FactAlign: Long-form Factuality Alignment of Large Language Models

Chao-Wei Huang **Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
f07922069@csie.ntu.edu.tw y.v.chen@ieee.org

Abstract

Large language models have demonstrated significant potential as the next-generation information access engines. However, their reliability is hindered by issues of hallucination and generating non-factual content. This is particularly problematic in long-form responses, where assessing and ensuring factual accuracy is complex. In this paper, we address this gap by proposing FactAlign, a novel alignment framework designed to enhance the factuality of LLMs' long-form responses while maintaining their helpfulness. We introduce fKTO, a fine-grained, sentence-level alignment algorithm that extends the Kahneman-Tversky Optimization (KTO) alignment method. Leveraging recent advances in automatic factuality evaluation, FactAlign utilizes fine-grained factuality assessments to guide the alignment process. Our experiments on open-domain prompts and information-seeking questions demonstrate that FactAlign significantly improves the factual accuracy of LLM responses while also improving their helpfulness. Further analyses identify that FactAlign is capable of training LLMs to provide more information without losing factual precision, thus improving the factual F1 score.[1]

1 Introduction

Generating natural language provides a natural interface for humans to communicate with artificial intelligence. With the emergence of large language models (LLM) (Brown et al., 2020), they immediately demonstrate the potential to become the next-generation engine for information access due to their ability to generate long-form natural language response to human queries. Given the large-scale pre-training on web-scale datasets, LLMs demonstrate impressive capabilities of answering diverse questions, showcasing the vast amount of knowledge they possess. The post training techniques, i.e., instruction tuning (Wei et al., 2022) and reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), further train LLMs to respond in a more human preferable way, e.g., generating coherent and detailed responses.

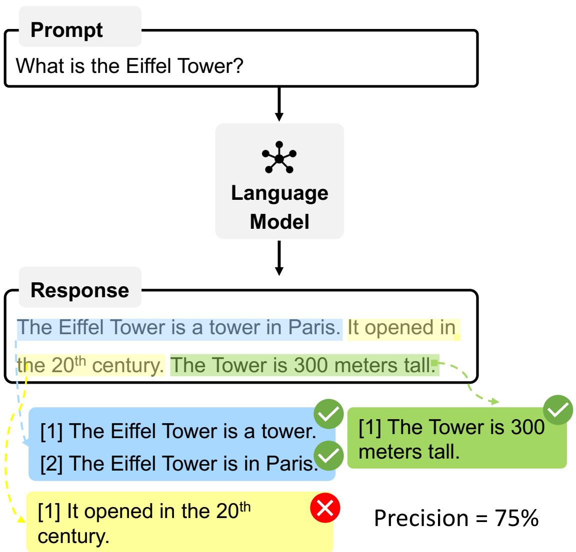

Figure 1: An example of the evaluation long-form factuality. The long-form response is broken down into subclaims and verified separately. The factual precision score can be calculated as the precision of all subclaims.

Despite their impressive reasoning capabilities and wide-range knowledge, research has shown that LLMs still struggle with hallucination (Xu et al., 2024b; Rawte et al., 2023) and generating non-factual content (Min et al., 2023). An example of long-form generation and factuality assessment is illustrated in Figure 1. These issues hinder the reliability of LLMs and make it hard to be adopted to real-world settings where factual accuracy is a crucial requirement for most applications. The long-form responses make these issues more complex as it is non-trivial to quantify the level of long-form factuality (Wei et al., 2024), let alone

[1] Our source code, datasets, and trained models are publicly available at https://github.com/MiuLab/FactAlign.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16363–16375
November 12-16, 2024 ©2024 Association for Computational Linguistics

to improve it. Meanwhile, most research focuses on improving the helpfulness of LLM chatbots and their reasoning capabilities, with little emphasis on the factuality of the responses.

In this paper, we aim to improve the reliability of LLMs by enhancing the factuality of their long-form responses. Recent advances of automatic factuality evaluators show that they are capable of providing factuality assessment at the atomic fact level (Min et al., 2023; Wei et al., 2024). To leverage those fine-grained factuality assessments, we propose FACTALIGN, an alignment framework designed to improve LLMs' long-form factuality while maintaining the same level of helpfulness. We introduce a fine-grained alignment algorithm, **fKTO**, which extends the Kahneman-Tversky Optimization (KTO; Ethayarajh et al. (2024)) alignment algorithm to sentence-level. We conduct experiments on both open-domain prompts and information-seeking questions and demonstrated that our proposed FACTALIGN can effectively improve long-form factuality of LMs while maintaining their helpfulness.

Our main contributions can be summarized as the following:

- We introduce fKTO, a sentence-level alignment algorithm that can leverage fine-grained signals provided by a long-form factuality evaluator.
- We propose FACTALIGN, a framework to align LMs with fine-grained signals to generate responses that are more factual, while keeping their helpfulness.
- The effectiveness of the proposed components are validated through detailed analyses.

2 Related Work

2.1 Language Model Alignment

Alignment, i.e., aligning language models to human values, has been a very popular research field recently. Prior work such as InstructGPT (Ouyang et al., 2022) and LLaMA-2 (Touvron et al., 2023) showcased that RLHF (Bai et al., 2022a) enhances models' ability to follow instructions significantly. Fine-grained RLHF (Wu et al., 2024) proposed to leverage fine-grained rewards for better alignment. Constituional AI (Bai et al., 2022b) and RLAIF (Lee et al., 2023) introduced AI feedback to eliminate the requirement of human annotation.

Another line of research focused on alignment without RL. DPO (Rafailov et al., 2023) derived a simple objective for alignment, thus attracting rapid adoption. KTO (Ethayarajh et al., 2024) eliminated the requirement of pairwise preference data. Our proposed alignment algorithm, fKTO, extends KTO to sentence-level, which can leverage the fine-grained signals provided by a long-form factuality evaluator.

2.2 Factuality of Langage Models

Factuality and hallucination have been long-standing issues for natural language generation (Lee et al., 2022; Ji et al., 2023). Lee et al. (2022), Li et al. (2023), and Chuang et al. (2024) proposed decoding techniques that improved factuality of LMs. Shuster et al. (2021) reduced hallucination by retrieval-augmented generation. (Dhuliawala et al., 2023) proposed chain-of-verification to reduce LLM hallucination. SelfCheckGPT (Manakul et al., 2023) proposed a method to self-check factuality by sampling multiple generations. FactScore (Min et al., 2023; Chiang and Lee, 2024) and LongFact (Wei et al., 2024) both introduced frameworks for evaluating factuality of long-form generations. FAVA (Mishra et al., 2024) introduced fine-grained hallucination categories to evaluate the models and provided a detailed view of the hallucination issues of LLMs. Our proposed method also utilize a long-form factuality evaluator, while focusing on leveraging the provided factuality assessments for better factuality alignment.

Prior work has also worked on training LMs to be more factual. FactTune (Tian et al., 2024) leveraged FactScore to construct preference pairs and demonstrated improvement on the bio generation task. FLAME (Lin et al., 2024) introduced factuality-aware alignment which combines FactTune with open-domain prompts. KnowTuning (Lyu et al., 2024) proposed knowledge augmentation which constructs synthetic pairs for DPO training. On the other hand, recent work has shown that fine-tuning LMs on new knowledge might encourage hallucinations (Gekhman et al., 2024; Kang et al., 2024). Our work additionally proposes fKTO for fine-grained factuality alignment, which achieves superior performance.

3 Preliminaries

In this paper, we aim to improve the long-form factuality of LLMs by factuality alignment. In this section, we introduce an overview of the task of long-form factuality and alignment algorithms.

3.1 Long-form Factuality

LLMs excel at generating long-form responses with detailed description and explanation. However, evaluating the factuality of long-form generations is non-trivial. In this paper, we define the factuality score of a long-form response as an aggregation of the factuality score of each individual atomic fact, following FactScore (Min et al., 2023) and LongFact (Wei et al., 2024). More formally, given a knowledge corpus $\mathcal{C}$, an user prompt x and the response $y = \mathcal{M}(x)$ generated by a model $\mathcal{M}$, we first decompose y into atomic statements $A = \{a_1, \cdots, a_{|A|}\}$. For each atomic statement a_i, its factuality score $f(a_i)$ is defined as whether it is supported by the knowledge in $\mathcal{C}$, i.e., $f(a_i) = \mathbb{1}[a_i \text{ is supported by } \mathcal{C}]$. Then, the factuality score of the long-form response y can be defined as $f_{\mathcal{A}}(y) = \mathcal{A}(\{f(a_1), \cdots, f(a_{|A|})\})$, where $\mathcal{A}$ is an aggregation function that can be defined in various ways.

In this paper, we adopt two metrics for long-form factuality: factual precision as defined in FactScore (Min et al., 2023) and factual f1 score as defined in LongFact (Wei et al., 2024). Factual precision measures the overall precision of the atomic statements:

$$f_{prec}(y) = \frac{\sum_{i=1}^{|S|} f(a_i)}{|A|}.$$

While factual precision is simple, it could be easily exploited. A model could obtain a very high factual precision score by only generating one statement that has the highest confidence.

On the other hand, factual f1 assumes that a certain amount of information is desired by the user and additionally considers the factual recall:

$$f_{f1@K}(y) = \begin{cases} \frac{2 \cdot f_{prec}(y) \cdot f_{rec@K}(y)}{f_{prec}(y) + f_{rec@K}(y)} & \text{if } |A| > 0 \\ 0 & \text{if } |A| = 0, \end{cases}$$

where $f_{rec@K}(y) = \min(1.0, \frac{|A|}{K})$ is the factual recall score assuming that at least K statements are desired by the user. Factual f1 is less exploitable than factual precision as it punishes the model when it only generates few statements.

3.2 Kahneman-Tversky Optimization

Training LLMs that are aligned to human values typically involves three stages: 1) pre-training, 2) supervised fine-tuning, and 3) reinforcement learning from human feedback (RLHF). The first two stages maximize the sequence generation likelihood of the LM given a dataset of either diverse pre-training data or human-annotated instruction-following data. The third stage, RLHF, aims to maximize the expected reward of LM generations, where the reward usually is defined as human preferences (Ouyang et al., 2022). As a result, the RLHF stage enables LMs to generate responses that are more preferable by humans, which is vital for creating intelligent assistants.

While the success of the RLHF framework is eminent, its adoption is hindered by the complexity of the framework, the unstability of the training process, and the increased training time due to the requirement of online sample generation. To this end, prior work has proposed alignment algorithms that do not require RL, thus attracting mass adoption. Direct Preference Optimization (DPO; Rafailov et al. (2023)) derives a simpler objective from the RLHF, eliminating the requirement of a reward model and the RL optimization process. More recently, Ethayarajh et al. (2024) introduced Kahneman-Tversky Optimization (KTO), which derives a family of human-aware alignment loss functions. The objective function of KTO is even simpler than DPO. It only requires a binary label for each prompt-response pair (x, y), as opposed to DPO which requires pairwise preference labels for each triplet (x, y_1, y_2). This relaxed data requirement enables us to extend the algorithm to sentence-level, which we will discuss in Section 4.2. More formally, the KTO loss is defined as:

$$\mathcal{L}_{\text{KTO}} = \frac{1}{|\mathcal{B}|} \sum_{x,y \in \mathcal{B}} (\lambda_y - v(x, y)),$$

where $\mathcal{B}$ denotes the minibatch, λ_y denotes the weight of the chosen and rejected samples, and

$$v(x, y) = \begin{cases} \lambda_c \sigma(\beta(r_\theta(x, y) - z_0)) \text{ if } c(x, y) = 1, \\ \lambda_r \sigma(\beta(z_0 - r_\theta(x, y))) \text{ if } c(x, y) = 0, \end{cases}$$

$$z_0 = \mathbb{E}_{y' \sim \mathcal{D}}[\text{KL}(\pi_\theta(y' \mid x') \| \pi_{\text{ref}}(y' \mid x'))],$$

$$r_\theta(x, y) = \log \frac{\pi_\theta(x, y)}{\pi_{\text{ref}}(x, y)},$$

where $c(x, y)$ denotes the preference function, i.e., $c(x, y) = 1$ if the response y is *chosen*. Ethayarajh et al. (2024) demonstrated that KTO achieves on par or better alignment performance compared to DPO. KTO also works well under the scenario where the number of chosen and rejected samples are significantly unbalanced, e.g., 1:9.

4 FACTALIGN: Aligning Language Models for Long-form Factuality

In this section, we introduce our proposed framework FACTALIGN. An overview of our framework is illustrated in Figure 2.

4.1 Automatic Long-form Factuality Evaluator

Obtaining fine-grained factuality annotations for long-form responses by human annotation is very costly. For example, Min et al. (2023) estimated that evaluating one generation costs $4. In this work, we employ an automatic factuality evaluator for long-form responses. The factuality evaluator, following the design of FactScore (Min et al., 2023) and SAFE (Wei et al., 2024), is a workflow of 4 stages: 1) atomic statement decomposition, 2) query generation, 3) relevant knowledge search, and 4) final factuality assessment. Note that stage 2 and 3 can be run multiple times to enrich the searched knowledge.

Atomic Statement Decomposition The response y is first split into sentences $S = \{s_1, \cdots, s_{|S|}\}$, and each sentence is decomposed into atomic facts A. We add an additional step to revise the decomposed atomic statements into self-contained statements s_i' with GPT-3.5-TURBO following SAFE.

Query Generation We prompt GPT-3.5-TURBO to generate a search query given the revised statement s_i' and possibly the previously generated queries and found knowledge snippets.

Relevant Knowledge Search We employ Wikipedia as the knowledge corpus $\mathcal{C}$ following FactScore. While the coverage of Wikipedia is more limited compared to commercial search engines like Google Search, we opt for Wikipedia as this reduces cost and allows us to fully manage the knowledge search component under a controlled setting. We perform search with the generated query and obtain the top-k most relevant knowledge snippets.

Final Factuality Assessment We prompt GPT-3.5-TURBO to provide the final factuality assessment of a revised statement s_i', which is either *Supported* if the statement is supported by the knowledge snippets, or *Not Supported* otherwise. The statement-level score is then defined as $f(a_i) = \mathbb{1}[a_i \text{ is Supported}]$. Note that $f(a_i)$ represents whether the statement is supported with respect to Wikipedia, not whether it is globally true.

4.2 Long-form Factuality Alignment

At the core of the FACTALIGN framework is the alignment algorithm, which operates on two granularities: response-level and sentence-level.

4.2.1 Response-level Alignment

We employ the standard KTO loss $\mathcal{L}_{\text{KTO}}$ for response-level alignment. The preference labels $c(x, y)$ in the KTO loss can be defined and obtained in various ways. For instance, most prior work utilized human-annotated preference labels or pseudo labels provided by LLMs. In order to align for factuality, we treat a response y as a *chosen* sample if the factual f1 score of the response is greater than a threshold t:

$$c(x, y) = \mathbb{1}[f_{f1@K}(y) > t].$$

By minimizing the response-level loss, we align the LMs to generate responses that have higher factual f1 scores.

In addition to the data for factuality alignment, the response-level loss is compatible to other forms of preference data. For example, in order to make the model more helpful, we can include diverse preference datasets that are based on human preferences. In practice, we include general-domain alignment datasets during training to make sure the model is aligned to diverse human values.

4.2.2 Sentence-level Alignment

Since our factuality evaluator provides assessments at a finer granularity, we propose a fine-grained alignment algorithm, **fKTO**, to leverage these signals by extending the KTO alignment algorithm to sentence-level. The fKTO loss is defined as

$$\mathcal{L}_{\text{fKTO}} = \frac{1}{|\mathcal{B}|} \sum_{x,y \in \mathcal{B}} \frac{1}{|S|} \sum_{i=1}^{|S|} (\lambda_f - v(x \parallel s_{<i}, s_i)),$$

where $x \parallel s_{<i}$ denotes the concatenation of x and $s_{<i}$ which denotes sentences before s_i. In this objective function, a sentence s_i is treated as the completion given $x \parallel s_{<i}$. A sentence is *chosen* if the average precision of its atomic statements is higher than a threshold t_s.

$$c(x \parallel s_{<i}, s_i) = \mathbb{1}\left[\frac{\sum_{j=1}^{|A_{s_i}|} f(a_j)}{|A_{s_i}|} > t_s\right],$$

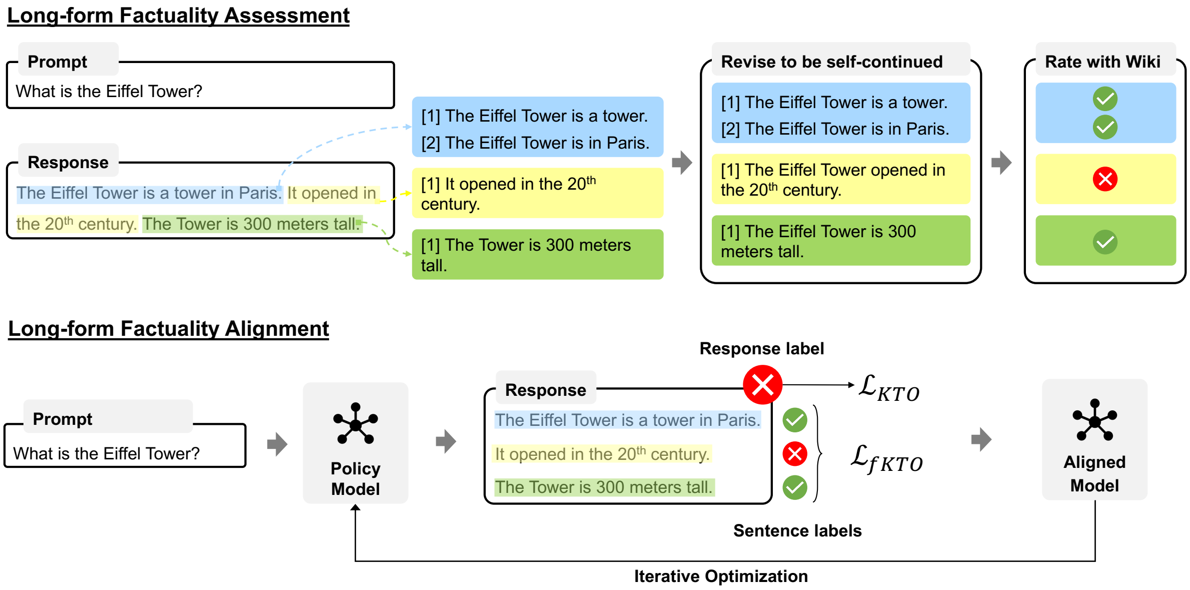

Figure 2: An overview of our FACTALIGN framework. Top: the pipeline for long-form facutality assessment. Bottom: the long-form factuality alignment process.

where $A_{s_i} = \{a_j \mid a_j \in s_i\}$ denotes the atomic statements in sentence s_i. The sentence-level loss provides training signals at a finer-grained level, thus enabling the model to be aligned more effectively. Note that the relaxed data requirement enables KTO to be easily extended to the sentence-level, as opposed to algorithms that require pairwise preference labels, e.g., DPO.

Finally, the loss function we optimize is the combination of the response-level and sentence-level losses:

$$\mathcal{L} = \mathcal{L}_{\text{KTO}} + \lambda \cdot \mathcal{L}_{\text{fKTO}},$$

where λ is the weight of the sentence-level loss.

4.3 Iterative Optimization

With the alignment algorithms introduced above, we can align LMs to be more factual and more helpful. However, the responses and factuality assessments are obtained in an offline fashion, i.e., we sample the responses and their factuality labels before training the model and use this data throughout training. This creates a discrepancy between the assessed responses and the model being trained, which would hinder the alignment process due to distributional shift. Hence, we employ an iterative optimization procedure, where we periodically sample new responses with the trained model and assess their factuality. The newly generated responses are then included in the training dataset for the next iteration.

5 Experimental Stetup

We conduct experiments to validate the effectiveness of our proposed framework FACTALIGN. Furthermore, we perform analyses to discuss the effectiveness of each component in the framework.

5.1 Datasets

Supervised Fine-tuning (SFT) We employ the Deita dataset (Liu et al., 2024) for supervised fine-tuning before performing alignment to ensure basic instruction-following capabilities of the model. The Deita dataset consists of high-quality data selected from UltraChat (Ding et al., 2023), ShareGPT[2], and WizardLM (Xu et al., 2024a).

General-domain Alignment We follow the Zephyr recipe (Tunstall et al., 2023) and employ the UltraFeedback dataset (Cui et al., 2023) as the general-domain alignment dataset. The UltraFeedback consists of prompts across multiple domains and completions generated from multiple LLMs to enrich diversity. We use the binarized version of the dataset[3] and decouple the pairs for the KTO loss.

Factuality Alignment We generate information-seeking prompts following the data creation procedure from LongFact (Wei et al., 2024). LongFact

[2] https://sharegpt.com

[3] https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized

consists of 38 topics chosen to ensure diverse coverage. For each topic, we generate 30 prompts with GPT-4-TURBO and sample generations with our policy model. The generations are then assessed by the long-form factuality evaluator and labeled with factuality assessments at an atomic statement level. For each iteration of iterative optimization, we generate a new set of prompts and sample generations with the currently aligned model.

5.2 Long-form Factuality Evaluator

We employ `gpt-3.5-turbo` to perform atomic statement decomposition, query generation, and final factuality assessment. The generation temperature is set to 0.1. We use the preprocessed Wikipedia corpus from the Dec. 20, 2021 dump released by Izacard et al. (2024) as our knowledge corpus $\mathcal{C}$, which consists of 33 million passages. A pre-trained retriever ColBERT-v2 (Santhanam et al., 2022) is used to encode all passages and perform retrieval given a query. We retrieve top-3 passages for each query and combine them with the previously retrieved passages for final factuality assessment. At most 2 search steps are performed to retrieve relevant passages for each statement. Detailed prompts can be found in Appendix A.

5.3 Models

We employ the pre-trained `gemma-2b` model (Team et al., 2024) as our policy model, which is an open-weight model pre-trained on large-scale datasets across diverse domains. The model is first fine-tuned with the Deita SFT dataset, and then aligned with the alignment datasets.

We also conduct experiments on `LLaMA-3 8B` (Meta, 2024) and `Phi3-Mini` models (Abdin et al., 2024), which are both open-weight models which were aligned with proprietary data.

5.4 Evaluation Procedure

The trained models are evaluated on two aspects: *long-form factuality* and *helpfulness*.

Long-form Factuality Evaluation We evaluate models' long-form factuality following the procedure of SAFE (Wei et al., 2024)[4]. We choose the *LongFact-object* subset following the original work, which consists of 38 topics. We change the Google Search API to our Wikipedia retriever due to resource and budget constraint. In preliminary experiments, we find that this change have very little impact on the evaluation outcome. Our evaluator has correlation scores of 0.93 and 0.82 with SAFE for the number of *Supported* and *Not Supported* assessments, respectively. We follow SAFE to add an postamble to each prompt to ask for the model to generate as many details and examples as possible. We report $f1@100$ as the main evaluation metric. We also report the factual precision and factual recall scores. In addition, we evaluate models with FactScore (Min et al., 2023). We run the evaluation from its official implementation[5] and use GPT-3.5-TURBO as the evaluator instead of InstructGPT. FactScore can be interpreted as the factual precision of bio generation.

Helpfulness Evaluation We evaluate models' helpfulness on MT-Bench (Zheng et al., 2023), a popular benchmark that includes challenging multi-turn open-ended questions for evaluating chat assistants. The automatic judgement is performed by GPT-4 with a score of 1 to 10, which is shown to be highly-correlated with human judgement. The evaluation is done with their official implementation[6].

5.5 Implementation Details

We set the threshold t to 0.75, meaning that the response is chosen if its f1@100 is higher than 0.75. The threshold for sentences t_s is set to 1.0, i.e., the sentence is only chosen if all of its atomic statements are supported. During training, we set $\beta = 0.1$ for KTO and $\beta_f = 0.5$ for fKTO. The weight of $\mathcal{L}_{\text{fKTO}}$, λ, is set to 2.0. The learning rate is set to 5e-7 with a linear learning rate schedule. We set the effective batch size to 16 and train for 1 epoch for each iteration. In order to reduce GPU memory consumption during training, we optimize the model with the 8-bit version of the AdamW optimizer. We iteratively optimize the LM as described in Section 4.3 for 3 iterations. All experiments are run on 4xV100 GPUs. Each training run takes 1 to 2 hours to finish. We estimate that each evaluation run costs $25 in API credits.

6 Results

We present the main results in Table 1, where we contrast FACTALIGN with both proprietary models (GPT-4-Turbo and GPT-3.5-Turbo), a prominent

[4] https://github.com/google-deepmind/long-form-factuality

[5] https://github.com/shmsw25/FActScore

[6] https://github.com/lm-sys/FastChat/tree/main/fastchat/llm_judge

	LongFact			FactScore		MT-Bench
	$f1@100$	Precision	# claims	FS	# claims	
GPT-4-Turbo	91.85	87.99	108.9	83.98	49.7	8.99
GPT-3.5-Turbo	61.59	89.25	52.4	75.79	18.1	7.94
Olmo-7B-Instruct	75.42	66.8	131.6	-	-	-
LLaMA-2-70B-Chat	83.78	79.33	112.5	59.06	65.4	6.85
Gemma-2B-SFT	61.74	77.41	66.8	57.02	27.4	4.41
+ FACTALIGN	**86.51**	**79.59**	**135.1**	**59.77**	**51.7**	**5.70**
Phi3-Mini	66.91	74.41	81.5	58.33	36.7	6.58
+ FACTALIGN	**82.31**	**78.29**	**106.7**	**61.21**	**49.6**	**6.63**
LLaMA3-8B-Chat	79.98	71.89	121.5	54.96	50.6	7.38
+ FACTALIGN	**87.32**	**78.46**	**132.6**	**62.84**	**55.7**	**7.44**

Table 1: Main results of our experiments. FS denotes the FactScore and # claims denotes the average number of claims. We report percentage points for $f1@100$, precision, and FS. We mark the best scores among the Gemma-2B models in bold.

open-weight model (LLaMA-2-70B-Chat (Touvron et al., 2023)), and a fully open-source model (Olmo-7B-Instruct) (Groeneveld et al., 2024). The comparison involves our baseline model, the Gemma-2B model[7], which has been fine-tuned using our SFT dataset, Deita. This model serves as the foundational policy model for all subsequent aligned models. Additionally, we benchmark against the rejection sampling fine-tuning method (Yuan et al., 2023), involving supervised fine-tuning with selected samples from our alignment dataset. This method shows modest improvements.

Remarkably, our FACTALIGN framework significantly improves the long-form factuality and helpfulness of the baseline model, achieving relative improvements of 40.1% and 29.2% in terms of $f1@100$ and average score on MT-Bench, respectively. These results demonstrate our capability to simultaneously refine LMs for enhanced factuality and utility. Moreover, FACTALIGN also boosts the FactScore of the baseline models and outperforms larger models like GPT-3.5-Turbo and LLaMA-2-70B-Chat in both $f1@100$ and FactScore metrics. This demonstrates the potential for smaller LMs, through precise alignment, to surpass general-domain large LMs in factual accuracy.

With a detailed examination of the metrics, it is evident that FACTALIGN primarily improves factual recall, increasing the output of factual claims from 66.8 to 135.1, while slightly improving factual precision from 77.41 to 79.59. This enhancement suggests that FACTALIGN primarily amplifies output volume while maintaining factual precision. This trend echoes findings from general-domain alignment research, which indicates that alignment algorithms typically promote longer outputs, likely due to a combined human and LM preference for more extensive responses (Dubois et al., 2024). A qualitative example of this can be found in Appendix B.

[7] https://huggingface.co/google/gemma-2b

	f1@100	Precision
FACTALIGN	**86.51**	**79.59**
- Iterative Optimization	77.10	78.44
- fKTO	73.12	73.27
- General Dataset	61.33	65.72
- Factuality Dataset	68.86	69.93
Rejection Fine-tuning	68.33	77.86

Table 2: Ablation study on LongFact (%).

6.1 Ablation Study

To validate the effectiveness of our proposed components, we conduct an ablation study to understand their contribution to the final improvement. The results are reported under FACTALIGN in Table 2.

Firstly, we remove the iterative optimization technique, where we only perform 1 iteration of training. As shown in the results, removing iterative optimization significantly degrades the performance, where $f1@100$ drops by over 10 points.

	Seen	Unseen
Gemma-2B-SFT	61.97	61.36
General-domain only	69.45	68.23
Seen topics only	76.49	72.23

Table 3: Performance on seen and unseen topics (%). We report the f1@100 score on LongFact.

β_f	threshold t	$f1@100$
0.5	0.65	74.32
0.5	0.75	77.10
0.5	0.85	73.66
0.1	0.75	75.86
1.0	0.75	75.12

Table 4: Performance with various number of β_f and threshold t (%).

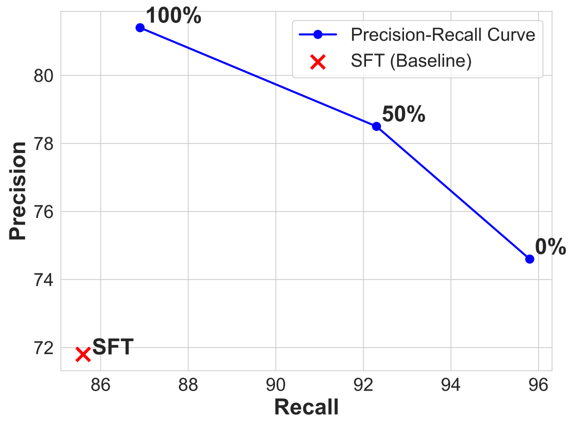

Figure 3: The precision-recall curve with varying ratios of data mixture. SFT denotes the supervised fine-tuned baseline. The labels denote the ratio of the precision data points used.

This result demonstrates that it is crucial to perform iterative optimization or online sampling in order to achieve better performance. We also observe that training on the same dataset for multiple epochs yields worse performance, showcasing that the alignment data quickly becomes stale and no longer is a good sample after 1 epoch of training. Note that for all other ablation experiments, we also only performs 1 iteration of training.

Next, we remove the fKTO loss $\mathcal{L}_{\text{fKTO}}$ and align the model with only $\mathcal{L}_{\text{KTO}}$. Without $\mathcal{L}_{\text{fKTO}}$, the factual f1 score degrades by 4 points from 77.10 to 73.12, demonstrating that the proposed fine-grained alignment objective fKTO can align LMs more effectively. Note that we observe that the fKTO loss occasionally makes the training process unstable. We hypothesize that this is due to the amount of factuality data being much less than the general-domain data, thus making the instances with fine-grained label sparse during training. Hence, the estimation of the fKTO loss becomes slightly unstable. We will also discuss the sensitivity to hyperparameters in Section 6.4.

We also conduct an experiment where we exclude the general-domain alignment dataset from our training data. The performance degrades significantly on all datasets after removing the general-domain alignment dataset. Upon further investigation, we observe that without general-domain data, LMs easily overfit and often generate repetitive outputs. This result indicates that a mixture of general-domain datasets and factualy-specific datasets is important to maintain a balance and prevent catastrophic forgetting.

Finally, we exclude the factuality dataset during training, i.e., only align the LM on general-domain datasets. As shown in the results, aligning with general-domain dataset also improves the long-form factuality and helpfulness of the baseline model. This indicates that factuality might be encoded in the diverse array of human values present in the general-domain alignment dataset. However, including the factuality dataset sill achieves significantly superior performance for long-form factuality.

6.2 Generalization to New Topics

Since the training data is created with the same set of topics in LongFact, all the topics should be considered *seen* during evaluation. Note that prompts used in evaluation are excluded during training. To validate whether FACTALIGN could generalize to unseen topics, we conduct an additional experiment where we split the topics into 19 seen topics and 19 unseen topics. We only include the data from the seen topics during training and perform evaluation on the unseen topics. The results are reported in Table 3. The results show that FACTALIGN performs slightly worse on unseen topics. Nonetheless, it still outperforms the baseline models significantly, showcasing that the alignment can generalize to unseen topics.

6.3 Relationship of Precision-Recall

By varying the ratio of data points using precision as the threshold and those using recall, we can control the tradeoff between the precision score and the recall score. We train models with different data mixture and plot the corresponding precision-recall curve in Figure 3. The model trained with 100% precision data achieves the highest precision score, and the model trained with 100% recall data achieves the highest recall score. Furthermore, we can achieve a specific level of factual precision and recall scores on the curve by changing the ratio. This result demonstrates that FACTALIGN enables control over the desired factual precision and recall scores.

6.4 Sensitivity of Hyperparameters

We report the performance of FACTALIGN under various hyperparamter settings. The results are reported in Table 4. We observe that the threshold t affects performance slightly, with 0.75 being the best setting. We also notice that with $t = 0.75$, the labels are balanced, i.e., the number of chosen samples is roughly equal to the number of rejected samples. This indicates that constructing a balanced dataset perform better for our alignment algorithm.

We also vary the hyperparameter β_f and notice that it degrades performance slightly. Note that the best β_f value is higher than the β value typically set for KTO, i.e., 0.1. Our hypothesis is that since fKTO operates on the sentence-level, the log probability difference naturally has a lower magnitude compared to the response-level case. Thus, a higher value of β_f is needed to promote the fine-grained loss to a similar level as the response-level loss.

7 Conclusion

In this paper, we address the issue of long-form actuality in LLMs by proposing a novel alignment framework, FACTALIGN. Our approach, which incorporates a proposed data construction process alongside the fine-grained alignment algorithm fKTO, significantly enhances the factuality of LLMs over long-form responses, while also boosting their helpfulness. Our analysis demonstrates that FACTALIGN enables detailed control over the desired level of factual precision and recall scores. We believe that the insights and methodologies presented in our work can motivate further advancements in the factuality alignment of LLMs.

Limitations

Our work focuses on the factuality aspect of LLMs, which we define as whether the generated response is supported by retrieved evidence. This definition makes the performance dependent to the performance of the retriever and the coverage of the knowledge corpus. Moreover, our data creation and evaluation pipeline rely on automatic factuality evaluators. Even though prior work has validated the effectiveness of these evaluators by showing high correlation with human judgements, the automatic evaluators inevitably might make incorrect judgements.

While FACTALIGN significantly improves the factuality of LLMs, they still are prone to generate non-factual content. A calibration method would be complimentary to our method to ensure the reliability of LLMs.

We focus on a controlled setting where the information-seeking prompts are all questions about a certain object. This is to ensure the reliability of the automatic evaluation process. Future work could extend the coverage of the information-seeking prompts to more diverse user queries.

Acknowledgements

We thank the reviewers for their insightful comments. This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002-013-MY3 and 112-2223-E002-012-MY5. We thank to National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan and Google PaliGemma Academic Program for providing computational and storage resources.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cheng-Han Chiang and Hung-yi Lee. 2024. Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations. *arXiv preprint arXiv:2402.05629*.

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, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *Preprint*, arXiv:2310.01377.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

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? *arXiv preprint arXiv:2405.05904*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2024. Atlas: few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24(1).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *arXiv preprint arXiv:2405.01525*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Yougang Lyu, Lingyong Yan, Shuaiqiang Wang, Haibo Shi, Dawei Yin, Pengjie Ren, Zhumin Chen, Maarten de Rijke, and Zhaochun Ren. 2024. Knowtuning: Knowledge-aware fine-tuning for large language models. *arXiv preprint arXiv:2402.11176*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

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. *arXiv preprint arXiv:2401.06855*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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*.

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

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2024. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024a. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024b. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Prompts Used

We use the following prompt for new prompt generation

```
Instructions:
1. Ask a general question about a specific
object (such as a person, place, event, act,
company etc.).
2. The object should be very specific and
niche within the topic of {topic}.
4. Do not ask about any particular aspect
of the object (such as history,
significance, impact, role, purpose, etc.).
5. Follow the question styles in the
provided examples.
6. Wrap the question in square brackets.
```

The following prompt is used for query generation:

```
Instructions:
1. You have been given a STATEMENT and
some KNOWLEDGE points.
2. Your goal is to try to find evidence
that either supports or does not
support the factual accuracy of the
given STATEMENT.
3. To do this, you are allowed to
issue ONE search query that you think
will allow you to find additional
useful evidence.
4. Your query should aim to obtain
new information that does not appear
in the KNOWLEDGE. This new information
should be useful for determining the
factual accuracy of the given STATEMENT.
5. Format your final query by putting
it in a markdown code block.

KNOWLEDGE:
{_KNOWLEDGE_PLACEHOLDER}

STATEMENT:
{_STATEMENT_PLACEHOLDER}
```

The following prompt is used for final answer assessment:

```
Instructions:
1. You have been given a STATEMENT and
some KNOWLEDGE points.
2. Determine whether the given STATEMENT
is supported by the given KNOWLEDGE.
The STATEMENT does not need to be explicitly
supported by the KNOWLEDGE, but should be
strongly implied by the KNOWLEDGE.
3. Before showing your answer, think
step-by-step and show your specific reasoning.
As part of your reasoning, summarize the main
points of the KNOWLEDGE related to the
STATEMENT.
4. If the STATEMENT is supported by the
KNOWLEDGE, be sure to show the supporting
evidence.
5. After stating your reasoning, restate
the STATEMENT and then determine your final
answer based on your reasoning and the
STATEMENT.
6. Your final answer should be either
"Supported" or "Not Supported". Wrap your
final answer in square brackets.
```

B Qualitative Examples

We include a qualitative in Table 5.

Model	Response	# Claims	$f1@100$
Prompt: Could you provide some details about the Hague Conventions?			
Gemma-2B-SFT	Sure! The Hague Conventions are a set of international treaties that were designed to protect the rights of civilians in international armed conflicts. Here are some specific details and examples related to the Hague Conventions: 1. The First Hague Convention: The First Hague Convention was signed in 1899 and was designed to protect the rights of civilians in international armed conflicts. It was signed by 31 countries and established the principle that civilians should not be targeted by military forces. Example: In 1904, during the Boxer Rebellion in China, British and French troops were sent to suppress the rebellion. However, they were ordered to avoid killing or injuring the Chinese civilians. The British and French troops complied with this principle and did not fire on the civilians. 2. The Second Hague Convention: The Second Hague Convention was signed in 1907 and established the principle that the use of prohibited weapons in war is prohibited. It was signed by 33 countries and established the principle that the use of prohibited weapons in war is prohibited. Example: During World War I, the use of poison gas by both the German and British armies was prohibited under the Second Hague Convention. However, the German army still used poison gas and was criticized for its use. 3. The Fourth Hague Convention: The Fourth Hague Convention was signed in 1907 and established the principle that the use of prohibited weapons in war is prohibited. It was signed by 33 countries and established the principle that the use of prohibited weapons in war is prohibited. Example: During World War II, the use of chemical weapons by both the German and Soviet armies was prohibited under the Fourth Hague Convention. However, the Soviet army used chemical weapons against the Polish army in 1939, which was not a violation of the Hague Convention. . . . These are just a few examples of how the Hague Conventions have been used to protect the rights of civilians in international armed conflicts. The Hague Conventions have been ratified by more than 190 countries and have been instrumental in shaping international law and protecting the rights of civilians in armed conflicts.	67	63.4
FACTALIGN	The Hague Conventions are a series of international treaties that were signed in the Netherlands city of The Hague in 1899. These conventions were designed to address the issue of international law and to promote cooperation between different countries. The Hague Conventions are divided into several different treaties, each addressing a different aspect of international law. Some of the most significant Hague Conventions include: 1. The Hague Convention for the Protection of Cultural Property in the Event of Armed Conflict (1954) 2. The Hague Convention Respecting the Laws and Customs of War on Land (1907) 3. The Hague Convention for the Protection of Victims of International Armed Conflicts (1977) 4. The Hague Convention on the Civil Aspects of International Litigation (1971) 5. The Hague Convention on the International Sale of Goods (1980) 6. The Hague Convention on the Protection of the Rights of Visigothic Migrants (1998) These conventions were signed by a large number of countries, and they have had a significant impact on international law. For example, the Hague Convention for the Protection of Cultural Property in the Event of Armed Conflict was signed in 1954, and it has been ratified by a large number of countries. This convention has been used in a number of different situations, including the conflict in Bosnia and Herzegovina. . . . In terms of specific examples, the Hague Conventions have been used in a number of different situations. For example, the Hague Convention for the Protection of Cultural Property in the Event of Armed Conflict has been used in a number of different conflicts, including the conflict in Bosnia and Herzegovina. The Hague Convention Respecting the Laws and Customs of War on Land has been used in a number of different situations, including the conflict in Afghanistan. Overall, the Hague Conventions are considered to be some of the most significant international treaties of the 20th century. They have had a significant impact on international law, and they continue to be relevant today.	65	78.9

Table 5: An example of model generations. The generations are cut short due to space limit.