

Do Factuality Methods Make LLMs More Factual? Reproducibility and Fair Comparison

Pratuat Amatya^{1,2} , Vinay Setty^{1,2} , Venkatesh V³ , Avishek Anand⁴ 

¹Factive AS, ²University of Stavanger, ³Stockholm University, ⁴TU Delft

pratuat@factiveverse.ai, vinay@factiveverse.ai, venkatesh.viswanathan@dsu.se, avishek.anand@tudelft.nl

Abstract

Large Language Models (LLMs) are aligned to human preferences, after pre-training and supervised fine-tuning steps. However, standard RL alignment steps such as RLHF do not focus on factuality of LLMs, hence they may generate non-factual information, which could result in unintended consequences. In response, a number of factuality alignment approaches have been proposed in the literature which claim to improve factuality of LLMs on their own benchmarks. However, these methods have not been reproduced independently and are also not comparable since each uses a different base model, training data, evaluation set, judge, and metric. In this paper, we present the first independent reproduction of five factuality alignment methods, which span on-policy RL (PPO, GRPO, and Dr. GRPO) and offline preference optimization (DPO and KTO), and compare them on a common benchmark with common evaluation metrics. We found that none of these methods reproduces in full, and on the common benchmark none of the methods significantly outperforms the base model. On closer inspection, the rewards optimize for format, length, or brevity rather than factuality and relevance, whereas the choice between on-policy and offline training matters little. Furthermore, the evaluation metrics and benchmarks are also part of the problem since they do not measure relevance of text to the prompt. Since there is no ground truth associated with the prompts, the reliance on LLM judges and retrieval for verification also makes the evaluation fragile. We release our code, data, and checkpoints.

1 Introduction

Large Language Models (LLMs) are aligned to follow instructions as part of their standard training process. But alignment methods such as Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017) are shown to de-

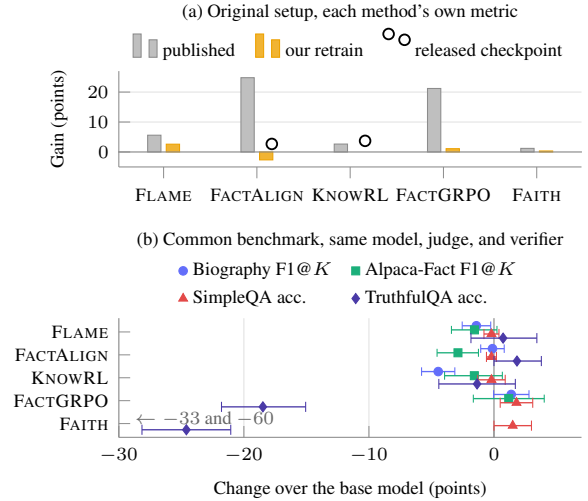


Figure 1: Published gains vs. reproduced. (a) Gain over each method’s own baseline and as we measure it in the original setup, with released checkpoints as circles. (b) Change over the base model on the common benchmark with 95% bootstrap intervals.

grade factuality of LLMs (Lin et al., 2024). In response, a number of factuality alignment methods have emerged: factuality-aware SFT and DPO (Lin et al., 2024), fine-grained KTO over sentence-level labels (Huang and Chen, 2024), reinforcement learning against a knowledge-boundary reward (Ren et al., 2026), a programmatic reward over long chains of thought (Chen et al., 2026), and PPO conditioned on an estimated knowledge state (Dong et al., 2026).

However, none of these gains have been independently reproduced, and it is also not feasible to compare them with each other since these five methods use different base models, training data, evaluation sets, judges, and metrics. The improvements reported by these papers are therefore neither replicable nor generalizable.

This paper is the first reproducibility effort on all five methods. Reproducing these methods is not straightforward. For example, the authors of

FLAME do not release the code, and even when the code is released, the configuration and setup used in the code deviate from what is documented in the paper (Table 5 summarizes what the authors released). We try our best to reproduce as faithfully as possible to the setup described in the papers and document every deviation (Figure 2 and Appendix A).

To make the comparison feasible and fair, we also train every method on a single base model, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), with the training regime and data that its paper specifies, and evaluate on a **common benchmark** with two long-form generation tasks scored with VERIScore (Song et al., 2024) and two short-form question answering tasks graded by an LLM judge.

Our study is driven by the following research questions:

- **RQ I:** Do the published gains hold when we reproduce each method, first in its own setup and then on a common benchmark?
- **RQ II:** How does the reward design affect long-form and short-form factuality?
- **RQ III:** Does the training paradigm, on-policy reinforcement learning or offline preference optimization, affect factuality?
- **RQ IV:** Do the current evaluation metrics measure factuality faithfully?

Our study yields four findings. First, **none of the methods reproduce the published results completely** when we rebuild the training pipeline independently. For instance, the factuality gain of FLAME reproduces, but its claim that factuality alignment costs no helpfulness does not, and FACTALIGN reproduces with its released checkpoint but not when retrained. On the common benchmark, none of the methods significantly outperforms the untrained base model.

Second, the rewards optimize for length, format, or brevity rather than factuality and relevance, and the models exploit this, which is commonly referred to as **reward hacking**. For instance, the FACTGRPO reward counts a claim as correct whenever evidence supports it, even if it is irrelevant to the prompt, and 36% of its biography responses contain such claims. The KNOWRL and FAITH rewards are built for short questions with a gold answer and break down on long-form prompts, where the answers of KNOWRL grow long and imprecise and those of FAITH shrink to one word.

Third, **the training paradigm matters less than the reward**. On-policy RL and offline preference

optimization perform similarly on the common benchmark, and the alignment step barely changes the model from the supervised fine-tuning stage.

Finally, **the evaluation itself is part of the problem**. The factuality metrics used in benchmarks such as FActScore (Min et al., 2023), SAFE (Wei et al., 2024b), and VERIScore (Song et al., 2024) credit claims supported by evidence irrespective of the claim’s relevance to the prompt. Filtering out irrelevant claims reduces precision by 4–15 percentage points across all methods that answer at length. Without ground truth, the scores also depend on the evidence source, the recall normalizer, and the LLM judge, each of which moves the results more than any method.

We make the code, configurations, and evaluation scripts¹ as well as data and checkpoints² publicly available (anonymously).

2 Related Work

2.1 Factuality Alignment

Factuality in LLMs refers to the capability of the LLMs to generate content that adheres to factual information grounded in common sense, domain-related facts, and world knowledge (Wang et al., 2026). While instruction-tuned LLMs have led to tremendous advances in a wide range of natural language tasks, they still suffer from issues such as hallucination, which might result in non-factual outputs. To improve factuality, among training-free strategies are self-consistency-based uncertainty estimation during decoding (Tian et al., 2023) and other decoding approaches (Kadavath et al., 2022). Alternatively, post-training methods such as SFT (Tao et al., 2024) and policy optimization (Dong et al., 2026; Chen et al., 2026; Ren et al., 2026; Huang and Chen, 2024) have been proposed for factuality alignment of such models. Policy optimization approaches can be on-policy or off-policy. Approaches like FactTune (Tian et al., 2024) construct preference pairs using generated responses and FActScore (Min et al., 2023) as a feedback signal. They then employ Direct Preference Optimization (DPO) in an off-policy fashion over this collected preference pairs dataset. FLAME (Lin et al., 2024) also proposes a similar approach but collects self-generated traces for factuality-aware alignment, combining factuality-aware SFT and a

¹<https://anonymous.4open.science/r/factuality-alignment-reproducibility/>
²<https://huggingface.co/factualityalignment>

factuality-aware RL loop through DPO. KnowTuning (Lyu et al., 2024) proposed fine-grained knowledge augmentation to identify difficult knowledge. They also employ DPO to help LLMs differentiate between reliable and unreliable knowledge. Alternatively, on-policy methods have demonstrated that training on self-generated traces during optimization is better than training on reference traces or static self-generated traces, which may cause more hallucination due to distribution mismatch (Kang et al., 2025; Gekhman et al., 2024; Ghosal et al., 2024). With the success of DeepSeek employing on-policy RL using GRPO, similar RL-based tuning approaches have been proposed for factuality alignment in LLMs (Zhang et al., 2024). While these approaches mostly focus on short-form generation, approach like (Chen et al., 2026) propose factuality-aware and aligned reasoning for long-form responses.

2.2 Factuality Evaluation

Several evaluation setups and metrics have been proposed for factuality in short-form and long-form responses from LLMs. Evaluating the factuality of long-form responses is hard, and hence prior approaches have proposed breaking them down into atomic claims to be verified (Kamoi et al., 2023; Gao et al., 2023; Guan et al., 2024; Wei et al., 2024b). FActScore (Min et al., 2023) extracts atomic facts, which are statements that each contain a single piece of information. However, its extraction method is specific to the domain of biographies and is inapplicable to other domains. SAFE (Wei et al., 2024b) extends to domains beyond biography and adopts a three-stage pipeline comprising a claim extraction stage, claim revision stage to resolve vague references, and a claim relevance check stage. While this leads to better evaluation compared to FActScore, Song et al. (2024) argue that the relevance check stage also significantly removes verifiable claims, undermining evaluation of the factuality of generated text. VERIScore (Song et al., 2024) was proposed to overcome these limitations and avoid the fallacy of treating all atomic claims as verifiable. VERIScore extracts only verifiable claims and is also one of the first methods to consider inter-sentence context when extracting claims. This also helps avoid the cost incurred by claim revision.

Existing factuality alignment approaches are not evaluated in a common setup with common benchmarks and a reliable metric. Our study primarily

contributes to the rigorous evaluation of factuality alignment methods for LLMs in such a standardized setup and to an analysis of the reliability of its factuality metrics.

3 Factual Alignment Methods

No base model and no evaluation dataset is common to all five papers, and their factuality metrics differ in kind, since three are precision-like, one credits refusals, and one excludes them from its denominator (Table 26 in the appendix).

3.1 Method Descriptions

Let π_θ denote the LLM-based policy being aligned and π_{ref} for its frozen initialization. A prompt / question x elicits a response y . For long-form evaluation y decomposes into atomic claims $c_1, \dots, c_T$, of which S are judged supported against an evidence source; precision is $P = S/T$, recall at a fixed budget K is $R_K = \min(1, S/K)$, and $F1@K$ is their harmonic mean. We follow Chen et al. (2026) in also calling S the *detail level*, which indicates the number of correct facts in the response. For short-form evaluation a response is graded CORRECT, INCORRECT, or NOT ATTEMPTED. We write $r(x, y)$ for a scalar reward and $d(y) \in \{+, -\}$ for a binary desirability label.

Figure 2 summarizes the five pipelines, which differ in the signal of the reward r , the granularity at which it is applied, and what is held fixed while it is optimized.

FLAME (Lin et al., 2024) makes both alignment stages factuality-aware. In the SFT stage, each instruction from seed data is classified as being fact-based or not. For fact-based instructions x , the SFT targets are the policy’s *own* generations instead of external text, to prevent hallucinations from training on unknown data. For non-fact-based instructions, the authors propose to leverage human responses. For DPO (Rafailov et al., 2023) training, responses to fact-based instructions are ranked by factual precision, $r_{\text{fact}} = S/T$. Responses to other instructions are ranked by an instruction-following reward. The preference data are generated before training and remain fixed as the policy is updated. Training is therefore off-policy, and the responses in the preference data may become less representative of the current policy.

FACTALIGN (Huang and Chen, 2024) replaces pairwise preference with KTO (Kahneman-Tversky Optimization) (Ethayarajh et al., 2024),

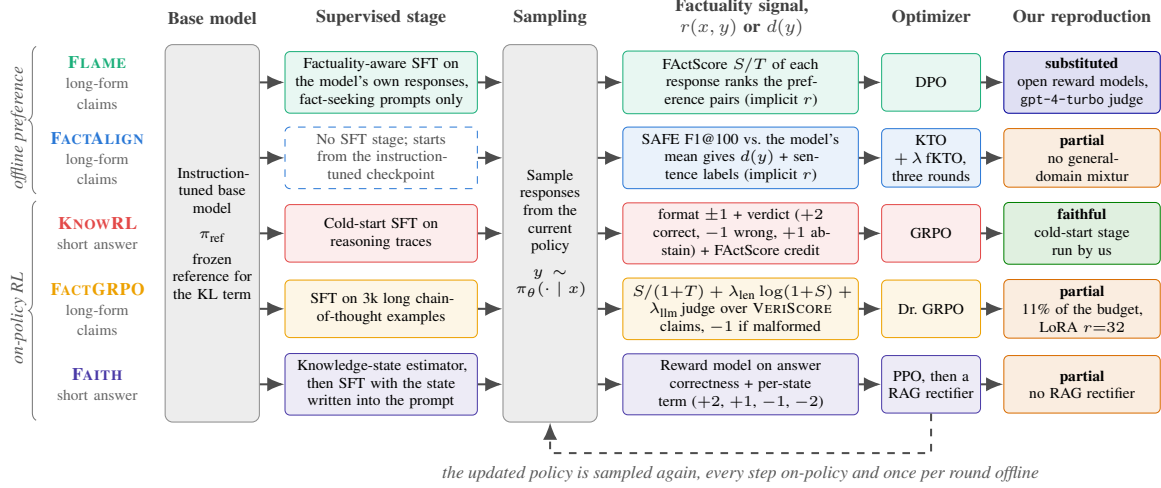


Figure 2: The five alignment methods and our reproductions. Gray boxes show shared stages, and colored boxes show differences between methods. FLAME and FACTALIGN use offline preference optimization; the other three methods use on-policy RL with explicit rewards. The final column records the main deviation in each reproduction. Further details appear in Section 4 and Appendix A.

which utilizes binary labels, $d(y)$ instead of pairwise preferences. Since the responses with corresponding rewards are collected offline which may cause discrepancy between assessed responses and the model being trained, the authors also propose an iterative optimization mechanism. Here, new samples are periodically sampled from the model being trained. A response is considered as desirable when its SAFE F1@100 exceeds a threshold set at the model’s own mean. Apart from response-level reward, the work’s contribution is also fine-grained sentence-level factuality alignment through *fine-grained* KTO (fKTO). Hence the loss also comprises a sentence-level term, $\mathcal{L} = \mathcal{L}_{\text{KTO}} + \lambda \mathcal{L}_{\text{fKTO}}$, with separate KL strengths β and β_{fg} .

KNOWRL (Ren et al., 2026) adopts GRPO (Shao et al., 2024) on top of a cold-start SFT stage, with factual supervision introduced into the chain-of-thought (CoT) reasoning process. The reward is of the form $r = r_{\text{format}} + r_{\text{correct}} + r_{\text{fact}}$. The correctness term gives +2 for a correct answer, −1 for an incorrect answer, and +1 for an explicit non-attempt. Abstaining therefore receives a higher reward than answering incorrectly. The authors argue that the cold-start stage is needed to keep training stable under this reward.

FACTGRPO (Chen et al., 2026) optimizes a *pro-grammatic* reward with Dr. GRPO (Liu et al., 2025) over a long chain of thought,

$$r = \frac{S}{1+T} + \lambda_{\text{len}} \log(1+S) + \lambda_{\text{llm}} \text{judge},$$

with $r = -1$ for a malformed chain of thought. The first term measures factual precision ($\mathcal{R}_{\text{fact}}$),

with 1 added to the denominator to avoid division by zero when no claims are extracted. The second term rewards detail ($\mathcal{R}_{\text{dtl}}$), measured by the number of supported claims S in the response, through $\log(1+S)$. The third term measures the response’s relevance to the question ($\mathcal{R}_{\text{rel}}$) using an LLM judge. The weights λ_{len} and λ_{llm} control the contributions of detail and relevance to the total reward. For $T > 0$, we can write the first term as $S/(1+T) = PT/(1+T)$, where $P = S/T$. At a fixed precision, this term increases with the number of claims. Thus, even with $\lambda_{\text{len}} = 0$, as in the released configuration, the reward retains a preference for more detailed responses of the same precision and does not remove incentives to produce more claims.

FAITH (Dong et al., 2026) conditions generation on an estimated *knowledge state*. For each training question, the method samples multiple responses and computes their consistency and semantic entropy. These statistics determine one of four states, defined by whether the model is estimated to possess the knowledge (K or $\neg K$) and whether its answer reflects that knowledge (H or $\neg H$). The paper calls these dimensions “trustworthiness” and “honestness.” The estimated state is written into the prompt, and PPO (Schulman et al., 2017) uses an additional state-dependent reward of +2, +1, −1, or −2.

The methods differ in both the unit of supervision and their treatment of abstention. FLAME, FACTALIGN, and FACTGRPO score claims extracted from long-form responses. KNOWRL and FAITH score short answers and give credit for ab-

stention. KNOWRL assigns +1 to a non-attempt and -1 to an incorrect answer, while FAITH rewards appropriate refusals. In FACTGRPO, $S/(1+T)$ rewards more claims at fixed positive precision.

4 Reproduction in the Original Setups

To answer **RQ I**, we first reproduce each method in its original setup, with the released code, data, and checkpoints (where available), and record for every component whether it is faithful, substituted, or partial (last column of Figure 2; Appendix A lists every deviation). Tables 27 and 26 in the appendix list the training stages, data, and metrics of each paper. We contacted the authors of all five papers about the discrepancies between the released artifacts and the papers, and only the FAITH authors had replied at the time of writing (Appendix A.5). Table 1 sets the published numbers against ours. We observe that all methods except FAITH and FACTGRPO follow a similar trend as their respective original implementations, although none reproduces their published results in full.

FLAME. We retrain all four models at 70B with open reward models and gpt-4-turbo in place of the unreleased reward models and the retired judge. The factuality claim reproduces, with a biography FActScore above the published one, whereas the claim that factuality alignment costs no helpfulness does not, since our AlpacaEval win rate falls well below the published one (Appendix A.1).

FACTALIGN. The released checkpoint improves over its base under two retrievers, as published. Our retrain on gemma-2b for three iterations without the general-domain mixture lands at the level of the paper’s own ablation without it, so the gain depends on that mixture rather than on the objective alone (Appendix A.2).

KNOWRL. Both our run with the cold-start stage followed by 150 GRPO steps and our evaluation of the released checkpoint reproduce the reported behavior without refusal collapse, albeit with lower converged rewards. In contrast, all runs using the released learning rate or omitting the cold-start stage collapse into refusals. However, closer inspection reveals that the released checkpoint abstains on most SimpleQA questions, and its score increases mechanically as refusals are excluded from the metric (Appendix A.3).

FACTGRPO. With 400 steps of Dr. GRPO, a tenth of the published optimization, and a rank-32 adapter, precision rises slightly while the supported

Quantity	Paper	Ours
FLAME		
Bio FActScore, SFT ^f +DPO ^f	47.9	54.2 ✓
AlpacaEval win rate vs. SFT+DPO	51.2	37.7 ✗
FACTALIGN, F1@100		
released checkpoint vs. base, paired	—	+2.68 ✓
our three-iteration run vs. its baseline	+24.8	-2.64 ✗
paper’s “— general-domain” ablation	-0.41	-2.64 ✓
KNOWRL		
refusal rate, final fifth of training	no collapse	0.2% ✓
combined reward at convergence	≈-0.1	-0.637 ≈
released checkpoint, SimpleQA	1.92→4.56	2.00→5.71 ✓
FACTGRPO (Chen et al., 2026), precision / detail level		
Biography precision	25.9→47.1	35.8→36.9 ✗
Alpaca-Fact precision	60.2→79.7	61.9→64.7 ✗
Biography detail level	10.3→14.8	12.1→10.9 ✗
Alpaca-Fact detail level	30.7→36.9	20.0→18.1 ✗
FAITH, precision, FAITH _{sft} →FAITH _{sft+ppo}	+1.20	+0.33 ≈
mean gain over four datasets (two reproduce)		
refusal rate	—	0.000 ✗

Table 1: Published and reproduced results under each method’s protocol, in points unless noted. Compare within rows only. ✓: reproduced; ✗: not reproduced; ≈: same direction, different magnitude.

count falls, against the published gains in both, and the reward stays flat, so this run bounds a tenth of the budget rather than the method (Appendix A.4). The rerun with the relevance term of the reward wired, which the first run had left out, behaves the same and is the variant used in Section 5.

FAITH. With four of the five stages on Llama-3-8B, without the rectifier and with the estimator-predicted state, the PPO gains reproduce on two of four datasets, at a fraction of the published size on average, and the model never refuses. Its SFT targets are the short gold answers of question answering sets and its reward model accepts a prediction that the gold answer contains, so both stages pull the policy toward one-word answers, which is what we measure on the long-form benchmarks (Appendix A.5).

5 Common Benchmark

5.1 Setup

Model and fine-tuning All five methods of Section 3 are trained on Llama-3.1-8B-Instruct, each alignment stage as a LoRA (Hu et al., 2022) adapter of rank 32 on top of the method’s own supervised stage with the hyperparameters of its paper. Since

the five methods carry too many settings to print, the full configuration of every run is listed in the code repository and summarized in Appendix D. All methods generate with greedy decoding from the bare question or instruction, with at most 1,024 new tokens for long-form prompts and 256 for short-form questions, and FAITH additionally receives its estimator-predicted knowledge state in the prompt, as its paper prescribes. **Benchmarks** Short-form evaluation uses the first 1,000 SimpleQA questions (Wei et al., 2024a) and all 817 TruthfulQA questions (Lin et al., 2022), graded as correct, incorrect, or not attempted by gpt-4o using the official SimpleQA prompt. We report accuracy, abstention rate, accuracy on attempted questions, and their harmonic mean F . Long-form evaluation uses 495 FACTSCORE biography prompts (Min et al., 2023) and 253 Alpaca-Fact instructions, selected from AlpacaEval by the FLAME fact-seeking classifier following Lin et al. (2024). We score responses with VERIScore (Song et al., 2024), which extracts verifiable claims and checks each against ten web search results. We report the average numbers of supported and unsupported claims per response, precision (assigning zero to responses with no claims), and $F1@K$, with K fixed at 20 and 19, the median claim counts of an earlier pool (Section 6). For each long-form table, we evaluate all models on the same subset of prompts; excluded prompts are listed in Appendix D. We test significance against the base model using McNemar’s exact test for short-form evaluation and paired bootstrap resampling for long-form precision, recall, and $F1@K$, with 10,000 samples. We apply Holm’s correction for multiple comparisons.

5.2 Overall Results

The common evaluation setup employs one base model, one decoding configuration, one judge, and one claim verifier, across all methods so that differences are attributable to training rather than to evaluation. We found that no method improves significantly over the untrained model on any of the four datasets, apart from two long-form cases discussed below. From Table 2, we observe that FAITH and FACTGRPO lose about 25 and 18 points of TruthfulQA accuracy, while the other three methods stay within two points of the base model, which is not statistically significant.

On long-form evaluation, FAITH has the highest precision on the biography task, but the gain is an artifact of 0.4 supported claims in a one-sentence

Method	Acc.	Abst.	Acc att.	F
SimpleQA ($n = 1,000$)				
base (untrained)	.034	.812	.181	.057
FLAME	.032	.851	.215	.056
FACTALIGN	.032	.813	.171	.054
KNOWRL	.032	.822	.180	.054
FACTGRPO	.052	.414	.089	.066
FAITH	.049	.000	.049	.049
TruthfulQA ($n = 817$)				
base (untrained)	.416	.108	.466	.440
FLAME	.424	.108	.475	.448
FACTALIGN	.435	.111	.489	.460
KNOWRL	.403	.124	.459	.429
FACTGRPO	.231	.368	.366	.284
FAITH	.170	.029	.175	.173

Table 2: Short-form evaluation, one judge (gpt-4o) across every method. *Abst.* is the share of questions the model declined; *Acc|att.* is accuracy over attempted questions only, which is the quantity a method can raise by declining more.

(very brief) response, which is why its relevance-filtered precision is significantly above the base model, but it has the lowest $F1@K$ (Table 3). Since FAITH is proposed as a general factuality alignment method for LLMs and not as specific to short-form tasks, the long-form benchmarks are within its stated scope. FACTGRPO has the highest $F1@K$ on both datasets with fewer claims than the base model. Its relevance-filtered $F1@K$ on biographies is .017 above the base model and is the only gain apart from the FAITH artifact that is statistically significant, which is a small fraction of the published improvement. On SimpleQA, FLAME raises its accuracy over attempted questions by declining more often while its overall accuracy stays unchanged, so the accuracy alone understates how differently these models behave. Hence, none of the methods provides substantial factuality gains in our common evaluation setting.

5.3 Reward Design

To answer **RQ II**, we analyze the effect of rewards on responses. The FACTGRPO reward is a function of the number of supported and total claims, with a relevance term that contributes little, so a claim that is true but does not address the prompt earns the same reward as one that does. We call such claims filler claims. They occur in about two fifths of the long-form responses of every method that answers at length (Figure 4 in the appendix), and the FACTGRPO reward rises with their number (Figure 3), since the factuality term $\mathcal{R}_{fact}$ dominates the relevance term $\mathcal{R}_{rel}$ (Table 21 in the appendix). The model can therefore hack the reward

Method	#Claims	VERIScore				VERIScore+ Relevance			
		#Sup.	P	R	F1	#Sup.	P	R	F1
<i>Biography</i> ($n = 490, K = 20$)									
base (untrained)	27.6	12.0	0.351	0.472	0.370	10.6	0.283	0.434	0.325
FLAME	23.5	10.6	0.338	0.436 [†]	0.356 [†]	9.7	0.294	0.407 [†]	0.324
FACTALIGN	27.1	11.7	0.343	0.468	0.368	10.4	0.284	0.431	0.326
KNOWRL	29.8	11.3	0.281 [†]	0.440 [†]	0.325 [†]	10.0	0.237 [†]	0.409 [†]	0.291 [†]
FACTGRPO	25.4	10.1	0.342	0.466	0.383	9.0	0.301	0.423	0.342 [†]
FAITH	0.6	0.4	0.374	0.019 [†]	0.037 [†]	0.4	0.372 [†]	0.019 [†]	0.036 [†]
<i>Alpaca-Fact</i> ($n = 248, K = 19$)									
base (untrained)	29.4	19.9	0.618	0.681	0.615	15.9	0.491	0.586	0.500
FLAME	26.6	18.2	0.621	0.653 [†]	0.599	14.5	0.500	0.565	0.492
FACTALIGN	28.1	18.1	0.598	0.651 [†]	0.586 [†]	14.4	0.484	0.556 [†]	0.478
KNOWRL	33.1	20.8	0.582	0.689	0.599	16.9	0.471	0.612	0.500
FACTGRPO	25.4	17.6	0.624	0.677	0.626	13.2	0.471	0.566	0.489
FAITH	0.1	0.1	0.101 [†]	0.006 [†]	0.012 [†]	0.1	0.097 [†]	0.006 [†]	0.012 [†]

Table 3: Pooled results for long-form factuality under VERIScore and VERIScore with relevance filtering evaluation. #Claims is the mean number of extracted claims per response and #Sup. the mean number of supported claims; bold marks the highest value per column within each dataset. [†] marks a significant difference from base (paired bootstrap test, Holm-corrected $p < 0.05$).

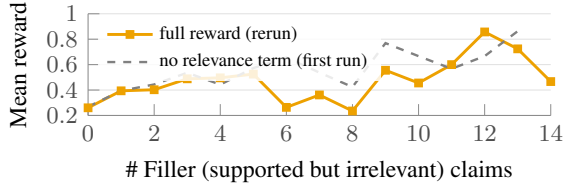


Figure 3: Mean FACTGRPO reward as a function of the number of filler (supported but irrelevant) claims in a long-form biography response, for the rerun with the full reward and for the first run without the relevance term. Bins beyond 14 hold one or two responses and are not shown.

by generating claims that are factual but off-topic. The KNOWRL reward adds a format term, a verdict term, and a factuality term, but on free-form prompts only the verdict term carries a signal, since the model never emits the expected tags and the factuality term retrieves Wikipedia pages by title, which questions do not have. The FAITH reward credits brevity, as described above.

When we score every method’s generations with every method’s reward, normalized by response length, the rewards disagree with each other and with the benchmarks (Table 19 in the appendix). Only the FACTALIGN and FACTGRPO rewards prefer their own generations, the latter only because every other method fails its format check, whereas the FLAME reward ranks the base model’s generations above FLAME’s own. Where the KNOWRL and FAITH rewards prefer a method over the base model, it is another method’s generations, as both rate FLAME’s generations above the base model’s, which none of the benchmarks does. The rewards

therefore credit length, format, or brevity rather than the factuality that the benchmarks measure.

5.4 Training Paradigm

To answer **RQ III**, we group the methods by paradigm. KNOWRL, FACTGRPO, and FAITH optimize on-policy with GRPO, Dr. GRPO, and PPO, whereas FLAME and FACTALIGN optimize offline with DPO and KTO on preference data. Neither group separates from the base model on the common benchmark, and the two large losses on TruthfulQA both come from on-policy methods whose training changed the output format. To separate the paradigm from the supervised data it starts from, Table 15 shows each supervised stage on its own and Table 2 the pipeline that follows it. On the short-form benchmarks the alignment step adds almost nothing, since no method moves by more than .007 in accuracy and none of the differences is significant, although individual answers do change, as FACTGRPO turns 38 TruthfulQA items from incorrect to correct and 42 the other way. On the long-form benchmarks (Table 16 against Table 3) the alignment step moves precision and F1@K by less than three points, and none of the differences is significant. We believe this is because the supervised stage already sets the format and the length of the answers, and the objective that follows has little room to change what the metrics count.

6 Reliability of Factuality Evaluation

To answer **RQ IV**, we examine short- and long-form evaluation pipelines whose components are

stochastic, lack ground truth, or depend on the model pool. We thus treat RQ IV as a measurement question rather than a flaw of any single metric.

Relevance. VERIScore credits supported claims even when they are irrelevant or repeated (Appendix, Table 23). Relevance filtering lowers precision by 4–7 points on biographies and 11–15 on Alpaca-Fact for methods producing long responses (Table 3). Filler claims occur in 178–223 of 490 biography responses per method, but only once for FAITH (Appendix Figures 4 and 5). The FACTGRPO reward also credits these claims, enabling the reward hacking discussed under RQ II (Figure 3).

Extraction and scoring conventions. Two scoring runs return identical claim lists for fewer than half of the same responses. We use paired tests, though these do not remove extraction noise. Zero-claim responses receive zero precision; excluding them raises FAITH’s Alpaca-Fact precision from 0.101 to 0.781.

Recall normalizer. Changing the model pool changes its median claim count K , shifting $F1@K$ by more than .1 for identical responses (Table 22). We fix K at 20 for biographies and 19 for Alpaca-Fact. Appendix C.3 examines the medians after adding the FACTGRPO rerun.

Evidence source. A true claim without retrieved support receives the same label as a false one. Re-verifying biography claims against the subject’s Wikipedia article lowers precision by about twenty points for methods producing long responses and changes their ranking (Appendix C.4). The article passages support only two fifths of the claims supported by web search. Changing the evidence source thus affects scores more than changing the method.

Judge. The short-form judge decides the absolute numbers, since regrading with gpt-oss-120b changes up to a third of the TruthfulQA labels (Table 20), although the large drops of FACTGRPO and FAITH persist.

Human audit. Two of the authors labeled the same 300 claims independently, 50 per variant and balanced over the automatic labels, without seeing those labels (Appendix C.5), and both judged 24 of them not to be claims at all. VERIScore agrees with their consensus on 84% of the claims (Cohen’s

$\kappa = 0.65$), counting claims that the evidence does not settle as unsupported, while the annotators agree with each other on 78% ($\kappa = 0.50$). Where both annotators agree, VERIScore is right for 97% of the claims it accepts but only 65% of those it rejects, so its precision is, if anything, an underestimate. The relevance labels agree with the consensus on 86% ($\kappa = 0.62$), and the LLM is right for 99% of the claims it keeps but only 55% of those it removes, so the drops under the relevance filter are an upper bound.

The limitations discussed above underscore that both long-form and short-form scores serve as mere proxies for factuality rather than direct measures. Long-form evaluation assesses the consistency of extracted claims with retrieved evidence, whereas short-form evaluation relies on a judge’s assessment of agreement with a reference answer. Consequently, the lack of improvements on these metrics does not rule out gains in factuality in other ways. Future comparisons should use a common evaluation, report relevance-aware scores with claim counts and abstention rates, and release implementations matching the paper.

7 Conclusions

In this paper, we examined the reproducibility and comparability of five factuality alignment methods, evaluating each under its original experimental setup and under a shared setup with a common base model, judge, and verifier. None fully reproduced its published results. Under the shared setup, no method significantly outperformed the untrained base model, except for an artifact associated with one-word answers and a small gain after relevance filtering. Our analysis further suggests that the rewards favor length, format, or brevity rather than the factuality measured by the benchmarks, and that on-policy and offline training perform alike. We also found that evaluation choices, including relevance filtering, evidence sources, recall normalization, and the judge, affected scores more than the methods themselves. These findings highlight the need for shared evaluation settings, relevance-aware metrics accompanied by claim counts and abstention rates, and released implementations consistent with the reported experiments. Future work could consider additional methods and larger models, expand the human evaluation to more annotators and claims, and investigate rewards that credit only relevant, supported claims.

Limitations

Our common evaluation uses Llama-3.1-8B-Instruct with LoRA adapters, four English datasets, and smaller training budgets than the original papers. Results may differ with larger models, longer training, or full fine-tuning. Each method was trained once, leaving variance across seeds unmeasured.

We document substitutions for unavailable artifacts, but cannot fully separate their effects from those of the methods. We also cannot inspect the training conditions of released checkpoints. Only one of the five author groups contacted had replied at the time of writing, leaving some discrepancies unresolved.

All scores rely on automatic claim extraction, web-based verification, commercial judges whose snapshots may change, and relevance labels from one LLM. Our human audit covers 300 claims checked by two annotators. Reference verification covers biographies only, as other datasets lack reference text. The fixed recall normalizer is one of several reasonable choices.

Ethical Considerations

All data come from public benchmarks. The released biography generations contain false claims about real people and are intended for factuality research, not as sources of biographical information.

We follow the models’ licenses and release code, configurations, and adapters rather than full weights. The appendix reports GPU usage and the code repository estimated API costs; billed amounts were not recorded. The human audit was carried out by two of the authors.

Our reward-hacking results show that higher automatic scores need not imply more truthful answers. These scores alone should not certify a system’s factual reliability.

Use of AI Assistants

We used Claude Code to help run experiments, to generate scripts for analyzing results and producing LaTeX plots and diagrams. We also used Opus 5.5 to rephrase parts of the paper to improve readability in addition to fixing typos and grammatical issues. All AI-assisted work was carried out under the authors’ supervision. The authors take full responsibility for the resulting code, experimental procedures, analyses, and claims made in the paper, including any errors that remain.

References

- Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. 2025. [FactBench: A dynamic benchmark for in-the-wild language model factuality evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 33090–33110. Association for Computational Linguistics.
- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. [Semantic parsing on Freebase from question-answer pairs](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.
- Mingda Chen, Yang Li, Xilun Chen, Adina Williams, Gargi Ghosh, and Scott Yih. 2025. [FACTORY: A challenging human-verified prompt set for long-form factuality](#). Preprint, arXiv:2508.00109.
- Xilun Chen, Iliia Kulikov, Vincent-Pierre Berges, Barlas Oguz, Rulin Shao, Gargi Ghosh, and Wen-tau Yih. 2026. [Learning to reason for factuality](#). In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 15003–15022.
- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.
- Xiaoning Dong, Chengyan Wu, Yajie Wen, Yu Chen, Yun Xue, Zhang Jing, Wei Xu, and Bolei Ma. 2026. [FAITH: factuality alignment through integrating trustworthiness and honesty](#). In *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pages 13975–13994. Association for Computational Linguistics.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling, COLM ’24*.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. [KTO: model alignment as prospect theoretic optimization](#). volume abs/2402.01306.
- Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. [RARR: researching and revising what language models say, using language models](#). In *Proceedings of the 61st Annual Meeting*

- of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pages 16477–16508. Association for Computational Linguistics.
- Zorik Gekhman, Gal Yona, Roei Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. [Does fine-tuning LLMs on new knowledge encourage hallucinations?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784, Miami, Florida, USA. Association for Computational Linguistics.
- Gaurav Rohit Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. 2024. [Understanding finetuning for factual knowledge extraction](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 15540–15558. PMLR / OpenReview.net.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. [Language models hallucinate, but may excel at fact verification](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 1090–1111. Association for Computational Linguistics.
- Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Dekai Sun, Zhicheng Zheng, Wenbo Su, and Bo Zheng. 2025. [Chinese SimpleQA: A Chinese factuality evaluation for large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 19182–19208. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Chao-Wei Huang and Yun-Nung Chen. 2024. [FactAlign: Long-form factuality alignment of large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 16363–16375. Association for Computational Linguistics.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *CoRR*, abs/2207.05221.
- Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. [WiCE: Real-world entailment for claims in Wikipedia](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7561–7583. Association for Computational Linguistics.
- Katie Kang, Eric Wallace, Claire J. Tomlin, Aviral Kumar, and Sergey Levine. 2025. [Unfamiliar finetuning examples control how language models hallucinate](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 3600–3612. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural Questions: a benchmark for question answering research](#). *Trans. Assoc. Comput. Linguistics*, 7:452–466.
- Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Scott Yih, and Xilun Chen. 2024. [FLAME: Factuality-aware alignment for large language models](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 3214–3252. Association for Computational Linguistics.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. [Understanding r1-zero-like training: A critical perspective](#). *CoRR*, arXiv:2503.20783.

- Youngang 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](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14535–14556. Association for Computational Linguistics.
- Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. 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, COLM ’24*.
- 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 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A graduate-level Google-proof Q&A benchmark](#). In *First Conference on Language Modeling, COLM ’24*.
- Baochang Ren, Shuofei Qiao, Ningyu Zhang, Da Zheng, and Huajun Chen. 2026. [KnowRL: Exploring knowledgeable reinforcement learning for factuality](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pages 39640–39658. Association for Computational Linguistics.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *CoRR*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. [VeriScore: Evaluating the factuality of verifiable claims in long-form text generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 9447–9474. Association for Computational Linguistics.
- Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. [When to trust LLMs: Aligning confidence with response quality](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 5984–5996. Association for Computational Linguistics.
- 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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5433–5442. Association for Computational Linguistics.
- Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Qipeng Guo, Xiangkun Hu, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Xuming Hu, Zehan Qi, Wenyang Gao, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2026. [Survey on factuality in large language models](#). *ACM Comput. Surv.*, 58(1):13:1–13:37.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. [Measuring short-form factuality in large language models](#). *Preprint*, arXiv:2411.04368.
- 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. 2024b. [Long-form factuality in large language models](#). In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, pages 94–106. Association for Computational Linguistics.
- 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)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 1946–1965. Association for Computational Linguistics.

A Reproductions in the Original Setups

This appendix gives, for each method, the setup we reproduced, the deviations from the paper, and the results under the paper’s own metric. The summary is in Section 4 and the published-versus-ours comparison in Table 1.

A.1 FLAME

The paper fine-tunes Llama-2 70B in four variants, SFT and SFT followed by DPO as the standard pipeline, and their factuality-aware counterparts SFT^f and FLAME, which is SFT^f followed by DPO^f. We retrain all four with data rebuilt from the published sources, 3,200 Open Assistant pairs for instruction following and 1,630 judge examples for evaluation fine-tuning, 20K self-generated instructions, and preference data ranked by an instruction-following and a factuality reward model. Full fine-tuning of the 70B model runs with DeepSpeed ZeRO-3 and optimizer offload on two H200 GPUs, at about 260 seconds per SFT step and 500 per DPO step. The paper mixes its instruction and evaluation fine-tuning data in the SFT stage, whereas our SFT^f set is built from the 3,200 instruction-following prompts alone, with the fact-seeking ones answered by Llama-2-70B samples. Three components are substituted. The paper’s reward models are not released, so the instruction-following reward is the 70B SFT model judging itself and the factuality reward is the recovered InstLLaMA checker. The factuality of open-ended instructions is scored with the same pipeline over an open retriever rather than with article-based FActScore. The AlpacaEval judge is gpt-4-turbo, since the paper’s gpt-4-1106-preview is retired.

Table 4 shows that standard DPO lowers biography FActScore from 53.9 to 51.6, while factuality-aware DPO lowers it from 54.6 to 54.2. FLAME exceeds SFT+DPO by 2.6 points on biographies and 2.5 on Alpaca-Fact, against a published biography gain of 5.6. SFT^f also improves TruthfulQA BLEU and ROUGE-L over SFT, but FLAME shows no FAVA advantage with our open retriever.

The helpfulness claim does not reproduce: our AlpacaEval win rate is 37.7%, or 38.6% length-controlled, against the reported 51.2%. Substituted reward models may contribute to this gap, but one training run and a newer judge do not allow us to isolate the cause.

Llama-2 70B	Bio	A-Fact	FAVA	Win%
SFT	53.9	44.3	54.7	26.1
SFT + DPO	51.6	43.7	51.5	50.0
SFT ^f	54.6	45.5	54.2	19.3
FLAME	54.2	46.2	50.9	37.7

Table 4: FLAME at 70B: FActScore on 183 biographies, 253 Alpaca-Fact prompts, and FAVA; AlpacaEval win rate against SFT+DPO using gpt-4-turbo. Published biography scores are 44.7, 42.3, 49.5, and 47.9 for the four rows.

A.2 FACTALIGN

The paper runs three iterations of fine-grained KTO, mixing newly generated responses with general-domain preference data, and evaluates SAFE F1@100 on LongFact. We test both the released checkpoint and retraining.

We score the checkpoint and Llama-3-8B-Instruct together on 190 prompts using one judge session and retriever state. Repeated evaluation of one model on 64 prompts gives a 1.51-point difference (CI: −2.05 to +5.04), encompassing our initial unpaired gain of +2.72. The paired gain is +2.68 (CI: +1.28 to +4.15; $p = .0003$). Claim counts remain near 46, while unsupported claims fall by 19% and supported claims rise by 6% (Table 6). Replacing web search with the rebuilt December 2021 Wikipedia index of 33 million passages gives a +3.13-point gain.

Our three-iteration gemma-2b run uses the paper’s effective batch size of 16 but omits the general-domain mixture. F1@100 falls from 0.428 to 0.401, a 2.64-point loss (CI: −4.49 to −0.90); the first iteration scores 0.366. This agrees in direction with the paper’s no-mixture ablation, which falls from 61.74 to 61.33.

Across 570 generations per model, mean length falls from 705 to 593 words, and responses repeating a sentence at least three times fall from 33% to 19%. Claim density nevertheless rises from 7.7 to 10.0 per 100 words, with 1–253 claims per response. Unsupported claims grow faster than supported ones. Beyond 100 supported claims, recall cannot increase, while unsupported claims can still

Method	Code	Training data	Checkpoints	Reward or judge	Released code vs. paper
FLAME	✗	prompts only; SFT and DPO data are self-generated	✗	reward models ✗	no code to compare
FACTALIGN	✓	self-generated; general-domain mixture?	✓	SAFE pipeline ✓	evaluates a LongFact variant regenerated in 2026, which offsets every absolute number by about 19 points
KNOWRL	✓	✓	✓	API judge, snapshot drifts	cold-start stage via LLaMA-Factory with the released data; the released RL configuration sets learning rate 10^{-4} and rank 128 where the paper gives 10^{-5} and 256, with the base model left as a placeholder; the fact reward receives the response without its opening reasoning tag
FACTGRPO	✓	SFT data ✓; RL prompts recovered from the DPO pairs (3,706 of 4k)	?	programmatic (VERIScore) ✓	$\lambda_{\text{len}} = 0$ in the released configuration
FAITH	✓	✓, with the estimator’s predicted state in every prompt	?	reward model trained from released code	one epoch for the reward model instead of two and 0.45 PPO epochs instead of about two in the Llama-3 configurations; the rectifier ships a Mistral configuration but no training data (provided by the authors on request) and no Llama-3 configuration

Table 5: What each method releases and where the released code departs from its paper. ✓ released, ✗ not released, *partial* when only a part is usable, and ? when it is unclear.

lower precision.

Our LongFact version was regenerated in 2026 rather than 2024, shifting absolute scores by about 19 points on identical weights. We therefore compare models within the same evaluation.

	Llama-3-8B-Inst.	FACTALIGN
claims	46.00	45.22
supported	30.54	32.39
not supported	14.71	11.98
precision	0.675	0.730
F1@100	0.4145	0.4413

Table 6: FACTALIGN’s released checkpoint against its base, both rated in one paired SAFE pass over 190 prompts. Counts are per response.

A.3 KNOWRL

The paper applies cold-start SFT followed by GRPO. Its reward combines format (± 1), correctness ($+2$ correct, -1 incorrect, $+1$ refusal), and FactScore over reasoning for incorrect, non-refused answers. The repository provides cold-start data and refers to LLaMA-Factory, but leaves the RL model path unspecified. Its learning rate and adapter rank are 10^{-4} and 128, against 10^{-5} and

256 in the paper.

Our first four runs reach 97%–100% refusals in the final fifth of training, with completion length falling from 344 to 89 tokens (Table 7). Two use released hyperparameters, with and without cold start; two use the paper’s hyperparameters without cold start. Cold-start SFT followed by 150 GRPO steps at 10^{-5} and rank 256 instead keeps refusals near 0.2% and reward near -0.64 . Cold start alone does not prevent collapse at the released learning rate.

The correctness term gives an attempt expected reward $3p - 1$, against $+1$ for refusal, favoring refusal when $p < 2/3$. A separate formatting error omits the opening reasoning tag from the factuality reward’s input, reducing extracted facts from 11.8 to 5.3 per response and leaving two fifths without facts. Repairing the tag restores the signal but does not prevent collapse. Our first cold-start model reduces refusals from 22.7% to 3.7% on 300 questions.

The checkpoint’s SimpleQA gain reproduces: 1.92 to 4.56 in the paper, and 2.00 to 5.71 on our 500 items (Table 8). However, accuracy excludes refusals. Rejection rises from 10.2% to 93.0%,

and correct answers fall from nine to two. The higher score therefore concerns a much smaller set of attempted questions.

Run	1	2	3	4	5
KNOWRL, end to end	0.2	0.2	0.1	0.2	0.2
KNOWRL-ZERO, paper hyperparameters	7.2	14.2	33.7	93.1	96.8
+ repaired fact input	7.4	15.2	82.2	98.3	98.3
released hyperparameters, raw model	65.2	99.6	99.9	100.0	100.0
released hyperparameters, after cold start	0.9	15.7	85.2	99.2	98.4

Table 7: KNOWRL refusal rate in percent by fifth of training. The method as specified does not abstain, whereas every run at the released learning rate or without the paper’s cold start collapses, and repairing the input of the fact reward does not prevent the collapse, which locates the cause in the cold-start stage and the learning rate rather than in the fact signal.

	Base distill	KNOWRL
Reported accuracy	1.92	4.56
our measurement	2.00	5.71
Rejected	10.2%	93.0%
Correct answers, of 500	9	2

Table 8: The released KNOWRL checkpoint on 500 SimpleQA items under the authors’ protocol. The reported gain reproduces, but rejection rises from 10% to 93% and the number of correctly answered questions falls.

A.4 FACTGRPO

Mean rewards across 80-step blocks are 0.704, 0.704, 0.686, 0.741, and 0.718: a net rise of 0.014 against batch fluctuations up to 0.9. No malformed response occurs in 400 steps. Before evaluation, we predicted a gain far below the published result. Precision rises by 1.1 points on biographies and 2.8 on Alpaca-Fact, roughly a tenth of the published gains, while detail level falls (Table 11).

Although $S/(1 + T)$ rewards more claims at fixed positive precision, our reduced-budget run does not recover the reported joint improvement. The required format appears in 83% of biography and 60% of Alpaca-Fact responses, but only 6% of SimpleQA and under 1% of TruthfulQA responses. SimpleQA answers are almost five times the base model’s length. Most short-form outputs would therefore incur the model’s own format penalty.

Reward ablation. The first run omitted the relevance term. We repeated training with the full

reward and unchanged data, budget, and adapter; all common benchmark tables use this rerun. The runs receive identical grades on 80% of TruthfulQA and 81% of SimpleQA items, with no significant difference under McNemar’s exact test (Tables 9 and 12). Long-form scores differ by less than two points. The judge prefers the rerun’s responses to the base model’s in about one third of cases, contributing .03 to a mean reward near .3. We find no measurable benefit from this term under our budget.

Method	Acc.	Abst.	Acc att.	F
SimpleQA ($n = 1,000$)				
without the relevance term	.053	.401	.088	.066
full reward	.052	.414	.089	.066
TruthfulQA ($n = 817$)				
without the relevance term	.237	.355	.368	.289
full reward	.231	.368	.366	.284

Table 9: Short-form results of the two FACTGRPO runs on the common benchmark, graded by gpt-4o. The first run was trained without the relevance term of the reward, and the rerun with the full reward is the variant used in the main tables.

Table 10: Our FACTGRPO run against the published configuration of Chen et al. (2026). The last three rows are the deviations that matter.

	Paper	Ours
Base model	Llama-3.1-8B-Inst.	same
SFT data	3k synthetic	their released 3k
RL prompts	4k synthetic	3,706 recovered
Rollouts	4	4
Learning rate	1e-6	1e-6
Objective	Dr. GRPO	Dr. GRPO
Training	~3,706 steps	400 steps (11%)
Fine-tuning	full	LoRA $r=32$
Hardware	40 H100	1 H100 + 1 A100, 3 A100 (rerun)

Table 11: VERIScore precision (P) and detail level (S) on the long-form benchmarks, base model against our 400-step run, with the published effect for reference.

	Biography		Alpaca-Fact	
	P	S	P	S
Base	35.8	12.1	61.9	20.0
+ our GRPO	36.9	10.9	64.7	18.1
Paper, base	25.9	10.3	60.2	30.7
Paper, +GRPO	47.1	14.8	79.7	36.9

Table 12: FACTGRPO with and without the relevance reward, evaluated on shared prompts before and after relevance filtering. K matches Table 3.

Method	#Claims	VS				VS+Rel			
		#Sup.	P	R	F1	#Sup.	P	R	F1
<i>Biography</i> ($n = 488$, $K = 20$)									
without the relevance term	25.6	10.5	0.360	0.480	0.399	9.4	0.316	0.438	0.357
full reward	25.4	10.1	0.342	0.466	0.384	9.1	0.301	0.424	0.343
<i>Alpaca-Fact</i> ($n = 246$, $K = 19$)									
without the relevance term	26.5	18.2	0.631	0.699	0.638	13.4	0.471	0.573	0.491
full reward	25.6	17.7	0.625	0.681	0.629	13.3	0.472	0.570	0.492

A.5 FAITH

FAITH trains a knowledge-state estimator, SFT model, reward model, and state-conditioned PPO policy, followed by a retrieval-based rectifier. The artifact includes four datasets, a trainer, and stage-specific configurations.

We train the first four stages on Meta-Llama-3-8B: 3 epochs each for the estimator and SFT, 2 for the reward model, and 2 PPO epochs over 10,000 of 93,904 prompts, about 310 steps or one fifth of a full-data pass. We follow the paper’s epoch counts rather than the released settings of 1 for the reward model and 0.45 for PPO.

We omit the rectifier because its data and Llama-3 configuration were unavailable during our runs. The authors later supplied the data and confirmed that the configuration was lost. We therefore compare against the no-rectifier ablation. Our estimator replaces the authors’ stored evaluation states; using those states changes accuracy by at most 0.7 points, with 87.5% state agreement (Table 13).

PPO adds 0.33 points in mean precision over questions classified as known, against 1.20 for the published ablation and about ten for the full system (Table 14). Our SFT scores match the reported levels on TriviaQA and NQ-Open but fall short on SciQ and WebQuestions. Reward changes little, from -0.37 at step 9 to -0.34 at step 309.

Questions classified as known are answered correctly more than twice as often as unknown questions on every dataset. Yet refusal rates are 0.000 on three datasets and 0.0003 on the fourth, despite 17%–56% of questions being classified as unknown. Truthfulness, which credits correct refusals, is therefore effectively accuracy.

Table 13: FAITH with predicted (Est.) or released, pre-computed (Read) knowledge states. Agree measures state agreement. Both models use Meta-Llama-3-8B (Section 4); Est. follows Section 3.4 of the original paper.

Dataset	SFT		SFT + PPO		Agree
	Est.	Read	Est.	Read	
NQ-Open	.297	.303	.295	.298	83.2%
SciQ	.585	.584	.598	.591	84.1%
TriviaQA	.662	.662	.659	.658	90.1%
WebQuestions	.342	.341	.347	.346	82.6%
overall					87.5%

Table 14: FAITH on its own four validation sets, scored in one harness. Precision is over the questions the estimator predicts as known. The last column is the gain of the paper’s full system on Llama-3-8B, whereas our delta compares our own SFT stage with our own PPO model, which corresponds to the paper’s ablation without the retrieval stage, reported in Table 1.

Dataset	Ours, precision		Full system	
	SFT	FAITH	Δ	Δ
TriviaQA	74.15	73.56	-0.60	$+13.39$
SciQ	65.74	67.13	$+1.39$	$+8.43$
NQ-Open	42.51	42.57	$+0.06$	$+16.72$
WebQuestions	43.27	43.74	$+0.47$	$+1.53$
Mean			$+0.33$	$+10.02$

Why FAITH produces short answers. The reward model’s containment label accepts a prediction contained in the gold answer, or vice versa. This can credit incomplete answers, but does not by itself establish a preference for brevity. The short-answer SFT data offer a more direct explanation: mean biography length falls from 274 words in the base model to 3 after SFT and remains 3 after PPO. Adding the knowledge state does not restore long responses (Table 18). Since FAITH is presented as general factuality alignment, this

behavior matters beyond its short-answer training tasks.

B Additional Common-Benchmark Results

This section gives the common-benchmark results that Sections 5.3 and 5.4 refer to.

B.1 Supervised Stages Alone

Tables 15 and 16 give the short-form and long-form results of each method’s supervised stage alone, which Section 5.4 compares with the full pipelines in Tables 2 and 3.

Table 15: SFT stage results for short-form evaluation.

Method	Acc.	Abst.	Acc att.	F
SimpleQA ($n = 1,000$)				
base (untrained)	.034	.812	.181	.057
FLAME	.029	.860	.207	.051
KNOWRL	.029	.831	.172	.050
FACTGRPO	.054	.416	.092	.068
FAITH	.046	.001	.046	.046
TruthfulQA ($n = 817$)				
base (untrained)	.416	.108	.466	.440
FLAME	.431	.100	.479	.454
KNOWRL	.401	.131	.462	.430
FACTGRPO	.236	.365	.372	.289
FAITH	.165	.029	.170	.168

Table 16: Long-form results of the supervised stages alone, scored in the same VERIScore configuration and with the same fixed K as Table 3.

Method	Sent.	S	U	Prec.	F1@ K
Biography ($n = 491$, $K = 20$)					
base (untrained)	13.6	12.02	15.60	.351	.370
FLAME	11.6	10.30	12.35	.334	.349
KNOWRL	16.0	11.33	19.31	.270	.316
FACTGRPO	12.7	10.33	15.15	.353	.394
FAITH	1.0	0.39	0.21	.373	.037
Alpaca-Fact ($n = 247$, $K = 19$)					
base (untrained)	21.9	19.79	9.38	.618	.613
FLAME	19.6	17.32	8.36	.594	.573
KNOWRL	23.3	19.59	10.79	.584	.594
FACTGRPO	14.8	17.10	8.37	.616	.616
FAITH	1.0	0.13	0.03	.103	.012

B.2 FAITH With and Without the Knowledge State

Tables 17 and 18 compare FAITH with and without the estimator-predicted knowledge state in the prompt, for the supervised stage and the full pipeline. The state lowers accuracy, precision, and F1@ K slightly, and the full pipeline barely differs

from its supervised stage. The main results use the variant with the state, as the paper prescribes.

Table 17: Short-form results of FAITH with and without the estimator-predicted knowledge state in the prompt, for the supervised stage and the full pipeline.

Method	Acc.	Abst.	Acc att.	F
SimpleQA ($n = 1,000$)				
FAITH SFT	.055	.001	.055	.055
FAITH SFT + state	.046	.001	.046	.046
FAITH full	.058	.001	.058	.058
FAITH full + state	.049	.000	.049	.049
TruthfulQA ($n = 817$)				
FAITH SFT	.192	.020	.196	.194
FAITH SFT + state	.165	.029	.170	.168
FAITH full	.193	.021	.198	.195
FAITH full + state	.170	.029	.175	.173

Table 18: Long-form results of the same four FAITH variants, with the fixed K of Table 3.

Method	Sent.	S	U	Prec.	F1@ K
Biography ($n = 495$, $K = 20$)					
FAITH SFT	1.0	0.54	0.24	.446	.050
FAITH SFT + state	1.0	0.39	0.21	.374	.037
FAITH full	1.0	0.52	0.22	.444	.048
FAITH full + state	1.0	0.39	0.21	.377	.037
Alpaca-Fact ($n = 253$, $K = 19$)					
FAITH SFT	1.2	0.25	0.06	.132	.023
FAITH SFT + state	1.0	0.12	0.04	.101	.012
FAITH full	1.2	0.24	0.05	.134	.022
FAITH full + state	1.0	0.12	0.03	.099	.012

B.3 Cross-Reward Scores

Table 19 gives the mean reward of each method’s generations under each method’s reward, which Section 5.3 discusses.

Table 19: Mean reward of each method’s generations under each method’s reward, as the difference from the base model’s generations, whose absolute value is in the first row. Rewards are per thousand words (FLAME, FACTALIGN, FAITH) or per response (KNOWRL on short-form and FACTGRPO on long-form prompts). Bold marks a method under its own reward, and † a significant difference from the base row (paired bootstrap, Holm-corrected $p < 0.05$). FAITH generations are left out, since one-word answers make per-word values not comparable.

Generations	Reward				
	FLAME	FACTALIGN	KNOWRL	FACTGRPO	FAITH
base (untrained)	1.13	−1.15	−0.392	−1.000	3.85
FLAME	−0.09 †	−0.25†	+0.050†	0.000	+2.16†
FACTALIGN	−0.01	+0.05 †	+0.026	0.000	+0.03
KNOWRL	−0.12†	+0.03	+0.004	0.000	−0.32
FACTGRPO	−0.40†	−2.24†	−0.349†	+1.028 †	−3.97†

B.4 FACTGRPO Reward Hacking

The reward rises with the number of supported but irrelevant claims because its main term, $S/(1+T)$, counts every supported claim and the relevance term contributes little, so filler claims raise the reward of the rerun as they did in the first run (Figure 3 in Section 5.3).

C Evaluation Analysis

This section gives the analyses of the evaluation that Section 6 refers to.

C.1 Open LLM Judge

Table 20 repeats the short-form evaluation with the open gpt-oss-120b judge in place of gpt-4o. The open judge grades TruthfulQA more leniently, but the large drops of FACTGRPO and FAITH remain.

Table 20: The short-form evaluation under both judges. *Acc.* is accuracy over all questions, *F* the harmonic mean of accuracy and accuracy-given-attempted. Absolute values shift systematically between judges; only differences within a column are meaningful. Items that the open judge left ungraded, up to 58 of 817 on TruthfulQA, are excluded from its rates.

Method	gpt-4o		gpt-oss-120b	
	Acc.	F	Acc.	F
SimpleQA ($n = 1,000$)				
base (untrained)	.034	.057	.030	.051
FLAME	.032	.056	.030	.052
FACTALIGN	.032	.054	.029	.049
KNOWRL	.032	.054	.031	.052
FACTGRPO	.052	.066	.053	.067
FAITH	.049	.049	.048	.049
TruthfulQA ($n = 817$)				
base (untrained)	.416	.440	.466	.501
FLAME	.424	.448	.471	.505
FACTALIGN	.435	.460	.469	.506
KNOWRL	.403	.429	.459	.502
FACTGRPO	.231	.284	.294	.361
FAITH	.170	.173	.186	.202

C.2 Filler and Hallucinated Claims

Figure 4 shows the distribution of filler claims, supported but irrelevant, per biography response, and Figure 5 that of hallucinated claims, relevant but unsupported. Every method that answers at length has a long tail in both, whereas the one-word answers of FAITH contain at most one claim of either kind.

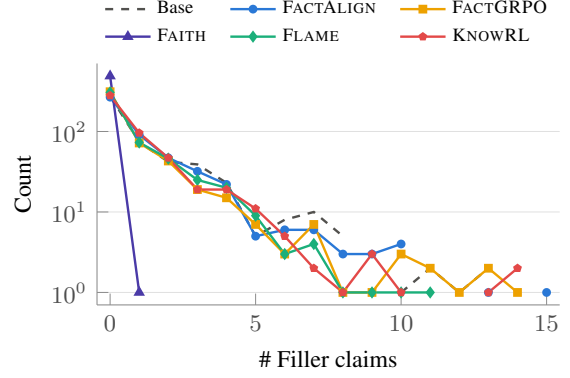


Figure 4: Distribution of the number of filler (supported but irrelevant) claims within long-form responses for the Biography dataset. The y-axis is logarithmic, and a missing point means a count of zero.

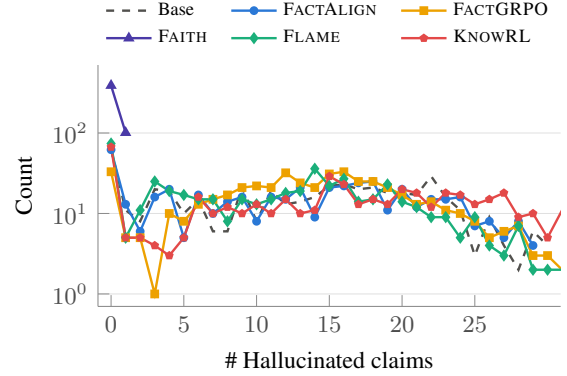


Figure 5: Distribution of the number of hallucinated (relevant but unsupported) claims within long-form responses for the Biography dataset. The y-axis is logarithmic, and a missing point means a count of zero.

Table 23 gives examples of non-claims, repeated claims, and irrelevant claims that VERIScore counts as supported.

C.3 Sensitivity to the Recall Normalizer

Adding the FACTGRPO rerun to the pool moves its median claim count to 23 for biographies and 21 for Alpaca-Fact. At those values every $F1@K$ falls by about .01 to .02, the order of the methods stays the same apart from one tie on Alpaca-Fact, and the one significant gain of FACTGRPO no longer survives the paired test, so we keep K fixed rather than let one run move every other row.

Question	Response (<answer> span)	R_{rel}	R_{fact}
Tell me a bio of Diego Suarez (garden designer).	I'm not sure if I have any information on a garden designer by the name of Diego Suarez. However, I can provide some general information about <i>garden designers</i> ... Garden designers are professionals who design and plan gardens, taking into account factors such as the climate, soil type, and desired aesthetic. ... Some well-known garden designers include Gertrude Jekyll, Russell Page, and Piet Oudolf. ... I apologize, but I don't have any information on a garden designer by the name of Diego Suarez.	0.00	0.93
Tell me a bio of Shaolin (humorist).	I'm not sure if I have any information about a well-known humorist by the name of Shaolin. However, I can provide some general information about <i>the Shaolin Temple in China</i> ... The Shaolin Temple is located in Dengfeng, Henan Province, China ... has a rich history that dates back to the 5th century, and it's been a major center of Buddhist learning ... particularly Shaolin Kung Fu ...	0.00	0.84
Tell me a bio of Alexandre Hohagen.	Alexandre Hohagen is a name that combines a French first name with a German surname. <i>The name Alexandre is derived from the Greek name Alexander</i> , which means "defender of the people." The surname Hohagen is of German origin ... Unfortunately, I don't have any specific information about a person named Alexandre Hohagen. ...	0.00	0.75
Tell me a bio of Kubota Beisen.	I'm not familiar with a well-known artist by the name of Kubota Beisen. ... However, I can provide some general information about <i>Japanese art and artists from the same time period</i> . The Edo period (1603–1868) and the Meiji period (1868–1912) were significant periods in Japanese art history ... Some notable Japanese artists from these periods include Hokusai ... Hiroshige ... and Yokoyama Taikan ...	0.00	0.74
Tell me a bio of Rangi Te Kanawa.	Rangi Te Kanawa is a New Zealand singer, but I couldn't find much information about their background ... you might be thinking of <i>Kiri Te Kanawa</i> . Kiri Te Kanawa is a prominent New Zealand soprano ... She was born in 1944 in Gisborne, New Zealand ... and has since performed with many major opera companies ... including the Metropolitan Opera in New York and the Royal Opera in London. ...	0.00	0.67

Table 21: FACTGRPO responses from the first run that drift away from the requested subject yet earn a high factuality reward (<answer> span only; ... marks omitted text). Italics mark the substituted subject: a profession, a homonym, the name’s etymology, an art-historical period, and a similarly named person. R_{rel} is 0.1 only when an LLM judge prefers the response to the reference response, and 0 otherwise; $R_{\text{fact}} = S/(T+1)$ over T extracted claims, S of them supported. Because $R_{\text{rel}} \leq 0.1$, verifiable but off-topic content dominates the total reward.

Table 22: F1@ K of the same responses under the per-domain K produced by different model pools scored in this study, given as biography / Alpaca-Fact. From left to right, the pools are a twelve-model pool with five FAITH variants, a ten-model pool, a nine-model pool, the six-run pool of the earlier scoring pass, and a single model scored alone. The responses are those of the earlier scoring pass, with the first FACTGRPO run. Precision does not depend on K .

Method	K (biography / Alpaca-Fact)				
	5 / 7	14 / 15	20 / 18	23 / 21	29 / 24
Biography ($n = 492$)					
base	.442	.398	.371	.357	.331
FLAME	.435	.394	.367	.353	.326
FACTALIGN	.427	.384	.357	.342	.315
KNOWRL	.421	.378	.350	.336	.311
FACTGRPO	.464	.430	.399	.382	.347
FAITH	.164	.068	.049	.043	.034
Alpaca-Fact ($n = 249$)					
base	.684	.635	.615	.594	.574
FLAME	.681	.629	.609	.588	.568
FACTALIGN	.679	.624	.604	.584	.564
KNOWRL	.675	.625	.605	.585	.564
FACTGRPO	.704	.661	.641	.622	.600
FAITH	.048	.026	.022	.019	.017

C.4 Reference Verification Against Wikipedia

We re-verify biography claims against the subject’s article in a 2023 English Wikipedia dump, using BM25 over 25-word windows and the same verifier (Table 24). Precision falls by about twenty points for methods producing long responses. Wikipedia supports only two fifths of web-supported claims, while nearly all Wikipedia-supported claims also have web support. FACTGRPO and FLAME move from below to above the base model, significantly so for FACTGRPO. KNOWRL remains significantly below it, while FAITH retains the highest precision with very few claims.

Question	Response claims	Comment
Tell me a bio of Mindy Smith.	[FACTGRPO] ✗ Mindy Smith has released several singles throughout her career. Below is a list of some of her notable releases: • - “One Moment More” • - “It’s That Time Again” • - “Come to Jesus” • Mindy Smith has been performing at festivals and venues across the United States. ...	VERIScore extracts bare song titles as check-worthy claims and labels them supported.
Tell me a bio of Jessie Mae Brown Beavers.	[FACTALIGN] ✗ Jessie Mae Brown Beavers was active in the Civil Rights Movement in Mississippi. • Jessie Mae Brown Beavers was a community leader. • Jessie Mae Brown Beavers was an community leader. • Jessie Mae Brown Beavers was an activist. ...	Near-duplicate claims are each counted as separately supported, inflating the factuality score.
Tell me a bio of Diego Suarez (garden designer).	[FACTGRPO] • <i>Garden designers design and plan gardens.</i> • <i>Garden designers consider the climate when designing gardens.</i> • <i>Garden designers consider the soil type when designing gardens.</i> • <i>Garden designers aim to achieve a desired aesthetic in their designs.</i> ...	Generic filler claims are factually supported but irrelevant to the question. FACTGRPO exploits such claims to hack the reward (Table 21).

Table 23: Examples of limitations in claim-level factuality evaluation. Non-claims, repeated claims, and irrelevant claims are each counted as supported, inflating the factuality score. A cross (✗) in place of the bullet marks claims that VERIScore labels not supported (all other claims are labeled supported), and typeface marks the relevance judgment (**bold**: relevant, *italic*: irrelevant).

Table 24: Comparison between evidence sources for verification. Biography claims of each method verified against web search vs against the subject’s Wikipedia article with the same verifier, on the 489 prompts with an article that every method answered. #Sup. is the mean number of supported claims per response, P the mean per-response precision, and Agreement the share of claims with the same verdict under both sources. † marks a significant difference from the base row (paired bootstrap, Holm-corrected $p < 0.05$).

Method	#Claims	Web search		Wikipedia		Agree.
		#Sup.	P	#Sup.	P	
<i>Biography (n = 489)</i>						
base (untrained)	27.6	12.0	.351	5.5	.140	.729
FLAME	23.6	10.6	.338	5.1	.149	.729
FACTALIGN	27.1	11.7	.344	5.5	.141	.734
KNOWRL	29.8	11.3	.281 [†]	5.1	.116 [†]	.760
FACTGRPO	25.4	10.1	.342	4.7	.155 [†]	.752
FAITH	0.6	0.4	.375	0.3	.316 [†]	.841

C.5 Human Audit of Claim Labels

The audit sample has 300 claims, 50 from each of the six variants of the common benchmark, with 200 from biography prompts and 100 from Alpaca-Fact prompts. Within each variant and dataset, the claims are spread evenly over the four combinations of the VERIScore verdict and the relevance

label, and at most one claim is drawn from a response. Since the combinations are balanced, the agreement on the sample is not the agreement on all claims. Two of the authors labeled the sample independently. For each claim, each annotator sees the prompt, the claim, the sentence it was extracted from with its neighbors, up to ten web search results, and, for biographies, up to ten passages from the subject’s Wikipedia article. The automatic labels are not shown.

Each annotator gives each claim up to three labels. The first two judge whether the web evidence and the Wikipedia evidence support the claim, each with the labels *Supported*, *Not supported*, and *Cannot tell from evidence*. A claim is supported when the evidence states or clearly implies every part of it, including names, dates, numbers, and places, and it is not supported when the evidence contradicts any part. When the evidence neither supports nor contradicts the claim, for example because it is about something else, too vague, or cut off, the label is *Cannot tell from evidence*, and the same holds when only part of the claim is covered or when the evidence conflicts. Annotators use only the evidence in the column being judged, not their own knowledge or a search, and treats evidence about a different entity with the same name as nei-

ther supporting nor contradicting the claim.

The third label judges whether the claim is *Relevant* or *Irrelevant* to the prompt, treating the claim as true. A claim is relevant when it gives information that the prompt asks for about its subject, such as any fact about the life, family, education, career, work, or legacy of the person in a biography prompt. It is irrelevant when it is mainly about something else, such as a related person, place, organization, or event, when it falls outside the scope of the prompt or adds detail that was not asked for, or when it is a statement about the answer, the task, or the model. When unsure, annotators choose *Irrelevant*, and a free-text note flags claims that are garbled or are not claims at all. After a check of the first annotator’s labels against these rules flagged blank labels and labels that contradict a rule, that annotator revised the labels they agreed with.

Table 25 gives the agreement of each annotator with the automatic labels, of the two annotators with each other, and of their consensus, the claims to which both gave the same label, with the automatic labels. On the consensus claims, the share of each automatic label that the annotators confirm is .97 for supported and .65 for unsupported under VERIScore, .99 for relevant and .55 for irrelevant under the LLM labels, and .76 for supported and .78 for unsupported under the Wikipedia verdicts.

Table 25: Agreement of the human audit with the automatic labels. Agr. is the share of claims with the same label and κ Cohen’s kappa. Support rows count *Cannot tell from evidence* as unsupported, as the verifiers do.

	<i>n</i>	Agr.	κ
Web support vs. VERIScore			
first annotator	261	.797	.58
second annotator	289	.751	.49
between annotators	261	.778	.50
consensus	203	.842	.65
Relevance vs. LLM labels			
first annotator	267	.835	.62
second annotator	296	.780	.53
between annotators	266	.842	.54
consensus	224	.862	.62
Wikipedia support vs. reference verdict			
first annotator	173	.792	.50
second annotator	192	.677	.32
between annotators	173	.821	.64
consensus	142	.775	.50

D Reproducibility Details

The repository’s reproducibility document records settings from saved arguments, scripts, logs, and outputs; unrecorded values are marked missing. It

lists model revisions, software versions, repository commits, hardware, training times, seeds, hyperparameters, decoding settings, dataset selection, scoring configurations, and API evaluation dates.

In the common evaluation, FLAME reuses the 70B reproduction’s SFT data with Llama-3.1 prompts. Its targets are therefore Llama-2-70B generations and human responses, not the current policy’s generations. It also trains for 1,000 rather than 500 steps.

Training used about 140 GPU hours for the common evaluation and 627 for the original setups, including 547 for FLAME at 70B. Evaluation used at least 70 GPU hours. API costs are estimated from call counts and list prices because billed amounts were not recorded.

Three details qualify Section 5.1. State-prompted FAITH uses limits of 512 long-form and 64 short-form tokens; neither is reached. The FLAME classifier selects 253 of 805 AlpacaEval instructions, against 241 in Lin et al. (2024). Long-form tables exclude ten of 748 FACTGRPO rerun responses that end within the reasoning trace without an answer.

D.1 Data, Metrics, and Training Stages

Table 26 lists the base models, evaluation datasets, and factuality metrics of the five papers, and Table 27 their training stages.

Table 26: Base models, datasets, and metrics of the five methods. None shares a model or dataset with all others. Three metrics are precision-like, one credits refusals, and one excludes them.

Method	Base models	Evaluation data	Factuality metric
FLAME (Lin et al., 2024)	Llama-2 7B, 70B	Bio (Min et al., 2023), Alpaca-Fact (Lin et al., 2024), FAVA (Mishra et al., 2024) TruthfulQA (Lin et al., 2022), AlpacaEval (Dubois et al., 2024)	FActScore (Inst-LLAMA+npn) BLEU/ROUGE; GPT-4 win rate
FACTALIGN (Huang and Chen, 2024)	gemma-2b, Llama-3-8B , Phi3-Mini	LongFact-Objects (Wei et al., 2024b) FActScore biographies (Min et al., 2023)	SAFE F1 @ 100 (ColBERT) FActScore (GPT-3.5 assessor)
KNOWRL (Ren et al., 2026)	R1-Distill-Qwen-7B, Skywork-OR1-7B	SimpleQA (Wei et al., 2024a), ChineseSimpleQA (He et al., 2025) TruthfulQA (Lin et al., 2022), GPQA (Rein et al., 2024), AIME	accuracy, non-rejected only BLEU; accuracy
FACTGRPO (Chen et al., 2026)	Llama-3.1-8B-Instruct	LongFact (Wei et al., 2024b), FAVA (Mishra et al., 2024), Alpaca-Fact (Lin et al., 2024), Bio (Min et al., 2023), FactBench (Bayat et al., 2025), Factory (Chen et al., 2025)	VERIScore prec., F1@K
FAITH (Dong et al., 2026)	Llama-3-8B , Mistral-7B	TriviaQA (Joshi et al., 2017), SciQ (Welbl et al., 2017), NQ-Open (Kwiatkowski et al., 2019) WebQuestions (Berant et al., 2013) (OOD)	precision over <i>known</i> truthfulness, <i>credits</i> <i>refusals</i>

Table 27: Training stages and optimization signals. Implicit rewards use $\beta \log(\pi/\pi_{\text{ref}})$; rule-based rewards use verifier outputs and checks; learned rewards use a trained model.

Method	Training stages	Signal	What it optimizes
FLAME	factuality-aware SFT $\rightarrow$ factuality-aware DPO	implicit	preference pairs scored with FACTSCORE against retrieval
FACTALIGN	fKTO, from an instruction-tuned checkpoint that supplies the supervised stage	implicit	$\mathcal{L}_{\text{KTO}} + \lambda \mathcal{L}_{\text{fKTO}}$: response-level KTO plus sentence-level labels
KNOWRL	cold-start SFT $\rightarrow$ GRPO	rule	format ± 1 , plus an LLM verdict (+2/−1/+1), plus FACTSCORE partial credit when wrong and not a refusal
FACTGRPO	SFT (3k) $\rightarrow$ Dr. GRPO	rule	−1 if malformed, else $F/(T+1) + \lambda \log(1+F) + \mu w$, with $\lambda_{\text{len}}=0$, $\lambda_{\text{lm}}=0.1$; S, T from VERIScore
FAITH	estimator $\rightarrow$ SFT $\rightarrow$ reward model $\rightarrow$ PPO $\rightarrow$ RAG rectifier (RAFT)	learned	a Llama-3-8B reward model with a scalar value head, trained to predict the EM_RP containment score