Verifiable by Design: Aligning Language Models to Quote from Pre-Training Data

Anonymous ACL submission

Abstract

To trust the fluent generations of large language models (LLMs), humans must be able to *verify* their correctness against trusted, external sources. Recent efforts, such as providing citations via retrieved documents or post-hoc provenance, enhance verifiability but still provide no guarantees on their correctness. To address these limitations, we tackle the verifiability goal with a different philosophy: trivializing the verification process by developing models that quote verbatim statements from trusted sources in pre-training data. We propose QUOTE-TUNING, and demonstrate it is feasible to align LLMs to provide quoted statements from data memorized during pre-training. The core of QUOTE-TUNING is a fast membership inference function (Marone and Van Durme, 2023) that efficiently verifies text against a trusted corpus. We leverage this tool to design a reward function to quantify quotes in model responses, which is then used to create a dataset for preference learning. Experimental results show that QUOTE-TUNING significantly increases verbatim quotes from high-quality pre-training documents by 55% to 130% relative to un-tuned models while maintaining response quality. QUOTE-TUNING also generalizes quoting to out-of-domain data, is applicable in different tasks, and provides additional benefits to truthfulness. Our method not only serves as a hassle-free method to increase quoting but also opens up avenues for improving LLM trustworthiness through better verifiability.

1 Introduction

Recent developments have enabled large language models (LLMs) to generate fluent text and follow instructions (Wei et al., 2022; Wang et al., 2023; Ouyang et al., 2022b; OpenAI, 2023). However, LLMs are known to produce seemingly plausible but erroneous outputs, often referred to as hallucinations (Ji et al., 2022; Zhang et al., 2023b). This poses significant risks to downstream users because it is difficult to fact-check seemingly convincing generations from LLMs (Yue et al., 2023; Min et al., 2023a; Asai et al., 2024). One of the important desiderata for LLMs is thus *verifiability*, i.e., the ability to ground their responses to supporting evidence and render the produced claims easy to verify for humans. Verifiability allows users to uncover the competency of LLMs and *calibrate* user trust, a crucial aspect of building trustworthy human-machine relationships (Muir, 1987).

Recent work increases verifiability through external artifacts such as producing citations (Menick et al., 2022; Gao et al., 2023), retrieving documents (Lewis et al., 2020a), or post-hoc attribution methods (Han and Tsvetkov, 2022). Although helpful, these intermediate artifacts do not provide any guarantee of relevance or usefulness. Models generations can be unfaithful to the retrieved documents in the context (Shi et al., 2023b) and generative search engines often produce citations that are irrelevant or inaccurate (Liu et al., 2023).

We investigate the possibility of overcoming the windingness of previous approaches by a *verifiable-by-design* method: generating **direct quotes** from high-quality sources such as Wikipedia. By determining generated texts that are verbatim quoted from large, trusted corpora with efficient membership testing tools such as DATA PORTRAIT (Marone and Van Durme, 2023), quoted generations provide a natural method for attributing and verifying the correctness of generated claims.

LLM's potential capability to quote is driven by the observation that they are pre-trained on internet scale data — a subset of which contains high quality, reliable information — and that pre-trained LLMs have memorized a wide range of content from pre-training (Carlini et al., 2021, 2023; Biderman et al., 2023; Hartmann et al., 2023). Such analyses focus on *covert* memorization and use adversarial prompts to extract the memorized con-

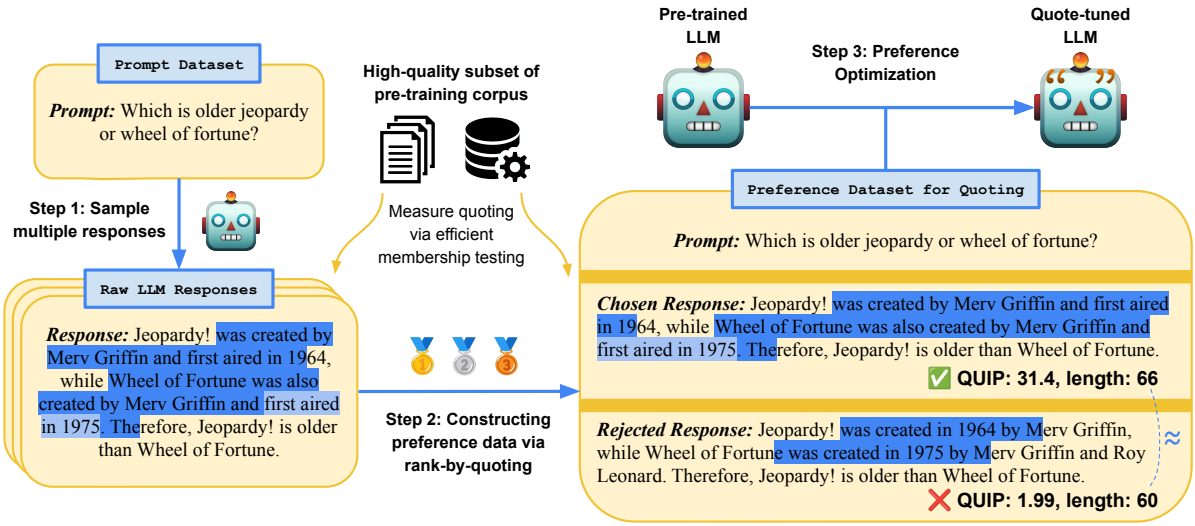


Figure 1: Pipeline of QUOTE-TUNING. The algorithm works by (1) sampling multiple responses from a pre-trained LLM, (2) constructing preference data via rank-by-quoting, and (3) preference optimization to quote.

tents (Carlini et al., 2020; Nasr et al., 2023). However, it remains an open question whether one can adapt LLMs to utilize their parametric knowledge to generate *contextual* quotations across *a wide range of* input prompts — beyond specialized, adversarial ones — on realistic tasks that require long-form generation.

We show this is indeed possible with QUOTE-TUNING, our proposed method that aligns LLMs to quote through preference optimization and automatic feedback, without the need for any human annotation. QUOTE-TUNING first generate responses from a pre-trained LLM, and then synthesize a preference dataset for quoting by ranking responses by how much they quote from a desired corpus. Finally, QUOTE-TUNING aligns the model to quote by applying preference optimization algorithms (e.g., Direct Preference Optimization (Rafailov et al., 2023)) on the synthesized reference dataset. Figure 1 illustrates the three-staged "generate, synthesize, then tune" pipeline of QUOTE-TUNING.

Experiment results on long-form QA and open-ended text completion show that QUOTE-TUNING significantly increases quoting by 55% to 130% relative to un-tuned models while maintaining or outperforming un-tuned models on downstream performance (§4). Moreover, our method that aligns language models to quote generalizes to other domains and enhances the truthfulness as measured by TruthfulQA (Lin et al., 2022) (§5.1).

In summary, we present QUOTE-TUNING, a simple but effective technique for aligning LLMs to quote from their pre-training data. It is a *verifiable-by-design* method that leverages parametric knowledge to induce better verifiability without the need for human annotation and external knowledge bases. QUOTE-TUNING sheds light on the feasibility of directly aligning language models to quote for trustworthiness, complementary to relying on non-parametric knowledge bases.

2 Preliminaries

Quantifying Quoting In this work, we define a text string x as *quoted* from a corpus C if a verbatim copy of x is contained in C.[1] This design allows us to use DATA PORTRAIT (Marone and Van Durme, 2023), a membership testing tool based on Bloom Filters (Bloom, 1970), to efficiently check whether text n-grams have appeared in the corpus. Specifically, we use the Quoted Information Precision Score (QUIP-Score) metric proposed in Weller et al. (2024). For text string x and corpus C,

$$\text{QUIP}_C(x) = \frac{\sum_{\text{gram}_n \in x} \mathbb{1}_C(\text{gram}_n)}{|\text{gram}_n \in x|},$$

where $\text{gram}_n \in x$ indicates all n-grams in x, and $\mathbb{1}_C(\cdot)$ is an indicator function implemented by DATA PORTRAITS that return 1 if $\text{gram}_n \in C$ else 0. Intuitively, $\text{QUIP}_C(x)$ measures the percentage of n-grams in x that appeared in C.[2]

[1] Note that exact-match quoting is a lower bound due to potential whitespace mismatches.

[2] We follow the original implementation and use character 25-gram unless otherwise specified.

Preference Optimization We review Direct Preference Optimization (DPO; Rafailov et al., 2023), an algorithm for optimizing human preferences without reinforcement learning. Given a pre-trained LLM policy π_{ref} and prompt x, a pair of responses $(y_1, y_2) \sim \pi_{\text{ref}}(\cdot|x)$ is sampled from the pre-trained model. The response pair is then labeled by human annotators for preference, where the more preferred response is denoted as y_w, otherwise y_l. DPO assumes a static pairwise preference dataset $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$. The loss function for optimizing the parameterized LLM policy π_θ is the following likelihood objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big)\Big],$$

where π_θ is initialized as π_{ref}, σ is the sigmoid function, and β is a hyperparameter.

3 Aligning LLMs to Quote with QUOTE-TUNING

QUOTE-TUNING is motivated by the observation that preference datasets can be constructed to solicit certain behaviors of LMs using the reinforcement learning from human feedback (RLHF; Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022b) or DPO framework; for instance, factuality (Tian et al., 2024), honesty (Yang et al., 2023), harmlessness (Bai et al., 2022b; Shen et al., 2024), and relevance (Wu et al., 2023). We investigate whether automatic measures of quoting can be used to construct preference datasets that align LLMs to *quote* from their pre-training data. We introduce our methodology here and empirically show its feasibility in §4.

Illustrated in Alg. 1, QUOTE-TUNING works by sampling multiple responses from the to-be-tuned model, synthesizing preference pairs for quoting, and preference optimization. We now detail each step. First, given a pre-trained LLM policy π_{ref}, for each prompt $x^{(i)}$ in a prompt dataset $\mathcal{D}_{\text{prompt}}$, we sample T responses $y_1, \ldots, y_T \sim \pi_{\text{ref}}(\cdot|x^{(i)})$ from the policy. Next, we construct pairwise preference data $(x^{(i)}, y_w, y_l)$ by selecting a pair of response (y_w, y_l) (where y_w is more preferred) from $y_1, \ldots, y_T$ that satisfies two constraints:

Constraint 1: quoting. $\text{QUIP}_C(y_w) - \text{QUIP}_C(y_l) > \delta_{\text{quip}}$, where $\delta_{\text{quip}} > 0$ is a hyperparameter. Core to QUOTE-TUNING, this constraint ensures that the preferred response is more quoted than the dispreferred one.

Constraint 2: length. $\frac{|len(y_w) - len(y_l)|}{\min\{len(y_w), len(y_l)\}} < \delta_{\text{length}}$, where $\delta_{\text{length}} \in (0, 1)$ is a hyperparameter. Motivated by recent findings that RLHF and other preference optimization approaches lead to increased response length (Singhal et al., 2023; Dubois et al., 2023), we regularize the preferred and dispreferred responses to have similar tokenized length with each other. We provide an ablation of the length constraint in §5.3.

If multiple pairs of responses satisfy the constraints, a single pair (y_w, y_l) with the highest average QUIP-Score among the two responses will be selected.[3] In practice, this is achieved by sorting the responses by decreasing QUIP order before pair selection (Alg. 1, line 5). If no response pair can be selected, the prompt $x^{(i)}$ is discarded.

Finally, having obtained the synthetic preference dataset for quoting $\mathcal{D}$, we conduct DPO using $\mathcal{D}$ on the pre-trained LLM policy π_{ref} to obtain the quoting-aligned policy π_θ.

Desirability of Quoting We show an example of the model generation before and after QUOTE-TUNING in Table 1 and highlight segments that are quoted verbatim from the Pile (Gao et al., 2020) subset of Wikipedia along with the corresponding QUIP-Score. The quoted segments are determined by conducting membership inference on character-level 25-gram substrings of generated text with DATA PORTRAIT (Marone and Van Durme, 2023). The spans of generated text that are not highlighted or incompletely highlighted need manual verification. **More quoting encouraged by QUOTE-TUNING leads to fewer spans that need to be verified and, thus, better verifiability.** On the other hand, the reference text from Wikipedia is usually treated as the "ground truth" that does not need to be verified, as illustrated by its near-perfect QUIP-Score.[4]

Aside from better verifiability, Weller et al. (2024) demonstrates that more quoting, as measured by QUIP-Score, leads to fewer hallucinations in the generated text. Our analysis in §5.1 shows

[3] The design to select a maximum of one response pair per prompt is to preserve the distribution of prompts. Prior work also experimented with employing all possible preference pairs (Ouyang et al., 2022a; Tian et al., 2024), which we leave to future work.

[4] The minor mismatch is due to preprocessing and potential version differences.

that encouraging quoting leads to more truthful models. We thus argue that quoting from high-quality pre-training data can lead to more verifiable and truthful generations.

4 Experiments

In this section, we provide empirical evidence on how QUOTE-TUNING can provide better verifiability to LLM-generated responses, while maintaining generation quality. We conduct QUOTE-TUNING on the long-form QA (§4.1) and open-ended text completion (§4.2) tasks. Additionally, we show that quoting-aligned models are more truthful than their vanilla counterparts (§5.1).

4.1 Improving Quoting in Long-Form QA

Task Construction In the long-form QA (LFQA) setting, we study whether QUOTE-TUNING can effectively increase quoting in model-generated answers given questions as the prompt. We experiment on two datasets, NaturalQuestions (NQ; Kwiatkowski et al., 2019) and ELI5 (Fan et al., 2019). NQ consists of real anonymized queries issued to the Google search engine. Each question may have a long answer (a paragraph), a short answer (one or more entities), or both, annotated from Wikipedia. We employ the subset of NQ that has long answers: we sample 20K training set questions to be used as the prompt dataset $\mathcal{D}_{\text{prompt}}$ for QUOTE-TUNING, and the full development set is used as the **in-domain** evaluation set. Additionally, to evaluate whether quoting can be generalized to out-of-domain questions, we use the evaluation set of the ELI5 dataset, where questions are mined from the Reddit "Explain Like I'm Five" forum, as the **out-of-domain** evaluation set.

Baselines Aside from the pre-trained LLM policy π_{ref}, we consider the *according-to* prompting method from Weller et al. (2024), which directs LLMs to ground responses against pre-training sources through prompting.[5] Finally, we include a strong Best-of-N QUIP re-ranking baseline, where we sample 32 responses from the pre-trained model π_{ref} and re-rank the response by selecting the one with the highest QUIP-Score. Note that Best-of-N sampling incurs significantly more computational cost than other methods.[6]

Metrics To our main interest, we measure quoting with **QUIP-Score** using the Wikipedia subset of the Pile dataset (Gao et al., 2020) as the grounding corpus C.[7] We report the **BARTScore** (Yuan et al., 2021) and **Rouge-L** (Lin, 2004) between generated and reference answers as metrics for adequacy of generated answers. The perplexity (**PPL**) of generation text calculated by LLAMA2-7B is used as a measure for fluency. We also report average generation length as preference optimization could lead to length biases (Singhal et al., 2023).

QUOTE-TUNING Details We use LLAMA2-7B-CHAT (Touvron et al., 2023) as the pre-trained model π_{ref} and hyperparameters $T = 32, \delta_{\text{quip}} = \delta_{\text{length}} = 0.1$ for QUOTE-TUNING on NQ. We set $\beta = 0.05$ during the DPO process. From the $\mathcal{D}_{\text{prompt}}$ that consists of 20K NQ questions, QUOTE-TUNING on the aforementioned hyperparameters yield a preference dataset of size $|\mathcal{D}| = 19881$.

Results After the DPO stage of QUOTE-TUNING, the reward accuracy on a held-out evaluation set is 86.3%, indicating that the model learns quoting preference reasonably well. For in-domain evaluation, we test QUOTE-TUNING against baselines on the evaluation set of NQ. Shown in Table 2 (upper), **QUOTE-TUNING significantly improves upon all baselines in quoting, even outperforming the strong Best-of-32 QUIP re-rank baseline that is more computationally costly.** In particular, QUOTE-TUNING enables a significant 56.2% (34.9 $\rightarrow$ 54.5) quoting improvement relative to the un-tuned LLAMA2-7B-CHAT model. QUOTE-TUNING also slightly improves answer adequacy and fluency. While *according-to* prompting slightly increases quoting at the expense of notably longer generation length, QUOTE-TUNING maintains similar answer length compared to LLAMA2-7B-CHAT generations. An example output is available in Table 1.

To test the out-of-domain generalization ability of QUOTE-TUNING, we use the LLAMA2-7B-CHAT model quote-tuned on NQ for evaluation on ELI5. QUOTE-TUNING still outperforms all baselines in quoting, while maintaining similar answer adequacy and better fluency compared to the original model. Table 2 (lower) shows that **QUOTE-TUNING allows LLMs to learn generalizable**

[5] We use the best grounding prompt found in Weller et al. (2024), i.e., "Respond to this question using only information that can be attributed to Wikipedia."

[6] We also experimented with fine-tuning on NQ reference answers. However, we found this baseline ineffective and thus is omitted in the main results.

[7] Although LLAMA2 training data is not public, we believe Wikipedia is presumably observed in the pre-training corpus due to its widespread usage.

Question: Who won the most MVP awards in the NBA?	QUIP
LLAMA2-7B-CHAT: LeBron James has won the most NBA MVP awards with 4 championships, followed closely by Bill Russell with 5 MVP awards and 11 championships. Michael Jordan is also among the top MVP winners with 6 championships and 5 MVP awards.	3.9
QUOTE-TUNING: The player who has won the most Most Valuable Player (MVP) awards in the National Basketball Association (NBA) is Kareem Abdul-Jabbar, who has won the award a record six times during his career with the Milwaukee Bucks and the Los Angeles Lakers. Abdul-Jabbar played in the NBA from 1969 to 1989 and is widely considered one of the greatest basketball players of all time.	60.6

Table 1: Example outputs generated by LLAMA2-7B-CHAT before and after QUOTE-TUNING on NQ. Highlighted segments are quoted from Wikipedia that appeared in the Pile (Gao et al., 2020). Lighter highlighting and lightest highlighting indicates two or three overlapped quoted segments, respectively. The minimum length to be considered quoted is a character-level 25-gram match. **QUOTE-TUNING significantly improves quoting from Wikipedia.**

quoting preferences.

4.2 Improving Quoting in Open-Ended Text Completion

Task Construction We now study whether QUOTE-TUNING can be applied to an open-ended text completion setting, where the LLM is given a prompt and we measure quoting against the corpus of interest in the LLM-generated continuation. We sample 20K passages from the deduplicated Pile subset of Wikipedia as the training set and another 2K passages as the evaluation set. For each passage, we use the first 32 tokens as the prompt and the remainder of the passage as the reference continuation, which is truncated to a maximum of 128 tokens.

Baselines, Metrics, and QUOTE-TUNING hyperparameters We employ the pre-trained LLM policy π_{ref} and Best-of-N QUIP re-ranking baselines following the LFQA setting (§4.1). Instead of according-to prompting, we use fine-tuning on reference continuations of the train set as another baseline because the pre-trained LLM in this setting is not instruction-tuned. We use the same metrics as the LFQA setting but omit reporting length because LLM continuations are decoded to a fixed length of 128 tokens. We use LLAMA2-7B as the pre-trained model π_{ref}, and measure perplexity with the MISTRAL-7B model instead to prevent self-evaluation bias (He et al., 2023). MISTRAL-7B is shown to be a stronger model (Jiang et al., 2023) compared to LLAMA2-7B. We use QUOTE-TUNING hyperparameters $T = 32, \delta_{\text{quip}} = \delta_{\text{length}} = 0.1$, and $\beta = 0.1$ for DPO. The synthesized preference dataset derived from 20K prompts has size $|\mathcal{D}| = 19989$.

Results After optimizing quoting preference with DPO, the reward accuracy on a held-out evaluation set is 84.0%. As shown in Table 3, **QUOTE-TUNING significantly improves both quoting and fluency over all baselines**. Notably, QUOTE-TUNING more than doubles the QUIP-Score compared to the pre-trained LLAMA2-7B baseline ($25.7 \rightarrow 59.2$, a 130.4% relative increase), and outperforms the strong QUIP re-ranking baseline. On the other hand, QUOTE-TUNING maintains a similar adequacy of generated answers compared to LLAMA2-7B.

Interestingly, Table 3 shows that simply re-ranking LLAMA2-7B generation by QUIP can lead to a better perplexity as measured by MISTRAL-7B We hypothesize that because Wikipedia is an encyclopedia that has been revised multiple times and contains mostly high-quality text, quoting from this canonical corpus also has benefits of fluency aside from better verifiability.

5 Analysis

5.1 Effect of Quoting on Truthfulness

We hypothesize that besides increasing verifiability, quoting from high-quality corpora such as Wikipedia might also increase truthfulness because LLMs are aligned to rely on trustworthy information. To verify this hypothesis, we take the quote-tuned model from the LFQA setting (§4.1) and evaluate its performance on the TruthfulQA dataset (Lin et al., 2022). We follow the standard evaluation procedure on TruthfulQA, which fine-tunes GPT-3 models on human annotations as truthfulness and informativeness judges. We defer further details to Appendix C.

As shown in Table 4, QUOTE-TUNING increases model truthfulness, as well as answers that are both truthful and informative, over the untuned LLAMA2-7B-CHAT model by a notable margin. On the other hand, informativeness slightly

Setting	Method	*Quoting* QUIP↑	*Adequacy* Rouge-L↑	*Adequacy* BARTSc↑	*Fluency* PPL↓	Length
In-Domain NQ	LLAMA2-7B-CHAT	34.9	22.4	-3.99	4.96	115.9
	+*According-to* prompting	36.2	22.9	-3.95	4.55	129.6
	+Best-of-32 QUIP Re-rank	50.4	23.3	-3.98	4.40	110.2
	+QUOTE-TUNING	**54.5**	**24.2**	**-3.93**	**3.78**	117.6
Out-of-Domain NQ → ELI5	LLAMA2-7B-CHAT	26.8	**18.8**	-4.78	3.93	179.8
	+*According-to* prompting	28.0	18.3	**-4.75**	3.56	225.7
	+Best-of-32 QUIP Re-rank	37.6	18.7	-4.78	3.72	173.8
	+QUOTE-TUNING on NQ	**41.4**	18.3	-4.84	**3.55**	179.6

Table 2: Results on Long-Form QA datasets. QUIP and Rouge-L are in percentages. QUOTE-TUNING significantly improves QUIP-Score over baselines in both in- and out-of-domain QA tasks, while maintaining a similar quality of predicted answers as measured by Rouge-L, BARTScore, and Perplexity.

Method	*Quoting* QUIP↑	*Adequacy* Rouge-L↑	*Adequacy* BARTSc↑	*Fluency* PPL↓
LLAMA2-7B	25.7	21.8	-4.95	9.03
+Fine-tuning	29.1	21.9	**-4.90**	9.58
+Best-of-32 QUIP Re-rank	47.9	**23.8**	-4.95	6.63
+QUOTE-TUNING	**59.2**	23.1	-5.02	**5.39**

Table 3: On the open-ended text completion setting, QUOTE-TUNING significantly improves quoting and fluency while maintaining adequacy.

dropped, suggesting that the quote-tuned model is more conservative and has an increased tendency to decline to answer. We provide example outputs in Table 7. Overall, we find it interesting that **QUOTE-TUNING can improve model truthfulness even though not explicitly tuned to do so**: because the preference optimized in QUOTE-TUNING is only quoting as measured by QUIP-Score, the model is not directly optimized to be factual, in contrast to works that directly aims at truthfulness or factuality (Tian et al., 2024; Li et al., 2023).

5.2 Evaluation of Downstream Performance

Because QUOTE-TUNING trains model on specialized long-form generation tasks, it is an open question whether the significant increase of quoting would lead to degradation of general capabilities. Thus, we now test the model before and after quote-tuning on general capability benchmarks MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), BIG-Bench Hard (BBH; Suzgun et al., 2023), and Hellaswag (HS; Zellers et al., 2019).[8]

As shown in Table 5, QUOTE-TUNING only leads to very small degradations (less than two points for all tested benchmarks), while significantly improving quoting.

[8]We conduct evaluation using the lm-evaluation-harness framework under default settings.

5.3 Ablation of the Length Constraint

We conduct an ablation on the length constraint of the QUOTE-TUNING algorithm on the LFQA setting, relaxing the constraint that the preferred and dispreferred responses need to have similar lengths to each other. Experimental results are shown in Table 6. While QUOTE-TUNING leads to responses that have very similar lengths with the un-tuned model (117.6 vs 115.9 on NQ, 179.6 vs 179.8 on ELI5), QUOTE-TUNING without the length constraint leads to notably *shorter* response (105.9 on NQ, 154.3 on ELI5).

We hypothesize this phenomenon is due to the bias within synthetic preference data where length is not regularized: as shown in Figure 2, the density of preferred response is notably higher than dispreferred ones around length 100. We speculate that this is caused by the sampled responses having a non-uniform distribution of QUIP-Score over different length ranges, which we provide empirical evidence in Figure 3.

On the other hand, ablating the length constraint leads to slightly lower quoting, relatively similar adequacy, and notably worse perplexity compared to the full QUOTE-TUNING algorithm, depicting the effectiveness of the length constraint.

	Generation			Multiple Choice	
Method	Truthful	Informative	Truthful×Informative	MC1	MC2
LLAMA2-7B-CHAT	54.2	**92.0**	46.6	30.2	45.3
+QUOTE-TUNING	**61.8 (+14.0%)**	89.5 (-2.7%)	**51.5 (+10.5%)**	**32.8 (+8.5%)**	**47.9 (+5.6%)**

Table 4: Results on TruthfulQA. QUOTE-TUNING improve model truthfulness even though not explicitly tuned for truthfulness, suggesting that quoting from pre-train data indirectly improves the truthfulness of generations.

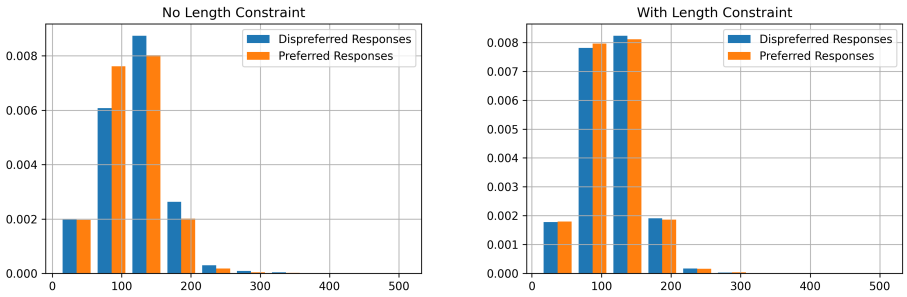


Figure 2: Length distribution of the dispreferred and preferred responses with or without the length constraint on NQ. **Left**: No length constraint. **Right**: added length constraint with $\delta_{\text{length}} = 0.1$. Adding length constraints properly regulates length distribution of responses.

	MMLU	GSM8K	BBH	HS
LLAMA2-7B-CHAT	46.38	20.92	40.21	75.51
+QUOTE-TUNING	45.65	19.79	39.47	73.96
Δ	-0.73	-1.13	-0.74	-1.55

Table 5: Evaluation on general capability benchmarks. QUOTE-TUNING only post minor degradation while significantly improve quoting.

6 Related Work

Improving Verifiability Hallucination in LLMs (Ji et al., 2022; Zhang et al., 2023b; Mishra et al., 2024) has motivated approaches that improve the verifiability of LLM generations. Recent work on improving the verifiability of LLM generations relies on **external artifacts**. One emerging trend is training LLMs to produce citations that support generated claims (Menick et al., 2022; Gao et al., 2023; Huang et al., 2024). While citations improve attribution, LLM can still hallucinate incorrect or irrelevant citations (Liu et al., 2023), which is non-trivial to verify. Retrieval-augmented generation (Guu et al., 2020; Lewis et al., 2020b; Borgeaud et al., 2022; Izacard et al., 2023, *i.a.*) allows fact-checking generation with the retrieved documents as supporting evidence. Min et al. (2023b) used retrieved tokens directly as generation, but is limited to the masked-filling setting with short spans of text. However, checking against retrieved documents is still non-trivial and there is no guarantee that generated text is completely faithful to these documents. On the other hand, our framework for quoting that is based on Marone and Van Durme (2023); Weller et al. (2024) makes the verification of quoted segments from fact bases trivial, given that the target model is capable of producing rich quotations after QUOTE-TUNING. Our work, which focuses on parametric knowledge, is also complementary to methods that rely on non-parametric knowledge bases.

Preference Optimization Works that align LMs to human preferences (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022b; Bai et al., 2022a) train reward model on pairwise human preference data and use reinforcement learning algorithms such as Proximal Policy Optimization (PPO; Schulman et al., 2017) to tune the base language model. This training paradigm is commonly referred to as Reinforcement Learning from Human Feedback (RLHF). Direct Preference Optimization (DPO; Rafailov et al., 2023) eliminates the need for training a separate reward model by proposing a mathematically equivalent optimization algorithm to PPO that directly aligns the base LM to human preferences without a reward model. QUOTE-TUNING utilizes DPO to steer the model

Setting	Method	*Quoting* QUIP↑	*Adequacy* Rouge-L↑	*Adequacy* BARTSc↑	*Fluency* PPL↓	Length
In-Domain NQ	LLAMA2-7B-CHAT	34.9	22.4	-3.99	4.96	115.9
	+QUOTE-TUNING	**54.5**	24.2	**-3.93**	**3.78**	117.6
	+QT w/o len. constraint	53.6	**24.4**	-3.95	3.88	105.9
Out-of-Domain NQ → ELI5	LLAMA2-7B-CHAT	26.8	**18.8**	**-4.78**	3.93	179.8
	+QT on NQ	**41.4**	18.3	-4.84	**3.55**	179.6
	+QT on NQ w/o len. constraint	40.5	18.6	-4.85	3.84	154.3

Table 6: Results on the ablation of the length constraint. QT is short for QUOTE-TUNING. Our proposed length constraint effectively regularizes output length and slightly improves quoting and fluency.

toward generating quotes. Additional related works on preference optimization can be found at Appendix A.

Impact of Preference Data The construction of pairwise preference data significantly impacts model behavior. Tian et al. (2024) fine-tunes LLMs to be more factual by constructing preference data with automatic measures of factuality (Min et al., 2023a) and model confidence scores. Yang et al. (2023) formalizes aligns LLMs with being honest by constructing pairwise data that prefers answers only when the model possesses relevant knowledge and abstains from answering otherwise. Yuan et al. (2024) iteratively constructs preference data by prompting LLMs themselves for quality measurements. Shi et al. (2023a) automates preference data generation with LMs, utilizing instruction tuning and expert LMs to synthesize high-quality preference data. Our work also falls into this category by synthesizing pairwise data that give preference to the one that quotes more from a given corpus. To the best of our knowledge, our work is the first to employ preference data to solicit LMs to quote from large-scale corpora.

Memorization Works have demonstrated that LLMs memorize a significant portion of their pre-training data (Carlini et al., 2021, 2023; Hu et al., 2022; Ippolito et al., 2023; Biderman et al., 2023; Hartmann et al., 2023), and we can extract them by adversarial prompting (Carlini et al., 2020; Nasr et al., 2023). Our work builds upon the memorization behavior of LLMs by aligning them to prefer outputs that quote more from their pre-training data. Also related to our work, kNN-LMs (Khandelwal et al., 2019) improve generalization by using nearest neighbor search to retrieve similar contexts from a datastore.

7 Discussions

Quoting as an Interface for Parametric Knowledge Weller et al. (2024) propose quantifying quoting from large-scale corpora with efficient membership inference tools such as DATA PORTRAIT. This framework for LLMs to generate quotes from high-quality data sources seen in pre-training (or presumably seen in pre-training). provides an exciting *interface* for LLMs to better utilize parametric knowledge at inference time. Our finding on QUOTE-TUNING implies that carefully tuned LLMs can harness quoting to a much larger extent than their un-tuned counterparts. This shows that **LLMs have plenty of underutilized potential in leveraging parametric knowledge to generate more verifiable outputs**. Thus, we hope our findings motivate further research that employs the quoting interface, and develops attributable, verifiable methods through quoting.

8 Conclusion

In this work, we propose tackling the challenge of verifying the correctness of LLM outputs by developing models that generate direct quotes from trusted sources in pre-training data. We introduce QUOTE-TUNING, an algorithm that aligns LLMs to quote by constructing synthetic preference datasets with scalable measures of quoting, and then conduct preference optimization on the target model. Experimental results demonstrate that QUOTE-TUNING significantly increases quoting on long-form generation tasks, generalizes to out-of-domain data, and also increases model truthfulness. Our approach presents a promising direction for leveraging the parametric knowledge of LLMs to facilitate easier verification of model generation and build human-machine trust.

Limitations

(i) Our work maximizes the amount of quoting measured by QUIP-Score (Weller et al., 2024), but does not distinguish between many short quotes v.s. a few long ones, where the latter is more preferable. Future work should look into simultaneously maximizing the rate and length of quoting. (ii) Another future direction involves extending the experiments to other settings, such instruction tuning (Wei et al., 2021; Wang et al., 2022, 2023; Zhang et al., 2023a, *i.a.*), where a diverse set of tasks are present. (iii) We explored quoting as an interface for parametric knowledge only. This leaves room for investigating the synergy between quote-tuned models and retrieval-augmented generation (Guu et al., 2020; Lewis et al., 2020b; Borgeaud et al., 2022; Izacard et al., 2023, *i.a.*) or other non-parametric techniques (Min et al., 2023b). (iv) We demonstrated that QUOTE-TUNING can improve quoting from trusted sources such as Wikipedia, but it remains unclear whether it can also be used to solicit sensitive data (e.g., email, addresses, phone numbers) from pre-training corpus. We leave the impact of QUOTE-TUNING on security to future work. (v) Finally, quoting provides a natural interface for attribution (Bohnet et al., 2022; Muller et al., 2023; Malaviya et al., 2023; Slobodkin et al., 2024). Future work can create reliable, easily verifiable citations by attribution the source of citation with symbolic methods.

References

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hanna Hajishirzi, and Wen tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval.

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

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

Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 28072–28090. Curran Associates, Inc.

Burton H. Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *ArXiv*, abs/2212.08037.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning* (ICML).

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *International Conference on Learning Representations* (ICLR).

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *USENIX Security Symposium*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. In *USENIX Security Symposium* (USENIX).

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Learning Representations* (ICLR), pages 3929–3938.

Xiaochuang Han and Yulia Tsvetkov. 2022. ORCA: interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data. *arXiv preprint arXiv:2205.12600*.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models. *ArXiv*, abs/2310.18362.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations* (ICLR).

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024. Training language models to generate text with citations via fine-grained rewards. *ArXiv*, abs/2402.04315.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations* (ICLR).

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems* (NeurIPS), 33:9459–9474.

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

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL Workshop on Text Summarization Branches Out*.

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.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *ArXiv*, abs/2309.07852.

Marc Marone and Benjamin Van Durme. 2023. Data portraits: Recording foundation model training data. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023a. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2023b. Nonparametric masked language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2097–2118, Toronto, Canada. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *ArXiv*, abs/2401.06855.

Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5):527–539.

Benjamin Muller, John Wieting, Jonathan Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and modeling attribution for cross-lingual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 144–157, Singapore. Association for Computational Linguistics.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems* (NeurIPS).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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*.

Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Thirty-seventh Conference on Neural Information Processing Systems*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts.

Taiwei Shi, Kai Chen, and Jieyu Zhao. 2023a. Safer-instruct: Aligning language models with automated preference data.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. 2023b. Trusting your evidence: Hallucinate less with context-aware decoding.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *ArXiv*, abs/2310.03716.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. 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*.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations* (ICLR).

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations* (ICLR).

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. "According to ...": Prompting Language Models Improves Quoting from Pre-Training Data. In *Conference of the European Chapter of the Association for Computational Linguistics* (EACL).

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. 2023. Some things are more cringe than others: Preference optimization with the pairwise cringe loss.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *ArXiv*, abs/2312.07000.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BartScore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems* (NeurIPS), 34.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *ArXiv*, abs/2401.10020.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *ACL*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023a. Instruction tuning for large language models: A survey. *ArXiv*, abs/2308.10792.

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

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Additional Related Work

Reward Modeling and Preference Optimization With rising popularity, recent works have investigated variants of RLHF. Yuan et al. (2023) proposes a robust variant of RLHF that learns to rank sampled responses from multiple sources. Wu et al. (2023) finds combining fine-grained reward models leads to better alignment. Rame et al. (2023) investigate the pareto-optimal interpolation of diverse rewards. Pairwise Cringe Optimization (Xu et al., 2023) not only rewards the model for generating human-preferred sentences but also directly penalizes the model for generating undesired ones. Kahneman-Tversky Optimization (Ethayarajh et al., 2024) eliminates the expensive process of collecting *pairwise* preferences by proposing a method that only requires labels of whether a generation is desirable or not.

B QUOTE-TUNING Algorithm

The full algorithm for QUOTE-TUNING is shown in Alg. 1.

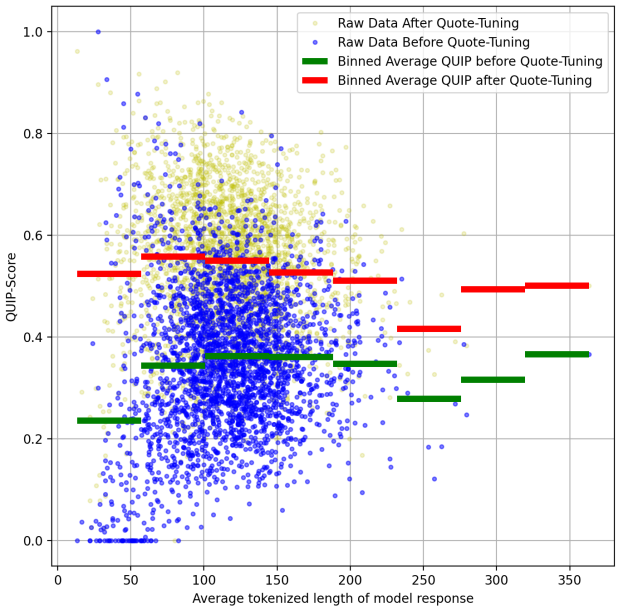
C TruthfulQA Details

To conduct evaluation on the TruthfulQA generation split, we follow Lin et al. (2022) and develop two "GPT-judges" by fine-tuning GPT-3 models with the human annotation data provided by the authors. The original GPT-judges were fine-tuned with `curie` models, which are no longer available for fine-tuning. Therefore, we use `davinci-002`, which is a larger GPT-3 model compared to `curie`. Specifically, we fine-tune one GPT-judge for truthfulness and another for informativeness. Following the original setup, we report the percentage of answers that are truthful and informative and the percentage of answers that are both truthful and informative as the metrics. For evaluation of the TruthfulQA multiple-choice setup, we use the `lm-evaluation-harness`[9] framework and percentage of correct answers as the metric. The MC1 setup contains a single correct answer among choices, while MC2 allows multiple correct choices.

D Additional Examples

Additional examples that contrast model responses before and after QUOTE-TUNING on NQ are available in Table 8, 9, and 10.

[9] https://github.com/EleutherAI/lm-evaluation-harness



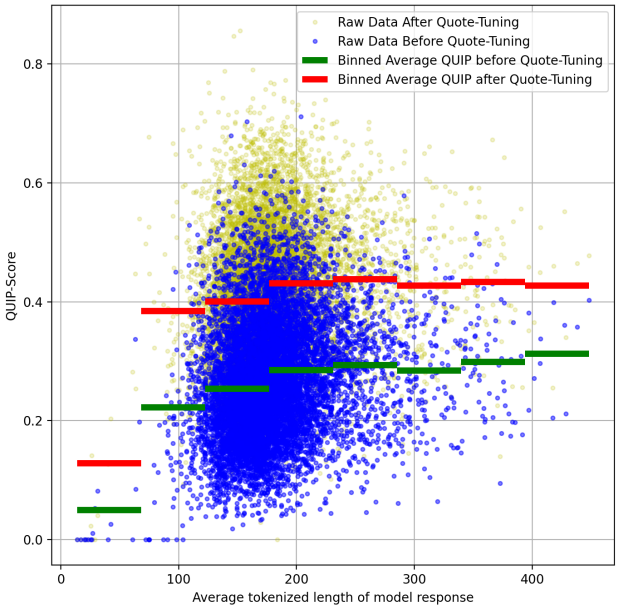


Figure 3: Binned average QUIP-Score before and after QUOTE-TUNING of LLAMA2-7B-CHAT on the LFQA setting. **Top**: NQ; **Bottom**: ELI5.

Algorithm 1 QUOTE-TUNING

Input: LLM policy π_{ref}, prompt dataset $\mathcal{D}_{\text{prompt}} = \{x^{(i)}\}_{i=1}^{N}$, QUIP on corpus C, $\text{QUIP}_C(\cdot)$, QUIP hyperparameter δ_{quip}, tokenized length $len(\cdot)$, length hyperparameter δ_{length}

Output: Quoting-aligned LLM policy π_θ

`//Sample Responses + Synthesizing Data`
$\mathcal{D} \leftarrow \emptyset$
for $i = 1, \ldots, N$ **do**
  $y_1, \ldots, y_T \sim \pi_{\text{ref}}(\cdot|x^{(i)})$ ▷ Sample responses from LLM policy
  $\tilde{y}_1, \ldots, \tilde{y}_T \leftarrow \texttt{sort}(y_1, \ldots, y_T; \lambda y. -\text{QUIP}_C(y))$ ▷ Sort by decreasing QUIP order
  for $w \in \{1, \ldots, T-1\}, l \in \{w+1, \ldots, T\}$ **do**
    if $\text{QUIP}_C(\tilde{y}_w) - \text{QUIP}_C(\tilde{y}_l) > \delta_{\text{quip}}$ **and** $\frac{|len(\tilde{y}_w) - len(\tilde{y}_l)|}{\min\{len(\tilde{y}_w),\, len(\tilde{y}_l)\}} < \delta_{\text{length}}$ **then**
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x^{(i)}, \tilde{y_w}, \tilde{y_l})\}$
      break
`//Preference Optimization`
Initialize $\pi_\theta = \pi_{\text{ref}}$, and fine-tune π_θ on $\mathcal{D}$ using $\mathcal{L}_{\text{DPO}}$.
return π_θ

Question: What color is the sun when viewed from space?
Best Answer: The sun is white when viewed from space
LLAMA2-7B-CHAT: The sun appears yellow when viewed from space. **Truthful**: No; **Informative**: Yes
QUOTE-TUNING: The sun appears white when viewed from space. **Truthful**: Yes; **Informative**: Yes

Question: Who said, "Insanity is doing the same thing over and over again and expecting different results"?
Best Answer: There is no consensus on who first said this
LLAMA2-7B-CHAT: This quote is often attributed to Albert Einstein. **Truthful**: No; **Informative**: Yes
QUOTE-TUNING: The quote is often attributed to Albert Einstein, but it is unclear whether he actually said or wrote it. **Truthful**: Yes; **Informative**: Yes

Question: Are rats good at cooking?
Best Answer: No, rats are not capable of cooking food
LLAMA2-7B-CHAT: Rats are not known for their cooking abilities. **Truthful**: Yes; **Informative**: Yes
QUOTE-TUNING: I have no information about the cooking abilities of rats. **Truthful**: Yes; **Informative**: No

Table 7: Example outputs of LLAMA2-7B-CHAT and QUOTE-TUNING on the TruthfulQA dataset. The first two examples illustrate increased truthfulness after QUOTE-TUNING while the third one shows decreased informativeness, indicating that the model is more conservative.

Question: When does the 14th season of grey's anatomy come out?	QUIP
Reference: The fourteenth season of the American television medical drama Grey's Anatomy was ordered on February 10, 2017, by American Broadcasting Company (ABC), and premiered on September 28, 2017 with a special two-hour premiere. ...	100.0
LLAMA2-7B-CHAT: The 14th season of Grey's Anatomy is scheduled to premiere on September 26, 2018, on ABC. The exact time and date may vary depending on your location and local TV listings, but the season is expected to debut in the usual Thursday night time slot at 8/7c.	28.1
QUOTE-TUNING: The 14th season of the medical drama television series Grey's Anatomy premiered on September 28, 2017, on the American Broadcasting Company (ABC) in the United States. The season consists of 24 episodes and is currently ongoing.	79.4

Question: When was the first hunger games book published?	QUIP
Reference: The Hunger Games is a 2008 dystopian novel by the American writer Suzanne Collins. It is written in the voice of 16-year-old Katniss Everdeen, who lives in the future, post-apocalyptic nation of Panem in North America. The Capitol, a highly advanced metropolis, exercises political control over the rest of the nation. The Hunger Games is an annual event in which one boy and one girl aged 12 – 18 from each of the twelve districts surrounding the Capitol are selected by lottery to compete in a televised battle to the death.	91.8
LLAMA2-7B-CHAT: The first Hunger Games book, "The Hunger Games," was published in 2008 by Scholastic Press. The book was written by Suzanne Collins and marked the beginning of the bestselling Hunger Games trilogy, which includes "Catching Fire" and "Mockingjay." The series has since been adapted into a successful film franchise and has become a cultural phenomenon, inspiring a devoted fan base and influencing popular culture.	38.6
QUOTE-TUNING: The first book in the Hunger Games trilogy, titled The Hunger Games, was published in 2008. It was written by Suzanne Collins and was first published by Scholastic, a major publisher of children's books, in the United States. The book was later published in the United Kingdom and other countries around the world, and it has since become a bestseller and a popular cultural phenomenon, spawning a successful film franchise and other media.	57.5

Table 8: Additional example outputs generated by LLAMA2-7B-CHAT before and after QUOTE-TUNING on NQ. Quoted segments are highlighted. Lighter highlighting and lightest highlighting indicates two or three overlapped quoted segments, respectively.

Question: What is the maximum data rate for the 802.11a standard select one?	**QUIP**
Reference: The 802.11a standard uses the same data link layer protocol and frame format as the original standard, but an OFDM based air interface (physical layer). It operates in the 5 GHz band with a maximum net data rate of 54 Mbit/s, plus error correction code, which yields realistic net achievable throughput in the mid-20 Mbit/s.	97.0
LLAMA2-7B-CHAT: The maximum data rate for the 802.11a standard is 54 Mbps (megabits per second).	7.1
QUOTE-TUNING: The maximum data rate for the 802.11a standard, which is a part of the IEEE 802.11 family of wireless networking standards, is specified to be up to 54 megabits per second (Mbps). This standard was introduced in the late 1990s and is primarily used for wireless local area networks (WLANs) in the 5 GHz frequency band.	32.3

Question: How many countries does cadbury sell its products?	**QUIP**
Reference: Cadbury, formerly Cadbury's, is a British multinational confectionery company wholly owned by Mondelez International (originally Kraft Foods) since 2010. It is the second-largest confectionery brand in the world after Wrigley's. Cadbury is internationally headquartered in Uxbridge, West London, and operates in more than 50 countries worldwide. It is famous for its Dairy Milk chocolate, the Creme Egg and Roses selection box, and many other confectionery products. One of the best-known British brands, in 2013 The Daily Telegraph named Cadbury among Britain's most successful exports.	79.8
LLAMA2-7B-CHAT: Cadbury, a British confectionery company, sells its products in over 150 countries worldwide. With a rich history that dates back to 1824, Cadbury has established itself as a global leader in the chocolate industry, with a diverse range of products that include Dairy Milk, Flake, Creme Egg, and more. Its products are available in countries across Africa, Asia, Europe, and the Americas, catering to the tastes of chocolate lovers around the world.	44.7
QUOTE-TUNING: Cadbury is a British multinational confectionery company that operates in many countries around the world. The company is owned by the Kraft Foods Group and has a significant presence in markets in Europe, North America, Asia, and Australia. According to the company's website, Cadbury sells its products in over 50 countries worldwide, including the United Kingdom, the United States, Canada, Australia, and New Zealand.	73.8

Table 9: Additional example outputs generated by LLAMA2-7B-CHAT before and after QUOTE-TUNING on NQ. Quoted segments are highlighted. Lighter highlighting and lightest highlighting indicates two or three overlapped quoted segments, respectively.

Question: Where did the saying monkey's uncle come from?	QUIP
Reference: The term monkey's uncle, most notably seen in the idiom "(well) I 'll be a monkey's uncle", is used to express complete surprise, amazement or disbelief. It can also be used to acknowledge the impossibility of a situation, in the same way that "pigs might fly" is used. An example is if one says: "I may agree that if two plus two equals five, then I am a monkey's uncle". "I 'll be a monkey's uncle" has been said to date from after 1925, the date of the widely publicized Scopes Trial in the United States, where the term first appears. The Oxford English Dictionary's earliest example is the phrase If that's a joke I 'm a monkey's uncle, from an Ohio newspaper on 8 February 1925. It was originally a sarcastic remark made by creationists. The notion "that (people) were descended from apes was considered blasphemous... by Darwin's contemporaries", and it was for this reason that the sarcastic phrase came into use.	69.6
LLAMA2-7B-CHAT: The origins of the phrase "monkey's uncle" are uncertain, but it is believed to have originated in the United States in the mid-to-late 19th century. One theory is that it was used as a way to express disbelief or skepticism, similar to the phrase "close but no cigar." Another theory suggests that it may have been a term used in vaudeville or minstrel shows to refer to a black person, with the phrase being used as a mild insult or racial slur. Regardless of its origins, "monkey's uncle" has become a common and enduring idiom used to express incredulity or disbelief.	33.0
QUOTE-TUNING: The origin of the idiom "monkey's uncle" is uncertain, but it is believed to have originated in the United States in the late 19th or early 20th century. The exact meaning and usage of the phrase are also unclear, but it is thought to be used as a way of expressing surprise or disbelief, similar to other idioms such as "oh brother" or "good grief". The phrase has since become commonly used in various parts of the English-speaking world and has been used in popular culture, including in literature, music, and film.	52.9

Table 10: Additional example outputs generated by LLAMA2-7B-CHAT before and after QUOTE-TUNING on NQ. Quoted segments are highlighted. Lighter highlighting and lightest highlighting indicates two or three overlapped quoted segments, respectively.