Boseman Died: August 28, 2020 (age 43ǒ0a0years), Los Angeles, CA. Chadwick Boseman Spouse: Taylor Simone Ledward. Chadwick Boseman Height: 6ž032 0ž033. Chadwick Boseman Parents: Leroy Boseman and Carolyn Boseman. Chadwick Boseman Alma mater: Howard University (BFA). ...

Title: Chadwick Boseman | Biography, Movies, Black Panther, & Death text: Chadwick Boseman
Text: Our editors will review what you've submitted and determine whether to revise the article.
- Chadwick Aaron Boseman
- **Born: November 29, 1976, Anderson, South Carolina, U.S.**
- Died: August 28, 2020, Los Angeles, California (aged 43)
- Notable Family Members: son of Leroy Boseman ...

Chadwick Boseman (born November 29, 1976, Anderson, South Carolina, U.S.—died August 28, 2020, Los Angeles, California) was an American actor and playwright who became a highly respected movie star with several iconic roles, notably that of T'Challa/Black Panther in the groundbreaking film Black Panther (2018). ...

Title: Chadwick Boseman - Wikipedia
Text: Chadwick Boseman Chadwick Aaron Boseman **(; November 29, 1976August 28, 2020)** was an American actor. During his two-decade career, Boseman received multiple accolades, including two Screen Actors Guild Awards, a Golden Globe Award, a Critics' Choice Movie Award, a Primetime Emmy Award, and an Academy Award nomination. After studying directing at Howard University, Boseman began his career in theatre, winning a Drama League Directing Fellowship and an acting AUDELCO ...

Title: Chadwick Boseman | Biography, Movies, Black Panther, & Death ...
...

Our Generated Critique and Label
Critique: The statement claims that Chadwick Boseman was born on November 29, 1977. However, all four documents consistently state that Chadwick Boseman was born on November 29, 1976, not 1977.

Label: **Contradictory**

Table 8: An example where we obtain multiple relevant source documents by calling a multiplicity of tools. The tool-extracted documents all contain the information about Chadwick's birthday, but have diverse formats.