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

Jingyu Zhang **Marc Marone** **Tianjian Li** **Benjamin Van Durme**♡ **Daniel Khashabi**♡
Johns Hopkins University
Baltimore, MD
`{jzhan237,mmarone1,tli104}@jhu.edu`

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 provide no guarantees of their correctness. To address these limitations, we aim to improve verifiability with a different philosophy: *trivializing the verification process by developing models that quote verbatim statements from trusted sources in their pre-training data.*

We propose QUOTE-TUNING, which demonstrates the feasibility of aligning models to quote. The core of QUOTE-TUNING is a fast membership inference function that efficiently verifies text against trusted corpora. We leverage this tool to design a reward function to quantify quotes in model responses and curate datasets for preference learning. Experiments show that QUOTE-TUNING significantly increases verbatim quotes from high-quality documents by up to 130% relative to base models while maintaining response quality. QUOTE-TUNING is applicable in different tasks, generalizes to out-of-domain data and diverse model families, 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]

> Trust, but verify.
>
> *Russian Proverb*

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, *i.a.*). This poses significant risks to downstream users due to the difficulty of fact-checking 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*, the ability to ground 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 approaches provide no guarantee of relevance or usefulness. Models generations can be unfaithful to the retrieved documents in the context (Shi et al., 2023b), generative search engines often produce citations that are irrelevant or inaccurate (Liu et al., 2023), and explanations alone do not lead to verifiability (Fok and Weld, 2024).

We overcome the windingness of previous approaches through a *verifiable-by-design* approach: generating **verbatim quotes** from high-quality sources such as Wikipedia. By determining verbatim quotes from large-scale and high-quality corpora with efficient membership testing tools (Marone and Van Durme, 2023), those generations provide a natural method for attributing and verifying the correctness of generated claims.

LLMs' potential 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 they have memorized a wide range of content from the pre-training stage (Carlini et al., 2020, 2023; Biderman et al.,

---

[1] `github.com/JHU-CLSP/verifiable-by-design`

♡ Equal advising.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3748–3768
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

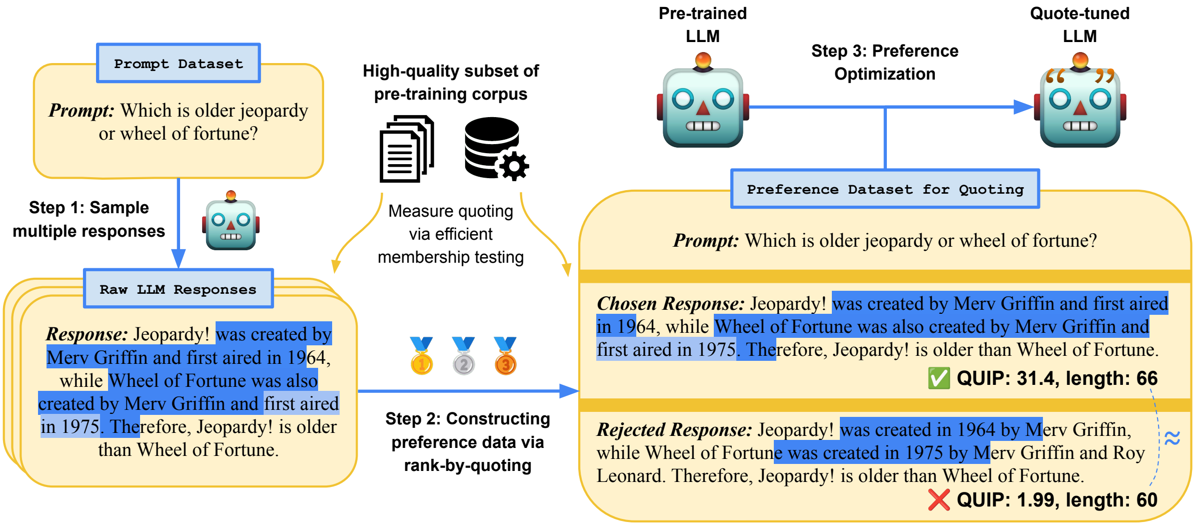
Prompt Dataset
Prompt: Which is older jeopardy or wheel of fortune?
Step 1: Sample multiple responses
Raw LLM Responses
Response: Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.
High-quality subset of pre-training corpus
Measure quoting via efficient membership testing
Step 2: Constructing preference data via rank-by-quoting
Pre-trained LLM
Step 3: Preference Optimization
Quote-tuned LLM
Preference Dataset for Quoting
Prompt: Which is older jeopardy or wheel of fortune?
Chosen Response: Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.
✅ QUIP: 31.4, length: 66
≈
Rejected Response: Jeopardy! was created in 1964 by Merv Griffin, while Wheel of Fortune was created in 1975 by Merv Griffin and Roy Leonard. Therefore, Jeopardy! is older than Wheel of Fortune.
❌ QUIP: 1.99, length: 60

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.

2023; Hartmann et al., 2023). Such analyses focus on *covert* memorization and use adversarial prompts to extract the memorized contents (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 (not just specialized or adversarial ones) on realistic tasks that require long-form generation.

We show this is indeed feasible with QUOTE-TUNING, our proposed method that aligns LLMs to quote through preference optimization and automatic feedback, without the need for human annotation. QUOTE-TUNING first generates responses from a pre-trained LLM, and then synthesizes a preference dataset for quoting by ranking responses by how much they quote from a desired high-quality corpus, e.g., Wikipedia. Finally, QUOTE-TUNING aligns the model to quote from trusted sources 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 up to 130% relative to base models while maintaining or improving downstream performance (§4). Moreover, QUOTE-TUNING generalizes to other domains and diverse model families, 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. The quoted responses are *verifiable-by-design* by inducing better verifiability without the need for human annotation and external knowledge bases (only leveraging parametric knowledge). 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 We define a text string x as *quoted* from a corpus C if a verbatim copy of x is contained in C.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 Quoted Information Precision Score (QUIP-Score) metric proposed by Weller et al. (2024):

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

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

---

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

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 len2gth $len(\cdot)$, length hyperparameter δ_{length}
Output: Quoting-aligned LLM policy π_θ

```
//Sample Responses + Synthesizing Data
$\mathcal{D} \leftarrow \emptyset$
for $i = 1, \dots, N$ do
    $y_1, \dots, y_T \sim \pi_{\text{ref}}(\cdot|x^{(i)})$    ▷ Sample responses from LLM policy
    $\tilde{y}_1, \dots, \tilde{y}_T \leftarrow \texttt{sort}(y_1, \dots, y_T; \lambda y. - \text{QUIP}_C(y))$    ▷ Sort by decreasing QUIP order
    for $w \in \{1, \dots, T-1\}, l \in \{w+1, \dots, T\}$ 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 $\pi_\theta$ on $\mathcal{D}$ using $\mathcal{L}_{\text{DPO}}$.
return $\pi_\theta$
```

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 $(y_1, y_2) \sim \pi_{\text{ef}}(\cdot|x)$ is sampled from the pre-trained model. The response pair is subsequently evaluated for preference (human annotators or automated metrics), with the more favored response labeled as y_w and the less preferred one as 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:

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

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

The design of QUOTE-TUNING is inspired by the observation that preference datasets can be used to elicit desired behaviors in LMs using preference alignment frameworks (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022b). The prior work, for example, has used this approach to address 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 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 $y_1, \dots, y_T \sim \pi_{\text{ref}}(\cdot|x^{(i)})$ 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 $y_1, \dots, y_T$ 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 direct 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 con-

Question: Who won the most MVP awards in the NBA?	**QUIP**
Reference: Every player who has won this award and has been eligible for the Naismith Memorial Basketball Hall of Fame has been inducted. Kareem Abdul-Jabbar won the award a record six times. Both Bill Russell and Michael Jordan won the award five times, while Wilt Chamberlain and LeBron James won the award four times. Russell and James are the only players to have won the award four times in five seasons. ...	94.1
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.**

straints, a single pair (y_w, y_l) with the highest average QUIP-Score among the two responses will be selected.[3] This ensures the dispreferred response still maintains relatively high quoting. 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.

The reason why we employ model self-generation as the preferred response, instead of simply using verbatim quotes (e.g., spans of Wikipedia text that contains the gold answer) is twofold: (1) Using self-generated responses keeps the synthesized preference data on-policy, which is crucial for preference optimization as shown in recent work (Tajwar et al., 2024). (2) Because quotes are a subset of the model response, they are constrained to be highly contextually relevant to both the query and the surrounding response context. This relevance would be harder to achieve with standalone quotes.

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] We provide an extended discussion of verifiability in §6.

We emphasize that the quality of the quoting corpus C is important: C needs to be carefully selected such that it contains high-quality, low-risk text, such as our choice of Wikipedia. We show in §5.1 that quoting from truthful corpus increase model truthfulness, and discuss implications of the selection of C on privacy and copyright in §7.

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 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.

---

[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.

Setting	Method	*Quoting* QUIP↑	*Adequacy* Rouge-L↑	 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 rerank	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 rerank	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.

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), and maintain downstream performance (§5.2).

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. To find settings relevant to QUOTE-TUNING, we select datasets that induce *long-form response* to measure quoting and also allow us to *verify the answers from trusted sources* such as Wikipedia. Accordingly, 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. 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 reranking baseline, where we sample 32 responses from the pre-trained model π_{ref} and rerank 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).

Results We run experiments with hyperparameters detailed in §B. We first employ LLAMA2-7B-CHAT (Touvron et al., 2023) as the pre-trained model π_{ref}. After DPO, the reward accuracy on a held-out evaluation set is 86.3%, indicating that the model learns quoting preference reasonably

---

[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.

	Quoting	*Adequacy*		*Fluency*	
Method	QUIP↑	Rouge-L↑	BARTSc↑	PPL↓	Length
LLAMA3.1-8B-INST	33.0	22.5	-3.97	5.03	136.7
+QUOTE-TUNING	**43.0**	**25.7**	**-3.82**	**3.13**	118.3
GEMMA2-9B-IT	30.0	21.4	-4.03	7.60	60.1
+QUOTE-TUNING	**44.9**	**24.4**	**-3.97**	**5.76**	57.4
STARLING-7B-BETA	33.8	22.7	-3.83	2.81	156.1
+QUOTE-TUNING	**44.4**	**23.8**	**-3.82**	**2.70**	150.8

Table 3: QUOTE-TUNING on diverse model family consistently improves quoting, adequacy, and fluency over their respective base models on NQ.

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 rerank baseline that is more computationally costly.** In particular, QUOTE-TUNING enables a significant 56.2% (34.9 → 54.5) quoting improvement relative to the un-tuned LLAMA2-7B-CHAT model. QUOTE-TUNING also slightly improves answer adequacy and fluency. Because QUOTE-TUNING significantly increases quoting from Wikipedia, which contains high-quality text thoroughly curated by human editors, the responses from quote-tuned models benefit from the high-quality nature on Wikipedia.[8] 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 LLAMA2-7B-CHAT quote-tuned on NQ for evaluation on ELI5. QUOTE-TUNING still outperforms all baselines in quoting, while maintaining adequacy and improving fluency compared to the original model. Table 2 (lower) shows that **QUOTE-TUNING generalizes quoting to out-of-domain prompts.**

Finally, we apply QUOTE-TUNING on LLAMA3.1 (Dubey et al., 2024), GEMMA2 (Team et al., 2024), and STARLING (Zhu et al., 2024) models, and find **QUOTE-TUNING consistently improve quoting, adequacy, and fluency across diverse model families** (Table 3).

4.2 Improving Quoting in Open-Ended Text Completion

Task Construction We now study whether QUOTE-TUNING can be applied to open-ended text completion, where we measure quoting on the candidate LLM's open-ended continuation of test prompts. 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 as the reference continuation, truncated to a maximum of 128 tokens.

Baselines and Metrics We employ the pretrained LLM policy π_{ref} and Best-of-N QUIP reranking baselines following the LFQA setting (§4.1). Instead of according-to prompting, we finetune π_{ref} on reference continuations of the train set as another baseline since π_{ref} 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 pretrained model π_{ref}, and measure perplexity with the MISTRAL-7B model instead to prevent self-evaluation bias (He et al., 2023).[9]

	Quoting	*Adequacy*		*Fluency*
Method	QUIP↑	R-L↑	BSc↑	PPL↓
LLAMA2-7B	25.7	21.8	-4.95	9.03
+Fine-tuning	29.1	21.9	**-4.90**	9.58
+Bo32 QUIP rerank	47.9	**23.8**	-4.95	6.63
+QUOTE-TUNING	**59.2**	23.1	-5.02	**5.39**

Table 4: On the open-ended text completion setting, QUOTE-TUNING significantly improves quoting and fluency while maintaining adequacy.

---

[8] Similar findings are also found in Li et al. (2024), which report improved perplexity by conducting speculative decoding to copy spans from a datastore.

[9] Moreover, MISTRAL-7B is shown to be a stronger model (Jiang et al., 2023) compared to LLAMA2-7B.

Method	Generation			Multiple Choice	
	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 5: 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.

Results After conducting QUOTE-TUNING with hypermaraters detailed in §B, the reward accuracy on a held-out evaluation set is 84.0%. As shown in Table 4, **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 → 59.2, a 130.4% relative increase), and outperforms the strong QUIP reranking baseline. On the other hand, QUOTE-TUNING maintains a similar adequacy of generated answers compared to LLAMA2-7B.

Interestingly, Table 4 shows that simply reranking 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 sources. 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 (Brown et al., 2020) on human annotations as truthfulness and informativeness judges. We defer further details to Appendix C.

As shown in Table 5, 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 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 8. Interestingly, **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). The increase of truthfulness likely attributes to the fact that Wikipedia is a relatively high-quality and reliable source, and quoting more from such reliable sources leads to more truthful responses. We thus posit aligning LLMs to reliable sources is a promising approach to increase their truthfulness.

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).[10]

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

As shown in Table 6, QUOTE-TUNING only leads to very small degradations (less than two points for all tested benchmarks), while significantly improving quoting. We therefore find QUOTE-TUNING a significantly better trade-off for verifiability with a small cost of general capability.

---

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

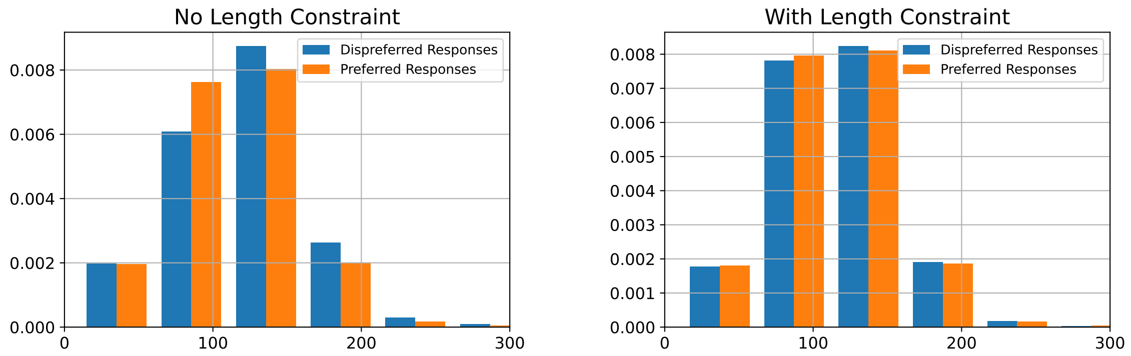

Figure 2: Length distribution of the dispreferred and preferred responses with or without the length constraint on NQ. **Left**: No length constraint. **Right**: added length constraint with $\delta_{\text{length}} = 0.1$. Adding length constraints properly regulates length distribution of responses.

		Quoting	*Adequacy*		*Fluency*	
Setting	*Method*	QUIP↑	Rouge-L↑	BARTSc↑	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 7: 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.

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 7. 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.

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. Khalifa et al. (2024) introduce intrinsic source citation to enable more faithful attribution to parametric knowledge, on a synthetic setting. Quote-Tuning is an approach complementary to citations that enhances verifiability. Hennigen et al. (2024) finds that explicit symbolic references to structured conditioning data, such as JSON tables, lead to faster human verification, further motivating our approach of increasing verifiability through verbatim quotes.

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 nontrivial and there is no guarantee that generated text is completely faithful to these documents. On the other hand, our framework for quoting, 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.

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 synthesizes 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., 2020, 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. We defer further related work to §A.

7 Discussions and Future Work

In §4, we provide rich empirical evidence that QUOTE-TUNING can significantly promote parametric quoting across diverse tasks, domains, and model families. Since quoted texts are autoregressively generated as spans of the model responses, and our evaluation quantitatively shows that these responses are of high adequacy and fluency (Tables 2, 3, 4), generated quotes are therefore constrained to be highly adequate, fluent, and context-relevant text. These findings demonstrate that it is not only possible but easily feasible to leverage parametric knowledge to generate more verifiable outputs. We thus argue **current LLMs have abundant underutilized potential in improving their own verifiability**, and call for future work that develop more attributable, verifiable approaches through our proposed "trivializing verification through quoting" framework.

As an early exploration of the quoting framework, this work focuses on investigating the ***feasibility*** of unlocking parametric quoting through fine-tuning, and the ***generalizability*** of enhanced quoting across different domains, tasks, and model families. We have provided positive empirical evidence supporting both aspects. Therefore, our setup focuses on measuring and improving the overall rate of quoting, as evaluated by QUIP-Score, leaving room for future work to enhance quoting reward signals by considering other desiderata of quoted texts, such as quote completeness and usefulness under the current context. We provide an extended discussion of limitations in the subsequent section.

Quoting has implications on privacy, security, and copyright. We focus on enhancing quoting from Wikipedia, a high-quality and low-risk corpus, for verifiability. As pre-training data contains diverse mixtures of data with varying risk levels (Elazar et al., 2024; Longpre et al., 2024), we argue that the grounding corpus for QUOTE-TUNING must be carefully selected and limited to trusted, public sources such as Wikipedia to prevent privacy violations or copyright infringement.

In conclusion, our approach presents a promising direction for leveraging the parametric knowledge of LLMs to facilitate easier verification of model generation and improve the calibration of 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 (Mishra et al., 2022; 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) 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.

Ethical Considerations

We have shown that increasing verbatim quotes from pre-training data through QUOTE-TUNING is a promising approach for enhancing verifiability. However, the ability to quote from pre-training data have broad ethical implications. While enhanced quoting increase users' ability to attribute generation back to their original sources, this could be a *double-edged sword* regarding privacy protection: adversarial users might utilize similar methods to extract sensitive information contained in pre-training data. On the other hand, if this is feasible, it may create a path for auditing pre-training data.

Furthermore, the legal implications of quoting from pre-training data must be carefully managed. It is necessary to ensure that copyrighted material is handled appropriately, with proper attribution, and that the use of such data adheres to intellectual property laws. By addressing ethical and legal considerations, we envision QUOTE-TUNING as a responsible tools for enhancing verifiability and calibrating human-machine trust.

Acknowledgements

This work is supported by ONR grant (N00014-24-1-2089) and an Amazon Faculty Research Award. We sincerely thank Yung-Sung Chuang, Andrew Wang, Jeffrey Cheng, Mahsa Yarmohammadi, Dongwei Jiang, and the broader JHU CLSP community for discussions and inspiration.

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).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* (NeurIPS).

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 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*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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 *Advances in Neural Information Processing Systems* (NeurIPS).

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data?

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.

Raymond Fok and Daniel S. Weld. 2024. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *AI Magazine*, 45(3):317–332.

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 *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

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).

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).

Lucas Torroba Hennigen, Zejiang Shen, Aniruddha Nrusimha, Bernhard Gapp, David Sontag, and Yoon Kim. 2024. Towards verifiable text generation with symbolic references. In *First Conference on Language Modeling*.

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.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. Source-aware training enables knowledge attribution in language models. In *First Conference on Language Modeling*.

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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems* (NeurIPS).

Minghan Li, Xilun Chen, Ari Holtzman, Beidi Chen, Jimmy Lin, Wen tau Yih, and Xi Victoria Lin. 2024. Nearest neighbor speculative decoding for llm generation and attribution.

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. In *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2303.03919*.

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.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Annual Meeting of the Association for Computational Linguistics* (ACL).

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 context. In *Annual Meeting of the Association for Computational Linguistics* (ACL) - *Findings*.

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.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. 2024. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Forty-first International Conference on Machine Learning*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda

Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

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, et al. 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 *Advances in Neural Information Processing Systems* (NeurIPS).

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 *Annual Meeting of the Association for Computational Linguistics* (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.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7b: Improving helpfulness and harmlessness with RLAIF. In *First Conference on Language Modeling*.

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*.

Supplemental Material

A Additional Related Work

Reward Modeling and 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 toward generating quotes. 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 Training details

We employ hyperparameters $T = 32, \delta_{\text{quip}} = \delta_{\text{length}} = 0.1$ for QUOTE-TUNING on NQ. We set DPO $\beta = 0.05$, which is tuned on a small dev set. We do not tune β for LLAMA3.1 (Dubey et al., 2024), GEMMA2 (Team et al., 2024), and STARLING (Zhu et al., 2024) models, and use the default recommended value $\beta = 0.1$. For open-ended text completion, 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 $|\mathcal{D}| = 19989$.

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`[11] 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.

---

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

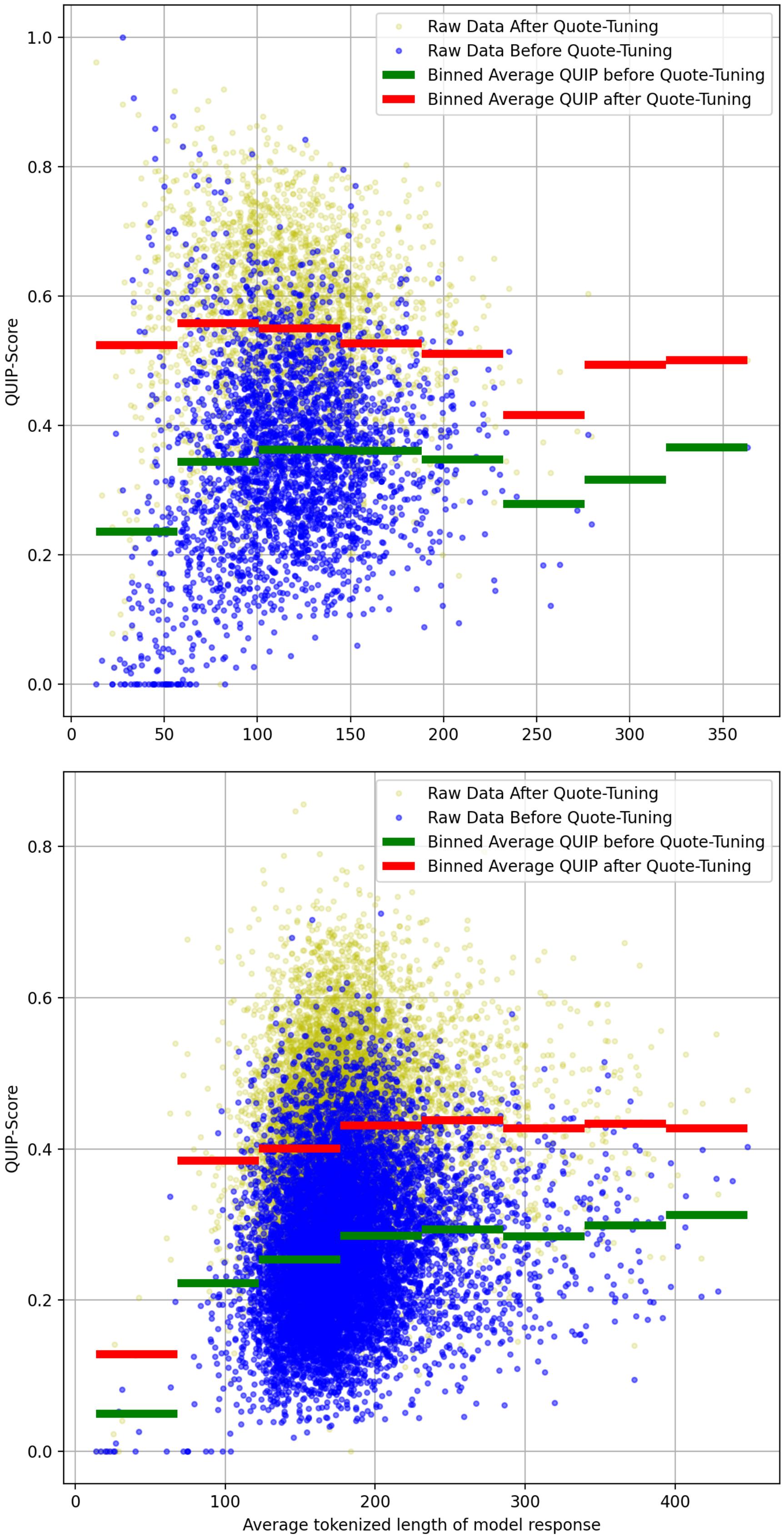

Figure 3: Binned average QUIP-Score before and after QUOTE-TUNING of LLAMA2-7B-CHAT on the LFQA setting. **Top**: NQ; **Bottom**: ELI5. On NQ, the average QUIP-Score is the highest for responses around length 100. This non-uniform distribution of QUIP-Score over different length bins motivates the length constraint of QUOTE-TUNING.

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 8: 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.

D Additional Examples

Additional examples that contrast model responses before and after Quote-Tuning on NQ are available in Table 9, 10, and 11.

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 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: 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 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.

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 11: 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.