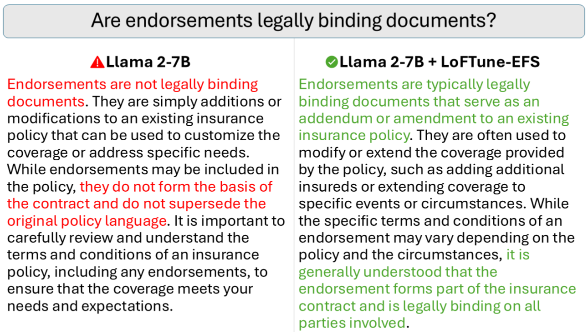
Towards Factual Large Language Models in Low-Resource Domains

Anonymous ACL submission

Abstract

Direct Preference Optimization (DPO) over automatically generated factuality preference rankings has been shown to significantly improve the factuality of large language models (LLMs). However, existing approaches often rely on assumptions, such as access to comprehensive reference or a strong correlation between model confidence and factuality, that do not hold in low-resource domains. To address these limitations, we propose a method for automatically constructing factuality preference datasets from domain-specific resources such as terminologies and knowledge graphs. We introduce two novel factuality estimators: one that links entities from arbitrary domain resources to Wikipedia entries, using their articles as proxy evidence, and another that uses a judge model to estimate factuality in the absence of reliable evidence. We also conduct a systematic study of key factors affecting factuality gains in representative domains, including estimator type, verification set, preference set size, and model scale. Experiments demonstrate significant improvements in in-domain factuality without degrading downstream task performance, while showing evidence of acquired domain knowledge.



Are endorsements legally binding documents?

⚠Llama 2-7B

Endorsements are not legally binding documents. They are simply additions or modifications to an existing insurance policy that can be used to customize the coverage or address specific needs. While endorsements may be included in the policy, they do not form the basis of the contract and do not supersede the original policy language. It is important to carefully review and understand the terms and conditions of an insurance policy, including any endorsements, to ensure that the coverage meets your needs and expectations.

✓Llama 2-7B + LoFTune-EFS

Endorsements are typically legally binding documents that serve as an addendum or amendment to an existing insurance policy. They are often used to modify or extend the coverage provided by the policy, such as adding additional insureds or extending coverage to specific events or circumstances. While the specific terms and conditions of an endorsement may vary depending on the policy and the circumstances, it is generally understood that the endorsement forms part of the insurance contract and is legally binding on all parties involved.

Figure 1: An example illustrating the difference between Llama 2 7B's non-factual output (in red) and factual generation after LoFTune (in green). More examples can be found in appendix A.13.

1 Introduction

Recent research (Tian et al., 2024) has shown to significantly improve the factuality of large language models (LLMs) for long-form generation tasks by fine-tuning on automatically generated factuality preference rankings through Direct Preference Optimization (DPO) (Rafailov et al., 2023). In that work, two approaches were explored involving reference-based and reference-free methods.

Based on FactScore (Min et al., 2023), their reference-based estimator showed good results in an experimental setting tailored to individuals and medical conditions with existing Wikipedia entries for the biography generation and medical question answering tasks. In that setting, the verification set perfectly covered the domain, contributing to demonstrating the effectiveness of the approach. However, it is unclear whether such results generalize to more realistic settings, including low-resource domains, which are typically underrepresented in general knowledge bases like Wikipedia.

In contrast, reference-free estimators using model confidence as an indication of factuality potentially eliminate the need for a verification set. Inspired by (Kuhn et al., 2023), this approach relies on prior findings showing that pre-trained LLMs tend to be well-calibrated (Tian et al., 2023), suggesting that the confidence of a model in a generated answer is highly correlated with the probability that such answer is factual. However, while calibration generalizes to some extent, out-of-distribution limitations (Kadavath et al., 2022) raise concerns about viability in arbitrary domains.

In this work, we target low-resource domains, where neither comprehensive verification sets nor a reliable LLM calibration can be assumed. In doing so, we address the following research questions:

RQ1: Can coverage gaps in general resources like Wikipedia be addressed for reference-based estimation in low-resource domains?

RQ2: Are reference-free estimators based on model confidence reliable in this setting?

RQ3: How do factors like the size of the automatically generated factuality preference set and LLM scale influence factuality outcomes?

RQ4: Do factuality improvements also reflect domain knowledge transfer into the model?

RQ5: Does factuality fine-tuning affect downstream performance?

RQ6: How does domain-specific factuality improvement affect other or general domains?

Our primary contribution is a systematic method for constructing factuality preference datasets in low-resource domains. We extend prior work by leveraging domain-specific structured resources such as terminologies and knowledge graphs to improve LLM domain alignment with the domain via DPO. To address current limitations, we introduce two novel factuality estimators. One extends verification coverage by linking underrepresented domain entities to semantically similar entities with existing Wikipedia articles, which are then used as proxy evidence; the second adopts a reference-free approach, where the factuality of LLM generations is estimated using a judge model.

We validate our approach in a low-resource domain, insurance, using health, a well-resourced domain for comparison. We systematically examine key factors affecting factuality improvements, such as the choice of factuality estimator, verification coverage, preference set size, and model scale. Both our reference-based and judge-based estimators outperform prior methods, with notable gains in in-domain settings, while showing evidence of domain knowledge acquisition by the LLM. The reference-based estimator benefits from increasingly large preference sets, quickly surpassing the judge-based approach, which is less sensitive to the volume of the preference set. Downstream task performance also improves. All datasets and code are publicly available.[1]

2 Related Work

Despite their capabilities, LLMs still struggle with hallucination (Ji et al., 2023; Rawte et al., 2023; Kandpal et al., 2023; Mallen et al., 2023; Lee et al., 2022), emphasizing the need for evaluation frameworks, such as reference-based fact-checking for long-form generation (Min et al., 2023; Chern et al., 2024; Wei et al., 2024), methods that leverage their internal knowledge to estimate factuality (Zhang et al., 2024; Manakul et al., 2023), and open-ended generation benchmarks (Vu et al., 2024; Muhlgay et al., 2024; Yin et al., 2023; Lin et al., 2022).

Training-free methods to improve factuality include external augmentation (Si et al., 2023; Jiang et al., 2023; Shuster et al., 2021), specialized decoding (Li et al., 2023b; Chuang et al., 2024), and chain-of-verification (Dhuliawala et al., 2024). In contrast, recent studies apply reinforcement learning to align LLMs with a factuality objective. FactAlign (Huang and Chen, 2024) introduces fKTO, a sentence-level algorithm extending Kahneman-Tversky Optimization (Ethayarajh et al., 2024). FLAME (Lin et al., 2024) enhances factuality while preserving instruction-following via supervised fine-tuning (SFT) and DPO. Kang et al. (2024) reduce hallucinations by generating succinct responses to unfamiliar queries.

FactTune (Tian et al., 2024) applies DPO using factuality preference pairs derived from FactScore and LLM confidence, but does not explore generalization to diverse or unseen domains. In contrast, motivated by recent findings linking hallucinations to low-resource settings (Luo et al., 2024; Kandpal et al., 2023; Guerreiro et al., 2023), we target low resource domains, addressing limitations in verification sources and LLM calibration in these settings, study factors like preference set size and LLM scale, and examine how factuality gains impact on downstream performance.

Factuality is crucial in critical domains such as finance, healthcare, and law, yet challenges persist (Chen et al., 2024). In law, grounding question answering in statutory provisions (El Hamdani et al., 2024; Louis et al., 2024) improves reliability, but hallucinations are common in responses and rationales. Clinical settings face similar issues, though progress in medical information extraction (Xu et al., 2024) and diagnosis (McDuff et al., 2025) is promising, with benchmarks like Med-HALT (Pal et al., 2023). In contrast, insurance remains under-resourced, with hallucination hindering LLM adoption (Balona, 2024). To our knowledge, our work is the first to systematically evaluate and improve factuality in this domain.

3 Preliminaries

Our approach for fine-tuning LLMs for factuality in specialized domains builds on the framework pro-

[1] https://github.com/anonloftune/LoFTune

posed by Tian et al. (2024), which leverages direct preference optimization (Rafailov et al., 2023) for preference-based reinforcement learning. In this section, we provide an overview of both methods.

3.1 Direct Preference Optimization

Let $\mathcal{P} = \{x, y_w, y_l\}$ be a preference dataset, where x is a prompt to an LLM, y_w its preferred (in our case, more factual) completion and y_l a less desirable option. According to (Bradley and Terry, 1952), the probability that y_w is preferred to y_l is as follows, with σ the logistic function and r an unobserved reward function:

$$p(y_w \succ y_l) = \sigma(r_{y_w} - r_{y_l}) \tag{1}$$

The goal of reinforcement learning is to maximize the expected reward for our prompts, usually combining that reward with a KL-divergence penalty (Kullback and Leibler, 1951) between the policy model π_θ and its initialization π_{ref}, with a hyperparameter β that controls the strength of the constraint. Unlike previous approaches such as PPO (Schulman et al., 2017), DPO (Rafailov et al., 2023) enables learning π_θ from $\mathcal{P}$ directly through supervised learning (equation 2), without fitting an explicit reward function or sampling from the policy during training. However, the challenge remains in the construction of a dataset of preference pairs that encourage greater factuality.

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(y_w, y_l, x)_{\sim\mathcal{D}}}\left[log\sigma(r_{y_w} - r_{y_l})\right]$$

where: (2)

$$r_{y_w} = \beta\, log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}$$

(3)

$$r_{y_l} = \beta\, log\, \frac{\pi_{ref}(y_w|x)}{\pi_{ref}(y_l|x)}$$

3.2 Factuality Tuning

Given a truthfulness estimator, FactTune (Tian et al., 2024) builds a factuality preference dataset $\mathcal{P}$ from a set of n unlabeled questions about Wikipedia entities regarding individual and medical conditions sampled from an LLM such as GPT-3.5. For each question, m candidate long-form responses are sampled from the target model that we aim to fine-tune for factuality, using temperature 1.0. For models without instruction tuning, few-shot prompting is used.

Each of the m responses is split into atomic claims (see Appendix A.5) and scored for truthfulness using either a reference-based or reference-free estimator. For the reference-based method, each atomic claim is checked against Wikipedia using FactScore by a smaller, more efficient model such as LLaMA-1. For reference-free estimation, each atomic claim is reformulated as a minimally ambiguous question using the larger LLM. For each atomic question, the target model is resampled 20 times, with the truthfulness score of each atomic claim being the frequency of the most common answer, reflecting model confidence. For either estimator, scores are aggregated for each of the n questions to compute the overall truthfulness score.

For each $\binom{m}{2}$ response pairs per question, the response with a higher score is chosen as the preferred one. For n total questions, $n\binom{m}{2} - k$ preference pairs are generated, with k the number of tied pairs. Finally, the target LLM is fine-tuned with DPO on the resulting $\mathcal{P}$, with all m responses as SFT targets.

4 Proposed method: LoFTune

Our method LoFTune, standing for Low-resource Factuality Tune, builds on the general framework for factuality tuning proposed by (Tian et al., 2024). However, we focus on low-resource domains, which we define as those with limited coverage in general resources such as Wikipedia, a lack of comprehensive domain-specific corpora like PubMed[2], and low representation in LLM pretraining data. These limitations challenge both reference-based and model confidence-based factuality estimators, reducing their effectiveness.

Unlike Tian et al. (2024), who focus solely on entities that are covered in Wikipedia, we leverage entities drawn from domain-specific (semi)structured resources. In doing so, we seek to enable more accurate in-domain factuality evaluation and better adaptation of the model to the domain. Let $R = (E, L)$ be a domain-specific resource, where E is a set of domain entities, and L is a set of links or relations among them. Given a factuality estimator, for each entity $e \in E$ we sample a set of questions Q_e from an LLM[3] and apply the FactTune pipeline. Note that the semantic structure of R can vary, from flat glossaries of terms, where $L = \emptyset$, to knowledge graphs where facts are rep-

[2] https://pubmed.ncbi.nlm.nih.gov

[3] For comparison with (Tian et al., 2024), we use GPT-3.5.

resented as a set of triples $F \subset E \times L \times E$. The approach naturally accommodates to the latter, including questions about each fact $(e_i, l_k, e_j) \in F$.

Additionally, we note that in low-resource domains the truthfulness score of the preferred responses across different factuality estimators tends to be lower compared to the scenarios considered in (Tian et al., 2024). To optimize the signal-noise ratio of the factuality preferences in $\mathcal{P}$, we filter out those pairs where the preferred response has a truthfulness score below a given threshold, which we fix empirically.

Next, we present two new factuality estimators introduced in this paper.

4.1 Expanded reference-based: LoFTune-EFS

Focusing on specialized resources provides a more detailed view of a domain. However, while large reference corpora like PubMed exist for fields such as health, such resources are lacking in others, like insurance. In such cases, general-purpose resources like Wikipedia may offer limited or mismatched coverage, reducing their effectiveness as verification datasets with FactScore. For example, the term *Excess Liability Insurance* from the LGIT glossary does not appear explicitly in Wikipedia. However, Wikipedia contains a semantically similar entity, *Excess Insurance*, with a corresponding article that can serve as proxy reference for verification.

Our method links the domain entities that are not explicitly represented in Wikipedia to similar entities with existing Wikipedia articles, providing FactScore with reference text to verify claims that may involve any entity $e \in E$. Given a domain-specific resource $R = (E, L)$, let $\mathcal{V}_R = \{w_e\}$ be a verification set where w_e is the Wikipedia article associated with entity e, as shown in equation 4.

$$w_e = \begin{cases} w_e & \text{if } e \in E_{wiki}, \text{ and otherwise:} \\ w_x & \arg\max_x \, wiki\_search(e, x) \end{cases} \tag{4}$$

We define $wiki\_search$ as a search function[4] over Wikipedia that retrieves candidate Wikipedia entities x based on their semantic similarity with e in the context of the domain of interest. Our goal is not to identify an exact match for e in the set of Wikipedia entities E_{wiki}, but to spot Wikipedia entities with articles that may contain relevant information for fact checking LLM generations about e. We therefore adopt a relaxed notion of semantic similarity and instruct the LLM powering $wiki\_search$ (appendix A.12) to retrieve the semantically closest entity to e.

4.2 Judge-based estimator: LoFTune-J

Recent studies (Kim et al., 2024; Zheng et al., 2023) report that large LLMs such as GPT-4, Claude 3, and Gemini 2.5 exhibit evaluation capabilities comparable to those of humans. Based on those findings, rather than resorting to a reference knowledge base or relying on model confidence, which can be troublesome for low-resource domains, the judge-based estimator leverages the evaluation capabilities of large LLMs to estimate factuality. Given a passage generated by L_t regarding a domain entity $e \in E$, we instruct a large LLM, in this case GPT-4o, to return a factuality score between 0 and 1, where 0 means absolutely false and 1 completely truthful (see the prompt in appendix A.10).

5 Experimentation

We address the research questions outlined in the introduction, with (Tian et al., 2024) as our baseline. We focus on insurance as a representative low-resource domain and use health for comparison in a well resourced setting. According to the domain taxonomy in (Wettig et al., 2025), health accounts for 6.5% of a pre-training set from Common Crawl, while insurance is estimated at just 0.1%. In a corpus like Dolma (Soldaini et al., 2024) this corresponds to ∼75B and ∼3B tokens, respectively. Additionally, just in the english Wikipedia we find 232K articles related to health, with only 9K for insurance. These estimates reflect typical LLM support for each domain. In terms of reference corpora, to our knowledge insurance is lacking domain-specific resources that suit reference-based estimation. In contrast, PubMed alone holds over 36M papers, offering extensive support for health.

5.1 Materials and resources

We use Llama 2-7B as our target model, following the setup in (Tian et al., 2024). For the analysis of the impact of model scale, we use the Pythia model suite. As judge model for factuality estimation, we use GPT-4o (*gpt-4o-2024-08-06*). Our expanded reference-based estimator and baselines use GPT-4o mini to fact-check atomic claims.

To generate SFT and DPO datasets with different estimators in insurance, we use the State of

[4] Built on OpenAI Web Search and GPT-4o mini `https://platform.openai.com/docs/guides/tools-web-search`

Maryland's LGIT glossary[5], with 113 entities that we split into train (91), validation (11), and test (11) sets. While domain-comprehensive, LGIT reflects a localized perspective on the domain, allowing us to test our approach in specialized settings. Alternative glossaries include, e.g., NAIC[6]. For verification, we use the FactScore Wikipedia dump (Min et al., 2023) for both LoFTune-EFS and the reference-based FactTune-FS baseline.

In health, we focus on COVID-19, a biomedical topic with abundant literature. We curate 4.7 million PubMed abstracts (2000–2022), selecting those with at least three highly influential citations according to Semantic Scholar (Valenzuela et al., 2015), which form our verification set. We then select abstracts mentioning *COVID*, *Coronavirus*, or *CoV-SARS-2*, and extract UMLS (Bodenreider, 2004) entities via named-entity-recognition following (Wright et al., 2022). The most frequent entities covering 50% of the distribution yield a glossary of 295 terms that are randomly split into training (237), validation (29), and test (29) sets. Preference set generation mirrors the insurance procedure, with prompts adapted to the domain.

For the train and validation splits, we generate $n = 6$ questions per entity using GPT-3.5 and the prompts in appendices A.1 and A.2. Using the prompts in A.3 and A.4, for each question we sample different numbers of responses $m \in \{5, 10, 20, 30, 40\}$ from the target model, resulting in different sizes of the factuality preference dataset $\mathcal{P}$. We then calculate the factuality score of each response set with the different estimators. For model confidence estimation, we use the prompts in A.6 and A.7 to transform atomic claims into questions and those in A.8 and A.9 to sample answers to such questions from the target model. Finally, we generate $\mathcal{P}$ for each estimator. For LoFTune, we fix a FactScore threshold $t = 50$ on the preferred option of each pair.

5.2 Factuality estimator analysis

This section addresses research questions RQ1, part of RQ2 and RQ3, and RQ4. We evaluate on ablations of $\mathcal{P}$ for $m \in \{5, 10, 20, 30, 40\}$ across several LoFTune variants based on the factuality estimators from section 4: Expanded reference-based LoFTune-EFS, with truthfulness threshold $t = 50$, and judge-based LoFTune-J. For comparison, we include the FactTune-MC model-confidence estimator by Tian et al. (2024) and FactTune-FS, standard reference-based without expansion or threshold. Each variant is compared to its SFT, trained on the same ablation of $\mathcal{P}$, and the base LLM.

We train both SFT and DPO models with LoRA (Hu et al., 2022), using rank $r = 8$, and $\alpha = 16$. For DPO training, we use $\beta = 0.1$, with batch size 64 and learning rate 1e-5, linearly warmed up from 0 to 1e-5 over the first 150 steps, followed by cosine decay. Training runs for up to 20 epochs, with evaluations performed every half epoch. We apply early stopping with patience of 4 evaluations. We report FactScore and the average number of correct, incorrect, and unverifiable (not enough information-NEI) atomic facts per response to account for limitations in the verification set.

As shown in Table 1, models trained with DPO on preference sets $\mathcal{P}$ built via LoFTune-EFS consistently outperform prior methods across all values of m and dataset sizes, driven by a more favorable ratio of correct to incorrect facts. In relation to RQ4, we observe a growing average number of correct facts in model outputs, with LoFTune-EFS generating nearly one more correct fact per response than the base model at $m = 40$, while reducing the number of incorrect facts. This suggests improved recall and a more effective use of domain knowledge from parametric memory.

LoFTune-EFS shows near-linear factuality gains as m increases, while LoFTune-J remains relatively insensitive to changes in m, being the only method to outperform the base LLM from the start, at $m = 5$, by a large margin. Indeed, FactTune-FS only outperforms LoFTune-J at $m > 20$. Although LoFTune-J shows slight improvements in the number of correct and incorrect facts at higher values of m, these gains do not consistently result in a higher FactScore due to a parallel increase in NEI.

LoFTune-EFS outperforms FactTune-FS across the board, with a gap between the two estimators that is near constant for $m > 5$. The expansion mechanism enables the verification of entities that would otherwise remain unaccounted for. Combined with the application of threshold t, this leads to larger and higher quality factuality preference datasets $\mathcal{P}$ for all values of m, which contribute to a more effective DPO training. We also observe particularly large $|\mathcal{P}|$ for FactTune-MC, which tends to generate overconfident estimates. However, FactTune-MC consistently underperforms across all values of m, always below the base model, sug-

[5] https://www.lgit.org/611/Glossary-of-Insurance-Terminology

[6] https://content.naic.org

gesting limitations in low-resource domains such as insurance. All models surpass the SFT baseline for $m > 5$.

m	Method	$\|\mathcal{P}\|$	Factuality				InsQA
			FactScore	#correct	#incorr↓	#NEI↓	
	llama-2-7b-hf	-	59.37	3.38	0.15	2.17	60.45
5	SFT	-	57.36	3.36	0.24	2.48	67.25
	FactTune-MC	5195	56.83	3.25	0.23	2.35	66.95
	FactTune-FS	1989	57.75	3.34	0.18	2.33	67.31
	LoFTune-EFS	2497	61.10	3.49	0.18	2.00	68.00
	LoFTune-J	3166	64.36	3.54	0.13	1.80	68.16
10	SFT	-	54.94	3.21	0.21	2.49	66.97
	FactTune-MC	23371	57.64	3.39	0.24	2.32	67,32
	FactTune-FS	8949	56.73	3.43	0.24	2.46	67.61
	LoFTune-EFS	11213	61.24	3.57	0.25	2.06	68.20
	LoFTune-J	14226	64.05	3.66	0.08	1.90	68.40
20	SFT	-	57.64	3.43	0.20	2.44	66.91
	FactTune-MC	98298	57.93	3.53	0.21	2.33	67.53
	FactTune-FS	37379	61.73	3.54	0.18	2.07	68.38
	LoFTune-EFS	47583	66.36	3.85	0.15	1.76	68.42
	LoFTune-J	59995	62.06	3.74	**0.08**	2.16	68.39
30	SFT	-	57.52	3.41	0.20	2.59	67.18
	FactTune-MC	225496	59.27	3.50	0.20	2.31	67.42
	FactTune-FS	85908	65.51	3.77	0.13	1.81	68.68
	LoFTune-EFS	109245	68.82	4.04	0.16	1.63	**69.07**
	LoFTune-J	136462	63.82	3.87	0.08	2.06	68.17
40	SFT	-	56.57	3.37	0.19	2.66	67.08
	FactTune-MC	404052	59.05	3.43	0.14	2.36	67.66
	FactTune-FS	154631	67.32	3.95	0.13	1.70	68.79
	LoFTune-EFS	196407	**70.82**	**4.24**	0.12	**1.55**	68.26
	LoFTune-J	244693	63.75	3.99	0.09	2.15	68.34

Table 1: Factuality results for LoFTune variants vs. base, SFT, FactTune. **Overall** best; best per m. Rightmost: InsuranceQA ground truth similarity via GPT-4o mini.

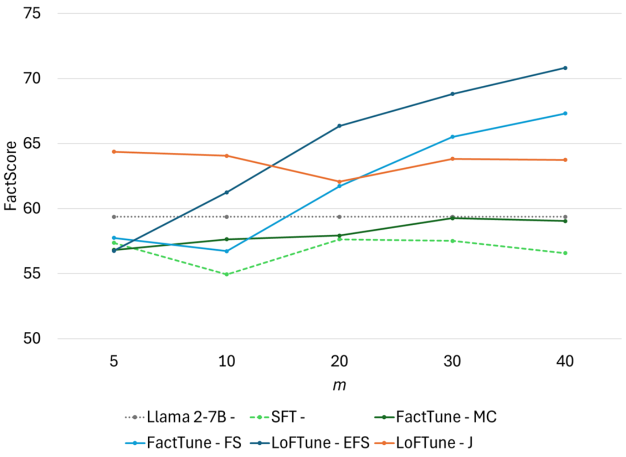


Figure 2: FactScore obtained by SFT, FactTune, and LoFTune models across different values of m.

(a) Pearson correlation

	MC	EFS	Judge
MC	1.00	-	0.024
EFS	-	1.00	0.16
Judge	0.024	0.16	1.00

(b) Spearman correlation

	MC	EFS	Judge
MC	1.00	-	0.036
EFS	-	1.00	0.146
Judge	0.036	0.146	1.00

Table 2: Pearson correlation (a), with p-value 0.073 for MC-Judge and 7.92e-30 for EFS-Judge. Spearman correlation (b): p-values 0.007 and 5.02e-25, respectively.

5.3 Model confidence reliability

The model confidence (MC) estimator proposed by Tian et al. (2024) builds on prior findings that pretrained LLMs tend to be well-calibrated (Kadavath et al., 2022), suggesting that a model's confidence in a generated response can serve as a proxy for its factual accuracy (Tian et al., 2024). However, LLM calibration can degrade across domains (Kadavath et al., 2022). The results obtained in section 5.2 raise our concern about the reliability of factuality estimates based on model confidence in a low-resource setting (RQ2). To examine this more closely, we compute the correlation between MC scores and judge-based factuality ratings, using the latter as ground truth, and compare with our expanded reference-based estimator (EFS). All estimators are evaluated with a fixed $m = 10$ responses per question $q_e \in Q_e, \forall e \in E$.

As shown in Table 2, both EFS and MC estimators moderately correlate with the judge. However, MC scores exhibit a significantly weaker correlation, approximately an order of magnitude lower in both Pearson and Spearman metrics, suggesting a low familiarity of the target model with the insurance domain, which is likely due to limited exposure to in-domain pre-training data. This result echoes the out-of-distribution and calibration challenges highlighted by Kadavath et al. (2022), underscoring the limitations of model confidence as factuality signal in low-resource domains at least at the scale of our base model, 7B parameters.

5.4 Model scale

In this section, we focus on RQ3 by investigating the impact of model scale in LoFTune. To this end, we use the Pythia suite (Biderman et al., 2023), which includes eight models with sizes ranging from 70M to 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, and 12B parameters. We reuse the SFT and DPO datasets corresponding to the best-performing configurations of LoFTune-EFS and LoFTune-J from Table 1, i.e., with $m = 40$ and $m = 5$, respectively, and apply the DPO procedure across all sizes.

To further examine the viability of model confidence as a factuality estimator in low-resource domains (RQ2), we also analyze how the limitations identified in earlier sections evolve with increasing model capacity, under the hypothesis that

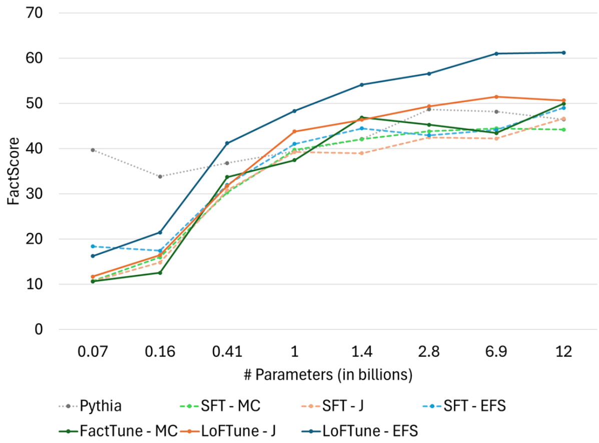


Figure 3: FactScore on Pythia across models sizes. We compare with three estimators: model confidence ($m = 30$), EFS ($m = 40$), and judge-based ($m = 5$).

larger models may mitigate these issues. To this purpose, we add FacTune-MC, generating new SFT and DPO datasets for each Pythia model size using a fixed $m = 30$, which yielded the best performance for FactTune-MC in Table 1. We then compute confidence-based factuality scores for each corresponding model. To train the SFT and DPO models we use the hyperparameters in section 5.2.

Figure 3 shows our results on the Pythia suite. While the 6.9B Pythia model underperforms LLaMA-2 (Section 5.2), our focus is on trends across Pythia scales. All models and their SFT baselines show rising factuality with scale, though SFT gains plateau around 1.4B. In contrast, LoFTune and FactTune models continue improving. LoFTune-EFS consistently outperforms all other methods but levels off at 12B, while LoFTune-J improves steadily up to 6.9B before plateauing. FactTune-MC shows a less stable upward trend, but eventually matches LoFTune-J at 12B, confirming the positive impact of model scale to correlate prediction certainty with factuality. Indeed, the last step between 6.9B and 12B suggests that model confidence may continuing improving with larger parameter counts.

5.5 Downstream performance

While previous sections examined the impact of LoFTune on model factuality, here we evaluate its effect on overall model performance through a domain-specific downstream task, assessing whether improvements in factuality translate into task-specific gains (RQ5). We use the 2,000 questions in the test set of InsuranceQA (Feng et al., 2015), a benchmark for non-factoid question answering in insurance with real-world user questions paired with answers from experts. All models are evaluated in a zero-shot setting, except for the base model, which is not instruction-tuned, for which we use few-shot. Evaluation is conducted on the full dataset, with similarity to ground truth answers measured using GPT-4o mini. [7]

As shown in the right-most column of Table 1, all models outperform the base LLM, with LoFTune-EFS achieving the highest similarity at $m = 30$. SFT models consistently surpass the base model across all m, showing alignment with the InsuranceQA task. However, SFT is also outperformed by all other methods, showing the effectiveness of our approach. Consistent with the factuality trends in Section 5.2, the next best performers are FactTune-FS at $m = 30$ and LoFTune-J at $m = 10$, followed by FactTune-MC at $m = 40$ and the SFT model at $m = 5$. Also, unlike for factuality, the results of FactTune-MC are significantly better than the base model. Combined, our results indicate that LoFTune (and FactTune) not only does not degrade downstream performance but can also enhance it.

5.6 Cross-domain analysis

This section analyzes LoFTune's cross-domain effects (RQ6) in the insurance and health domains. We use health, specifically the COVID-19 subdomain, which has abundant verification data, as a richer-resource comparison to insurance. To ensure consistency, we apply the same estimators and training procedure, fixing $m = 5$ for all experiments. As in insurance, DPO fine-tuning on a LoFTune-EFS-generated preference set yields the highest FactScore, outperforming models trained with alternative estimators (see Table 3). Based on the results shown in Table 1, increasing m would likely further widen this margin in health.

To further evaluate model performance on a relevant downstream task, we use the COVID-QA dataset[8], which consists of open-ended COVID-19 questions and answers. We focus on a subset of biomedical questions from 15 English news websites with technical and domain-specific content. As shown in the right-most column of Table 3, all fine-tuned models outperform the base LLM and SFT. FactTune-FS achieves the highest overall score. However, downstream performance does not fully align with the factuality scores: LoFTune-J

[7] All GPT-4o mini mentions refer to version 2024-07-18.

[8] https://github.com/xhluca/covid-qa

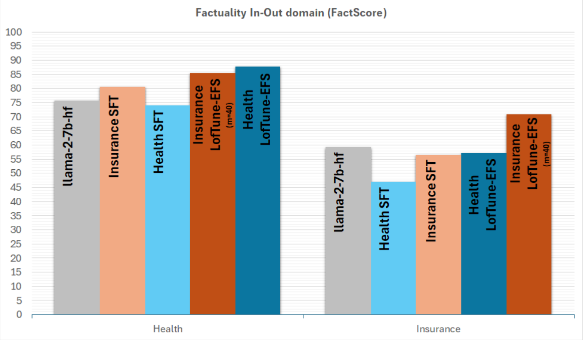


Figure 4: FactScore of insurance LoFTune-EFS model ($m = 40$) in health and vice-versa.

outperforms LoFTune-EFS on the downstream task despite the latter achieving a higher FactScore, suggesting a bias in FactScore toward certain types of responses. Further investigation with varying m values would be needed to verify this hypothesis.

We assess generalization to unseen domains in Figure 4. In insurance, only models trained on insurance data improve over the base model. In health, while LoFTune-EFS trained on health data achieves the highest factuality, models trained on insurance data, and even the SFT model, also outperform the base and SFT models. LoFTuned models remain competitive across domains, often surpassing baselines outside their training domain.

		Factuality				
Method	$\|\mathcal{P}\|$	FactScore	#correct	#incorr.↓	#NEI↓	CovQA
llama-2-7b-hf	-	75.65	3.55	0.22	0.86	39.96
SFT	-	74.05	3.74	0.25	1.07	40.88
FactTune-MC	5194	73.30	2.43	0.27	0.93	41.74
FactTune-FS	11838	86.38	4.43	0.11	0.58	**46.31**
LoFTune-J	10237	85.96	4.29	0.09	0.63	45.14
LoFTune-EFS	10621	**87.83**	4.11	0.12	0.47	44.73

Table 3: LoFTune factuality in health vs. base, SFT, FactTune. Right: COVID-QA similarity (GPT-4o mini).

5.7 General domain analysis

We analyze whether improving LLM factuality within a specific domain affects their overall factuality (RQ6). We select open-ended generation datasets from (Wang et al., 2024) that directly evaluate factual accuracy, excluding those focused on evaluating factuality methods or detecting hallucinations (Wang et al., 2023; Chen et al., 2023; Li et al., 2023a). Our focus is on datasets assessing the factuality of long-form generations, including Factscore-Bio (Min et al., 2023), Factool-QA (Chern et al., 2024), FreshQA (Vu et al., 2024), and SelfAware (Yin et al., 2023). Table 4 reports results on these datasets.

Across all general-domain datasets, except LoFTune-J in FreshQA, LoFTune methods outperform SFT. In SelfAware, where F1 measures self-knowledge, i.e., the ability of the model to identify unknowns, LoFTuned models consistently outperform SFT, with the largest gain obtained by LoFTune-J in insurance. In FreshQA, results are mixed: LoFTune-EFS shows modest improvements across both domains, while LoFTune-J declines, particularly in insurance. In Factool-QA, LoFTuned models, especially in health, achieve notable gains in claim-level accuracy. Finally, in FactScore-Bio, LoFTuned models consistently outperform SFT, with LoFTune-EFS leading in insurance and LoFTune-J in health.

	Model	SelfAware F1	FreshQA Acc.	FacTool-QA Acc.	Bio FactScore
Insur.	$\text{SFT}_{m=40}$	35.13	21.20	56.97	40.94
	LoFTune-EFS	36.02	21.40	**59.21**	**48.73**
	$\text{SFT}_{m=5}$	37.74	22.00	52.40	41.93
	LoFTune-J	**45.58**	18.40	58.71	46.83
Health	$\text{SFT}_{m=5}$	22.30	21.20	35.64	34.76
	LoFTune-EFS	24.21	**22.20**	51.67	38.82
	LoFTune-J	24.25	21.00	55.94	40.40

Table 4: Performance of insurance and health LoFTuned models in general domain factuality datasets.

6 Conclusion

While previous work improved LLM factuality by automatically generating factuality preference datasets and fine-tuning with DPO, such methods mostly focused on general-purpose settings. This paper explores their limitations in more realistic, domain-specific and low-resource contexts, where issues like incomplete references and calibration challenges can affect effectiveness. We introduce LoFTune, a method for automatically generating factuality preference sets from domain-specific resources along with two novel estimators. Our results show additional factuality gains compared to previous approaches. We also observe that larger factuality preference sets improve factuality, while gains from larger models level off at relatively small sizes, except for model confidence, whose trend suggests potential further gains with larger model sizes. LoFTuned models recall more domain-relevant facts and improve downstream performance, although not always does the most factual model yield the best results in those tasks. Finally, we find that models LoFTuned in one domain can generalize to others as well as to general-domain benchmarks.

Limitations

Our data and evaluation pipelines use datasets distilled by LLMs from underlying knowledge artifacts. While prior work shows strong alignment with human judgments, some generated questions may still be irrelevant or incorrect. Also, during our experimentation, our pipeline bootstrapped from entities contained in domain-specific resources. Future work could expand this to also include relations and facts from sources like knowledge graphs.

Our experiments are extensive and the domains we chose are representative of the different scenarios that can be encountered in terms of resource availability. However, GPU availability constraints prevented experimenting in a broader range of domains and with a model scale beyond the 12B parameters of the largest Pythia model. Experimentation with larger models could provide additional clarity. Also due to such limitations, all our training was conducted using LoRA. While effective and reliable, full fine-tuning could offer additional insights.

Although LoFTune has shown to significantly improve factuality in low-resource domains, nonfactual outputs can still occur. While this paper focused on low-resourced domains, future work aims at addressing extensions of LoFTune focused on improving LLMs for underrepresented languages.

References

Caesar Balona. 2024. Actuarygpt: applications of large language models to insurance and actuarial work. *British Actuarial Journal*, 29:e15.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: benchmarking factuality evaluation of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Zhiyu Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Ruth Petzold, and William Yang Wang. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law. *Transactions on Machine Learning Research*. Survey Certification.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2024. Factool: Factuality detection in generative AI - a tool augmented framework for multi-task and multi-domain scenarios.

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*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Rajaa El Hamdani, Thomas Bonald, Fragkiskos D. Malliaros, Nils Holzenberger, and Fabian Suchanek. 2024. The factuality of large language models in the legal domain. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 3741–3746, New York, NY, USA. Association for Computing Machinery.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 813–820.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

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 *International Conference on Learning Representations*.

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*, pages 16363–16375, Miami, Florida, USA. Association for Computational Linguistics.

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 Comput. Surv.*, 55(12).

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. 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, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. Cite arxiv:2207.05221Comment: 23+17 pages; refs added, typos fixed.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. In *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Solomon Kullback and R. A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2024. Zero-resource hallucination prevention for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3586–3602, Miami, Florida, USA. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

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.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale Webster, and Vivek Natarajan. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

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.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian's, Malta. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

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.

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

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.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. 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*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher 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*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI Workshops*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2023.

Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv preprint arXiv:2311.09000*.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 80756–80827. Curran Associates, Inc.

Alexander Wettig, Kyle Lo, Sewon Min, Hanna Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pretraining data curation. *ArXiv*, abs/2502.10341.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024. Mitigating hallucinations of large language models in medical information extraction via contrastive decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7744–7757, Miami, Florida, USA. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don‘t know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. 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)*, pages 1946–1965, Bangkok, Thailand. Association for Computational Linguistics.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Appendix

A.1 Generate questions dataset prompt (Insurance domain)

```
Write 6 different questions about "{ENTITY}" in the Insurance domain that have to be answered with a short paragraph containing factual knowledge. Write only the questions (without the answer) following the format "-- question_1\n-- question_2\n-- question_n"
```

A.2 Generate questions dataset prompt (Health domain)

```
Write 6 different questions about "{ENTITY}" in the covid-19 domain that have to be answered with a short paragraph containing factual knowledge. Write only the questions (without the answer) following the format "-- question_1\n-- question_2\n-- question_n"
```

A.3 Sample model prompt (Insurance domain)

```
INPUT: Answer the following question in a single paragraph: "What is liability insurance?"
OUTPUT: Liability insurance is a type of insurance coverage that provides financial protection to individuals or businesses in the event they are held legally responsible for causing harm or damage to others. It typically covers the costs associated with legal defense, as well as any judgments or settlements that may arise from lawsuits. Liability insurance is crucial for protecting assets and mitigating financial risks, as it helps policyholders manage the financial consequences of third-party claims, such as bodily injury or property damage, arising from their actions or negligence. This type of insurance is common in various fields, including auto insurance, business insurance, and homeowner's insurance, offering a safeguard against the potential financial burdens resulting from legal liabilities.
----
INPUT: Answer the following question in a single paragraph: "What types of liabilities are typically covered by insurance policies?"
OUTPUT: Insurance policies generally cover a range of liabilities, including but not limited to property damage, bodily injury, and personal liability. Property and casualty insurance often protect against financial losses resulting from damage to or loss of physical assets, such as homes, cars, or businesses. Liability insurance, on the other hand, typically addresses legal responsibilities arising from bodily injury or property damage caused by the policyholder or their possessions. This can include coverage for medical expenses, legal fees, and settlement costs. Additionally, professional liability insurance may protect individuals or businesses against claims related to errors, negligence, or malpractice in their professional services. The specific types of liabilities covered depend on the type of insurance policy and its terms and conditions.
----
INPUT: Answer the following question in a single paragraph: "How is liability coverage different from other types of insurance?"
OUTPUT: Liability coverage distinguishes itself from other types of insurance by primarily focusing on providing financial protection for the policyholder against claims or lawsuits filed by third parties due to injuries or property damage for which the insured is deemed responsible. Unlike property or health insurance, which primarily safeguard the policyholder's personal belongings or well-being, liability coverage steps in to cover legal expenses, court judgments, and settlements arising from incidents where the insured is found at fault. This type of insurance is crucial in shielding individuals or businesses from the potentially significant financial repercussions of legal actions, emphasizing protection against liabilities rather than the direct safeguarding of personal assets or well-being.
----
INPUT: Answer the following question in a single paragraph: "{question}"
OUTPUT:
```

A.4 Sample model prompt (Health domain)

```
INPUT: Answer the following question in a single paragraph: "How does COVID-19 spread between people?"
OUTPUT: Current evidence suggests that the SARS-CoV-2 virus spreads mainly between people who are in close contact with each other, for example at a conversational distance. The virus can spread from an infected person’s mouth or nose in small liquid particles when they cough, sneeze, speak, sing or breathe. Another person can then contract the virus when infectious particles that pass through the air are inhaled at short range (this is often called short-range aerosol or short-range airborne transmission) or if infectious particles come into direct contact with the eyes, nose, or mouth (droplet transmission). The virus can also spread in poorly ventilated and/or crowded indoor settings, where people tend to spend longer periods of time. This is because aerosols can remain suspended in the air or travel farther than conversational distance (this is often called long-range aerosol or long-range airborne transmission). People may also become infected when touching their eyes, nose or mouth after touching surfaces or objects that have been contaminated by the virus.
----
INPUT: Answer the following question in a single paragraph: "When do infected people transmit the SARS-CoV-2 virus?"
OUTPUT: Whether or not they have symptoms, infected people can be contagious and the virus can spread from them to other people. Laboratory data suggests that infected people appear to be most infectious just before they develop symptoms (namely 2 days before they develop symptoms) and early in their illness. People who develop severe disease can be infectious for longer. While someone who never develops symptoms can pass the virus to others, it is still not clear how frequently this occurs and more research is needed in this area.
----
INPUT: Answer the following question in a single paragraph: "Are there certain settings where COVID-19 can spread more easily?"
OUTPUT: Yes, any situation in which people are in close proximity to one another for long periods of time increases the risk of transmission. Indoor locations, especially settings where there is poor ventilation, are riskier than outdoor locations. Activities where more particles are expelled from the mouth, such as singing or breathing heavily during exercise, also increase the risk of transmission. The “Three C’s” are a useful way to think about this. They describe settings where transmission of the COVID-19 virus spreads more easily: Crowded places, Close-contact settings -especially where people have conversations very near each other-, Confined and enclosed spaces with poor ventilation. The risk of COVID-19 spreading is especially high in places where these “3Cs” overlap.
----
INPUT: Answer the following question in a single paragraph: "{question}"
OUTPUT:
```

A.5 Extract claim prompt

Extract a list with all the atomic facts about "{ENTITY}" extracted from the following paragraph. At all times when a pronoun is used instead of "{ENTITY}", replace the pronoun with "{ENTITY}". Write only the facts using the format "-- fact\n-- fact\n-- fact".
"{INPUT}"

A.6 Claim to question prompt (Insurance domain)

I will provide a statement containing one atomic fact about the insurance concept "Equipment Breakdown Insurance". Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example:
Statement: Motor failures are included in the coverage, such as engine, transmission, or alternator issues.
Question: Equipment breakdown insurance usually covers motor failures such as what?
Statement: Equipment Breakdown Insurance (EBI) covers equipment failure or mechanical breakdown due to internal failure or mechanical malfunction.
Question: Equipment Breakdown Insurance (EBI) covers equipment failure or mechanical breakdown due to what?
Statement: Specific endorsements or riders may be required to be added to the policy for equipment breakdown insurance coverage.
Question: What may be required to be added to the policy for equipment breakdown insurance coverage?
I will provide a statement containing one atomic fact about the insurance concept "Intentional Acts". Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example:
Statement: This exclusion is common in both life and health insurance policies, as well as in other forms of insurance.
Question: Are intentional acts a common exclusion in life and health insurance policies?
Statement: Intentional acts, such as fraud or forgery, are typically not covered by insurance policies.
Question: Are intentional acts, such as fraud or forgey tipically covered by insurance policies?
I will provide a statement containing one atomic fact about the insurance concept "Theft, Disappearance and Destruction". Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example:
Statement: Theft coverage may include coverage for the cost of replacing stolen property or the value of any damage caused to the item
Question: Theft coverage may include coverage for the cost of what?
Statement: Theft, Disappearance and Destruction can be excluded in insurance policies for intentional damage caused by the insured.
Question: Theft, Disappearance and Destruction can be excluded in insurance policies for intentional damage caused by whom?
I will provide a statement containing one atomic fact about the insurance concept {ENTITY}. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example:
Statement: {STATEMENT}
Question:

A.7 Claim to question prompt (Health domain)

I will provide a statement containing one atomic fact about the medical concept "Vaccines" in the context of COVID-19. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example:
Statement: COVID-19 vaccines can help people prevent mild and moderate illness from COVID-19 and greatly reduce the risk of severe illness that may lead to hospitalization and death.
Question: How effective are COVID-19 vaccines in preventing the infection of the virus?
Statement: Vaccination against COVID-19 by Pfizer, Moderna, and AstraZeneca requires a 21-day gap between doses. Question: What is the recommended spacing between doses of the COVID-19 vaccines?
Question: What is the recommended spacing between doses of the COVID-19 vaccines?
I will provide a statement containing one atomic fact about the medical concept "ACE2 gene" in the context of COVID-19. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example:
Statement: ACE2 gene may allow entry of SARS-CoV-2 into the cells of the airway epithelium.
Question: How does the ACE2 gene play a role in viral entry into cells in COVID-19?
Statement: Having a high level of ACE2 in the airways is one of the risk factors for severe COVID-19 disease.
Question: Which is the relation between ACE2 and the level of severity of a COVID-19 infection?
I will provide a statement containing one atomic fact about the medical concept "Mental Depression" in the context of COVID-19. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example:
Statement: The global prevalence of depression and anxiety symptoms increased during the pandemic.
Question: How does the prevalence of mental depression changed among the general population during the COVID-19 pandemic?
Statement: The disruptions and associated economic impacts are likely to increase the risk of developing depression in some people.
Question: What are some risk factors that can contribute to the development of mental depression in individuals during the COVID19 crisis?
I will provide a statement containing one atomic fact about the medical concept {ENTITY}. Please rephrase the following statement into a specific question testing knowledge of the key fact in the statement. For example:
Statement: {STATEMENT}
Question:

A.8 Answer questions prompt (Insurance domain)

INPUT: Answer the following question with a short answer: "What does liability insurance protect policyholders from?"
OUTPUT: Liability insurance protects policyholders from financial loss resulting from claims or lawsuits filed against them for injuries or damages they're deemed responsible for.

INPUT: Answer the following question with a short answer: "What expenses may be covered by liability insurance?"
OUTPUT: Liability insurance may cover expenses related to legal fees, settlements, and damages resulting from third-party claims against the insured party.

INPUT: Answer the following question with a short answer: "What is another name for Liability coverage?"
OUTPUT: Another name for Liability coverage is "Third-party coverage."

INPUT: Answer the following question with a short answer: "{question}"
OUTPUT:

A.9 Answer questions prompt (Health domain)

```
INPUT: Answer the following question with a short answer: "How does COVID-19 spread between people?"
OUTPUT: COVID-19 can be spread through the air by sneezing, talking, coughing, or by breathing.
----
INPUT: Answer the following question with a short answer: "Can someone spread COVID-19 if they are asymptomatic?"
OUTPUT: A person with COVID-19 who is asymptomatic is still contagious and can spread the infection to others.
----
INPUT: Answer the following question with a short answer: "How long does SARS-CoV-2 survive on surfaces?
OUTPUT: SARS-CoV-2 can persist on some surfaces for several days.
----
INPUT: Answer the following question with a short answer: "{question}"
OUTPUT:
```

A.10 Judge-based factuality estimator prompt

```
You are a helpful and critical assistant. Your task is to evaluate the factuality of a given text. You should give an overall factuality score, ranging from 0 to 1, 0 meaning absolutely false and 1 completely truthful. Be extremely critical when providing the score. Structure the output as a JSON object, ensuring that all fields are included and meticulously populated according to the following structure:

    class Output(BaseModel):
        factuality_score: float

Provide only the JSON object. Exclude comments and extra text.

"{TEXT}"
```

A.11 GPT similarity score prompt

```
### TEXT_A:
{TEXT_A}

### TEXT_B:
{TEXT_B}

### QUESTION:
How similar is the information (facts, names, dates, values) present in the TEXT_A with information present in the TEXT_B?
Answer with a number between 0 and 100.

### ANSWER:
Similarity Number:
```

A.12 Wikipedia search function prompt

Given an input term and its definition in a specific domain, give me the link to its corresponding Wikipedia article. If the input term is not covered in Wikipedia, give me the link to the Wikipedia article of the entity that is semantically closest to it. Otherwise, return the closest hypernym entity in the domain. Just give me the json output as in the following example:

```
input term: Special Endorsement
definition: Special Endorsement - Written language appended to an insurance policy that changes the coverage in regards to special circumstances.
domain: Insurance
[
{
"input term" : "Special Endorsement"
"definition" : "Special Endorsement - Written language appended to an insurance policy that changes the coverage in regards to special circumstances."
"retrieved term" : "Endorsement (insurance)"
"link" : "https://en.wikipedia.org/wiki/Endorsement"
}
]
```

A.13 Examples

Entity	Question	Llama 2-7B answer	LoFTune-EFS (Health) answer
Excess Liability Insurance	What is excess liability insurance and how does it differ from primary liability coverage?	Excess liability insurance, **also known as umbrella insurance**, is a type of liability coverage that provides additional protection above and beyond the limits of a primary liability policy. It typically kicks in when a claim or lawsuit exceeds the policy's maximum coverage limit, offering additional financial protection for the policyholder. (. . .)	Excess liability insurance is a type of insurance that provides additional coverage beyond the primary liability coverage. It is designed to provide additional protection against claims that exceed the primary liability coverage. Excess liability insurance typically covers claims that exceed the policy limit of the primary liability coverage, providing additional protection in the event of a large or catastrophic claim. In contrast to primary liability coverage, which provides protection against claims up to a certain limit, excess liability insurance provides additional coverage for claims that exceed that limit. (. . .)
Indemnification Agreement	What is an indemnification agreement in the insurance industry?	An indemnification agreement in the insurance industry is a contract between two parties, typically an insurer and an insured, that outlines the responsibilities and obligations of each party in the event of a covered loss or damage. The agreement typically states that the insurer will cover the cost of any damages or losses incurred by the insured, **while the insured agrees to indemnify the insurer for any damages or losses that exceed the policy limits**. In other words, **the insured agrees to reimburse the insurer for any additional costs incurred by the insurer as a result of the claim**.	An indemnification agreement is a contractual agreement between two parties that establishes the terms and conditions for **one party to compensate the other for any losses or damages that may occur**. In the insurance industry, an indemnification agreement is often used to transfer liability for certain risks or claims from one party to another. This type of agreement is commonly seen in situations where one party provides a service or product to another party and is exposed to potential liability for any damages or losses that may arise from that service or product. For example, a contractor may agree to indemnify a property owner for any damages or losses that may occur during the construction of a building. (. . .)
Umbrella Liability Insurance	Are there any limitations or exclusions to umbrella liability insurance coverage?	Umbrella liability insurance, also known as excess liability insurance, provides additional coverage beyond what is typically provided by standard liability insurance policies. However, it is important to note that umbrella insurance does have limitations and exclusions. For example, umbrella insurance may not cover certain types of liabilities, such as those related to professional negligence or intentional acts. Additionally, **umbrella insurance may only cover certain types of claims, such as those related to bodily injury or property damage**. (. . .)	Umbrella liability insurance is an additional layer of coverage that provides financial protection beyond the limits of a primary liability insurance policy. However, like any insurance policy, there are limitations and exclusions to this type of coverage. Generally, umbrella insurance provides additional coverage for liabilities that exceed the limits of a primary policy. However, it may not cover certain types of liabilities, such as those arising from intentional acts or illegal activities. Additionally, umbrella insurance policies may have specific exclusions, such as those related to certain types of claims or damages. It's important to carefully review the policy's terms and conditions to understand what is and is not covered. (. . .)

Table 5: Examples of questions and answers generated by the baseline (Llama 2-7B) and LoFTune-EFS in the insurance domain. Llama 2 7B's non-factual output appears in red and factual generation after LoFTune-EFS is in green. Note that factual outputs generated by LoFTune-EFS, which omit non-factual content present in the output of Llama 2-7B, appear normally.

Entity	Question	Llama 2-7B answer	LoFTune-EFS (Health) answer
Pharmaceutical Preparations	Can pharmaceutical preparations such as corticosteroids help in the treatment of severe COVID-19 cases?	Yes, corticosteroids can help to improve the clinical condition of severe COVID-19 cases, for example by reducing the need for mechanical ventilation. However, **they do not appear to improve survival.**	Yes, corticosteroids may be used to treat people with severe COVID-19. **They can help reduce the immune response that can cause inflammation and can help reduce the severity of COVID-19 illness.** Corticosteroids are also used to treat people who are on ventilators to reduce the inflammation in their lungs. However, they should not be used to treat people who have mild disease, as they can have serious side effects. Corticosteroids should only be used under the supervision of a health care provider.
M protein, multiple myeloma	Are individuals with multiple myeloma and elevated levels of M Protein at a higher risk of severe COVID-19 infection?	**There is currently no evidence that people with multiple myeloma are at higher risk of severe COVID-19 infection.** However, people with multiple myeloma may be at increased risk of complications if they become infected with COVID-19.	**Yes, individuals with multiple myeloma and elevated levels of M Protein may be at an increased risk of severe COVID-19 infection.** The COVID-19 pandemic has highlighted the importance of protecting vulnerable populations, including people with underlying health conditions. People with multiple myeloma and elevated levels of M Protein may be at an increased risk of severe COVID-19 infection due to their underlying health conditions. It is important for individuals with multiple myeloma and elevated levels of M Protein to take precautions to protect themselves from COVID-19, such as wearing a mask, practicing social distancing, and avoiding crowded spaces. If you are concerned about your risk of severe COVID-19 infection, you should speak with your healthcare provider.
Mutations	What impact can mutations have on the transmissibility of the virus?	Mutations can change the way that the virus behaves, including its ability to spread between people. However, **there is no evidence that mutations in the SARS-CoV-2 virus have changed the way it spreads.**	Mutations in the virus can affect its transmissibility. For example, **the Delta variant, which is currently the dominant variant circulating globally, has several mutations that may affect how easily it spreads.** Some of these mutations may affect the structure of the spike protein, which is the part of the virus that allows it to attach to cells in the body. Other mutations may affect how the virus interacts with the immune system, making it more difficult for the body to fight off infection.

Table 6: Examples of questions and answers generated by the baseline (Llama 2-7B) and LoFTune-EFS in the health domain. Llama 2 7B's non-factual output appears in red and factual generation after LoFTune-EFS is in green. Note that factual outputs generated by LoFTune-EFS, which omit non-factual content present in the output of Llama 2-7B, appear normally.