Knowledge Acquisition through Continued Pretraining is Difficult: A Case Study on r/AskHistorians

Jan Vincent Hoffbauer[1,2], **Sylwester Sawicki**[1,2], **Marc Lenard Ulrich**[3], **Tolga Buz**[2], **Konstantin Dobler**[2], **Moritz Schneider**[2], **Gerard de Melo**[2]
[1]SAP, [2]Hasso Plattner Institute / University of Potsdam, [3]University of Potsdam
jan.vincent.hoffbauer@sap.com, sawicki@uni-potsdam.de, marc.ulrich@uni-potsdam.de, tolga.buz@hpi.de, konstantin.dobler@hpi.de, moritz.schneider@guest.hpi.de, gdm@demelo.org

Abstract

Powerful LLMs such as ChatGPT master a wide array of tasks, but have notable limitations in domain-specific areas, especially when prompted to recite facts. This is of particular importance for knowledge workers who are increasingly adopting LLM-based tools. While there are various techniques that can help ingest knowledge into LLMs, such as instruction tuning and alignment, most have disadvantages. We examine the impact of prominent training techniques on LLMs' knowledge accuracy using a knowledge-dense dataset that we curate from `r/AskHistorians`, a rich source of historical knowledge. We evaluate the impact of different model sizes from 1.3B to 7B parameters and other factors such as LoRA adapters, quantization, overfitting, and the inclusion of Reddit data in pretraining. In addition, we measure linguistic metrics and human and LLM-based preferences. Our results suggest that pretraining and model size have a much stronger effect on knowledge accuracy than continued pretraining – except in cases of overfitting to the tested knowledge. Fine-tuning on our Reddit dataset introduces less complex, but slightly more toxic language. Our study explores the challenges of injecting domain-specific datasets into LLMs and has implications for practitioners, e.g., when LLMs are to be fine-tuned with company-specific datasets.

1 Introduction

Large Language Models (LLMs) have evolved far beyond mere natural language processing tools and are now widely used by knowledge workers seeking answers to knowledge-related questions. However, while these models incorporate a vast set of world knowledge due to their pretraining on trillions of tokens, they still often lack niche domain-specific knowledge, which can manifest in hallucinations or unspecific responses (Huang et al., 2023). In addition, an LLM's knowledge may need to be updated from time to time (Ovadia et al., 2023). These issues are especially critical for models deployed in professional settings to assist knowledge workers in performing knowledge-intensive tasks in particular domains.

There are various ways how one can try to ingest knowledge into LLMs, but each has its disadvantages: Unsupervised pretraining enables LLMs to learn immense amounts of knowledge, but without long-tail details (Kandpal et al., 2023). Supervised fine-tuning (SFT, or instruction tuning) can be used to expose the model to new knowledge when learning a new task, but niche facts do not seem to "stick" (Kandpal et al., 2023) and fine-tuning without the original training data can lead to *catastrophic forgetting* (Kirkpatrick et al., 2016; Kemker et al., 2018); Alignment using techniques such as Reinforcement Learning from Human Feedback (RLHF; e.g., as used in Touvron et al. 2023) or Direct Preference Optimization (DPO; Rafailov et al. 2023) can greatly improve the quality of generated texts and introduce safety mechanisms, but requires costly training due to very small learning rates for highly nuanced model adjustment. Retrieval Augmented Generation (RAG; Lewis et al. 2020) appears to be a promising workaround that avoids fine-tuning, but requires a more complex architectural setup and a greater number of prompts and tokens to operate, causing higher usage costs, while the result quality is highly dependent on the information stored in its database.

We investigate this area of research using a large dataset from `r/AskHistorians`, a strictly moderated online community on Reddit, that contains questions and long-form answers on diverse historical topics, often discussing very specific historical facts. As Reddit users can up- or downvote posts and comments, the dataset provides inherent human feedback that can be leveraged for aligning LLMs with DPO. Given that social media datasets often pose challenges with regard to issues such as

Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024), pages 96–108
August 16, 2024 ©2024 Association for Computational Linguistics

data quality and toxicity, we exercise special care to curate a high-quality dataset. Furthermore, we assess the impact of different model sizes (ranging from 7B to 1.3B parameters), the usage of LoRA adapters and quantization, and overfitting to the knowledge dataset. We present an approach to measure the knowledge accuracy of the models by manually creating a Knowledge Filling dataset. In addition, we conduct human and LLM-based evaluation, and consider more traditional NLP metrics such as text complexity, reading time, and toxicity. The main purpose of this work is to demonstrate how one can proceed when attempting to inject specific knowledge into LLMs and evaluate its success. Our code is publicly available on Github[1].

2 Background

2.1 Models, Datasets & Related Work

There is a variety of capable LLMs available, including proprietary solutions such as ChatGPT and Google's Gemini (Gemini Team, 2024), and open-source alternatives such as Meta's Llama-2 (Touvron et al., 2023) and Mistral's various models, e.g., `Mistral-7B-v0.1` (Jiang et al., 2023). In this work, we utilize leading open-source LLMs that fulfill our conditions along two dimensions: different model sizes that are sufficiently small to run on consumer-grade hardware with 1.3B (`pythia-1.4B`, Biderman et al. 2023) to 7B parameters (`Mistral-7B-v0.1`, Jiang et al. 2023; `zephyr-7B-beta`, Tunstall et al. 2023) and are either pretrained with Reddit data (`pythia-1.4B`; Biderman et al. 2023) or not (`phi-1.5`; Li et al. 2023). It should be noted that (1) `phi-1.5` is trained on textbook-style synthetic data exclusively, and (2) the training data for `Mistral-7B-v0.1` is not disclosed, but one can assume that it has seen various types of online data, including social media data from Reddit, based on its generated texts.

Reddit is a social media platform containing communities known as subreddits, where individuals share and discuss content on a wide range of topics. Users can up- or downvote posts and comments to indicate their preferences. This provides an inherent quality rating of posts that can be leveraged for fine-tuning, aligning, and evaluating LLMs. In recent years, social media datasets have become essential for training and evaluating LLMs. For example, Fan et al. (2019) present a large corpus for long-form question answering centered on the subreddit `r/explainlikeimfive` (ELI5), and Buz et al. (2024) utilize `r/Showerthoughts` to train LLMs for generating creative and witty texts that deceive human evaluators. Ayers et al. (2023) compare responses to patient questions written by physicians on the `r/AskDocs` subreddit to those generated by ChatGPT, finding that annotators prefer ChatGPT's responses in 79% of cases. Apart from work about Reddit communities, there are also very large internet datasets such as CommonCrawl (Common Crawl, 2024) and the Pile (Gao et al., 2020), which include social media data and have been used (in their entirety or after filtering) for pretraining a variety of LLMs, including GPT-3 (Brown et al., 2020). UltraChat (Ding et al., 2023) and UltraFeedback (Cui et al., 2023) are two noteworthy datasets, which have enabled the creation of `zephyr-7B-beta` from `Mistral-7B-v0.1` using SFT and DPO, respectively (Tunstall et al., 2023).

In summary, related research shows that social media datasets, specifically those from Reddit, can be valuable in the context of LLMs. However, there is no work yet that examines how domain-specific social media datasets can be curated to create knowledge datasets, nor how knowledge can be injected from such datasets into LLMs using different techniques.

2.2 Training

A full pipeline for training an LLM as a chatbot or question-answering system typically consists of the following steps, as outlined in Touvron et al. (2023): (1) Unsupervised pretraining on a large dataset (potentially trillions of tokens) to help the LLM identify common linguistic patterns, (2) SFT on a set of questions (or prompts) and best answers to teach the LLM specific tasks and ways to respond, and (3) alignment on a preference dataset (i.e., two answers of which one is rated as better than the other), e.g., using RLHF, to fine-tune the quality of the LLM's responses towards nuanced differences in human preference.

A common method for RLHF is Direct Preference Optimization (DPO; Rafailov et al. 2023), which avoids a reward model and instead utilizes preference scores directly, enabling a more efficient and stable model alignment.

[1]https://github.com/aiintelligentsystems/askhistorians-knowledge-filling

2.3 Knowledge Injection

As described above, there are various techniques for modifying LLMs and instilling knowledge, with each technique having its own advantages and disadvantages. Yu et al. (2020) distinguish between internal and external knowledge sources for LLMs:

Regarding internal knowledge, Kandpal et al. (2023) argue that unsupervised pretraining and SFT are good at making LLMs learn broad world knowledge and specific tasks, respectively, but fail at injecting specific facts and niche knowledge they consider *long-tail knowledge*. Other research indicates that fine-tuning on specific data can lead to catastrophic forgetting on previously learned tasks (Kirkpatrick et al., 2016; Kemker et al., 2018), while the concept of continual learning advocates approaches that aim to prevent this (Zhou et al., 2024; Scialom et al., 2022). In contrast, Liu et al. (2023a) present a model that is specifically fine-tuned on a dataset related to chip-design tasks – the authors show that a model specifically pretrained on a highly domain-specific dataset yields improved performance on related tasks. As an alternative, Jiang et al. (2024) propose pre-instruction tuning to inject knowledge before fine-tuning on documents, which seems to improve on this task, but is more difficult to implement correctly. Alignment techniques such as DPO (while more efficient than PPO; Schulman et al. 2017) are costly approaches that focus on nuanced alignment of LLMs using a very small learning rate. Furthermore, very recent work indicates that using LoRA adapters for training reduces the learning and forgetting effects (Biderman et al., 2024). In our experiments, we focus on internal LLM knowledge and investigate how strongly these techniques can affect an LLM's knowledge when trained and evaluated in the historical domain. We disregard more complex specialized techniques such as knowledge editing, which aims to modify a model's parameters (Wang et al., 2023) or its outputs through a smaller language model (Liu et al., 2024) or a steering vector (Rimsky et al., 2023), due to their complexity and lack of support in common libraries such as PyTorch.

While not the focus of this work, it is relevant to point out research on incorporating external knowledge – Retrieval-Augmented Generation (Lewis et al., 2020) is often presented in related work as a better alternative for knowledge injection (Ovadia et al., 2023). However, RAG requires a more complex architectural setup including a suitable database with a retrieval model that is connected to the main LLM and yields relevant excerpts of text fed to the latter via prompting, increasing the amount of input tokens. This increases the usage cost and introduces various risks – e.g., difficulties of inserting new information, or retrieval of unsuitable pieces of information. We consider methods incorporating extrinsic data sources at runtime as beyond the scope of this work.

2.4 Evaluation

Evaluating LLM-generated texts, especially in long form, in a scalable and reliable way remains an ongoing challenge at the time of writing. Human judgment is still the gold standard when it comes to assessing the generation quality of dialogue-tuned or question-answering models (Touvron et al., 2023)

A key idea when evaluating LLMs is to compare the output of a fine-tuned LLM to another LLM that is considered state-of-the-art or a valid baseline, e.g., Touvron et al. (2023) compare their results to GPT-4 (OpenAI, 2023) with human annotators. The LLM-as-a-judge approach aims to automate this by instead invoking high-quality LLMs such as GPT-4 (Zheng et al., 2023; Liu et al., 2023b) to perform the assessment – while there seems to be decent correlation with human preference, these approaches are subject to various biases, e.g., the judge LLMs preferring longer responses or those that are similar to what they are trained to respond.

In addition to the evaluation of text quality, various descriptive metrics are commonly used to measure simpler properties of texts, e.g., toxicity (Hartvigsen et al., 2022), text complexity, and reading time (Ward et al., 2023).

3 Methodology

An overview of our technical setup is shown in Figure 1: We process and curate a preference dataset from the raw `r/AskHistorians` data, utilize it for model training using SFT (`phi-1.5` and `zephyr-7B-beta`) and DPO (`zephyr-7B-beta`) and evaluate using different approaches, including `GPT-4-turbo` as LLM judge, and `Mistral-7B-v0.1` and `pythia-1.4B` for baseline comparisons.

3.1 Dataset

We retrieve our dataset from the Pushshift API (Baumgartner et al., 2020), which was freely accessible until mid-2023, when the Reddit API terms

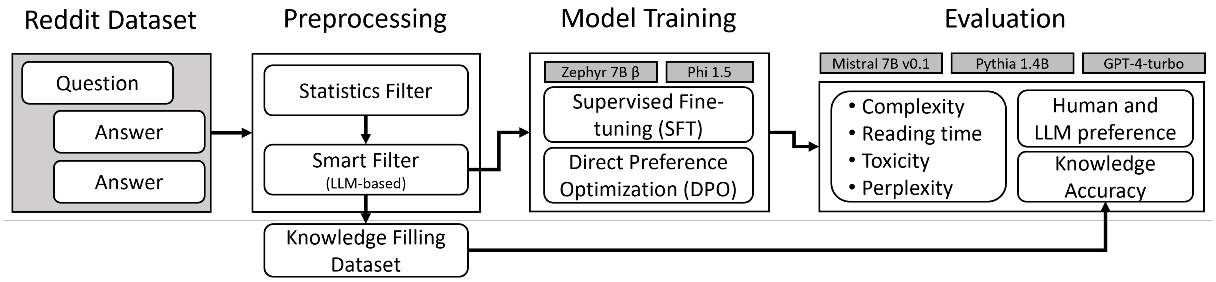

Figure 1: An overview of our experimental setup

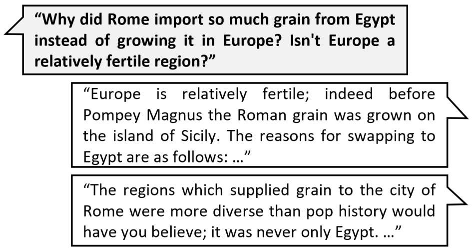

Figure 2: `r/AskHistorians` dataset example (accessible via `reddit.com/1cuvs50`)

were changed due to growing demand for training machine learning models. Therefore, our dataset ranges from the creation of the subreddit in August 2011 until the end of 2022 and contains approximately 116,452 questions and 384,491 answers. Figure 2 shows an example of the discussions on `r/AskHistorians` – questions are often about specific details that require in-depth historical knowledge to respond. The community is strictly moderated to ensure serious and factually correct discussions, resulting in a relatively small, but high-quality dataset.

To further enhance the data quality, we eliminate posts that (1) do not contain questions (e.g., recommendations or monthly reading lists), (2) are shorter than 55 characters, (3) have an upvote score lower than 4 (to focus on popular posts), or (4) have fewer than two top-level comments as answers (which we require to build a preference dataset).[2]

In a final step, we apply the baseline `zephyr-7B-beta` model as a smart filter to assign a quality rating to each question – for this purpose, we use a few-shot setting that explains criteria for good questions based on the subreddit's community guidelines (further details in the Appendix). We manually evaluate the smart filter's correlation with human judgement based on 100 randomly sampled questions and identify an agreement rate of 70%, which we deem sufficient. This yields a final dataset of 34,631 questions labelled as "good", and 100,429 answers.

3.2 The `r/AskHistorians` Knowledge Filling Task

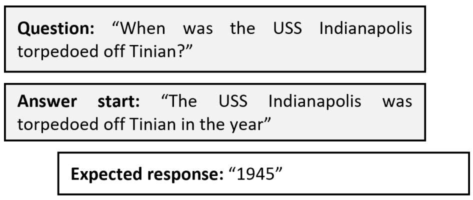

Figure 3: Knowledge Filling dataset sample

Accurate evaluation of an LLM's factuality in long-form answers is challenging and there are currently no existing frameworks that we could draw upon for this purpose. In order to facilitate and enable this evaluation, we create a Knowledge Filling dataset[3] inspired by the cloze procedure (Taylor, 1953): We rephrase facts from the dataset's discussions to ask about a specific fact and formulate an "answer start" prompt that is only missing the key fact at its end. The LLM is then prompted to only generate the missing fact using a very limited number of tokens (see Figure 3).

We select 100 random samples from the training dataset to ensure that our models have seen the data during SFT and DPO, as it is our goal to measure whether the further training helps in injecting knowledge. The resulting question–answer prompts are relatively short with an average length of 100 characters, while the average expected response length is 9.9 characters – this enables inexpensive evaluation.

[2] Lower-level comments are often posted in response to the first level comment instead of the question, which disqualifies them for our purpose.

[3] `https://huggingface.co/datasets/aiintelligentsystems/askhistorians-knowledge-filling`

It is important to note here that this procedure is critical to separate the evaluation of knowledge accuracy from linguistic style – a large number of currently popular evaluation frameworks such as MTBench (Zheng et al., 2023) and G-Eval (Liu et al., 2023b) inevitably evaluate linguistic style, as they prompt for preference or attempt to measure abstract linguistic properties. Likewise, the perplexity metric primarily measures how close a text is to a linguistic style rather than factual accuracy.

3.3 Model Training

Base models. For our experiments, we conduct SFT and DPO sequentially on `zephyr-7B-beta`, and SFT on `phi-1.5` (this model uses a different training procedure with custom code, therefore we do not perform DPO on it). These choices are motivated as follows: `zephyr-7B-beta` is a popular checkpoint[4] in the commonly used 7 billion parameter range for base models. We further use `phi-1.5` in an additional experiment because it has been trained on exclusively synthetic data, which does not include `r/AskHistorians`.

For evaluating the Knowledge Filling dataset, we additionally use `mistralai/Mistral-7B-v0.1` and `EleutherAI/pythia-1.4B` as baselines. While we know that `pythia-1.4B` included Reddit during pretraining, we assume that `Mistral-7B-v0.1` (and therefore also `zephyr-7B-beta`) has seen Reddit data as well, based on some of the texts it generates that resemble the structure and patterns seen in Reddit metadata (e.g., mentioning of a subreddit with "r/[subreddit name]").

After preliminary experiments with RLHF and PPO (which are highly dependent on the quality of the reward model), we choose Direct Preference Optimization due to its simple implementation and higher robustness.

Usage of LoRA and quantization. Low-Rank Adaptation (LoRA) reduces memory requirements by approximating the *update weight vector* during training (Hu et al., 2021). LoRA fine-tuning is a widely used method, which we employ to efficiently fine-tune the `zephyr-7B-beta` model, which would otherwise not fit into our GPU's memory during training. However, as the weight updates through fine-tuning are low-rank, this bears the risk of inhibiting knowledge ingestion. To mitigate this, we fine-tune `zephyr-7B-beta` using LoRA but utilize full-weight fine-tuning for `phi-1.5`. Quantization further reduces the memory footprint by using reduced precision for the parameters – we apply this to the `zephyr-7B-beta` and `Mistral-7B-v0.1` models by using `bfloat16`.

Experimental setup. We use the HuggingFace `transformers` Trainer for SFT and DPO, and conduct experiments either on Nvidia A6000 48GB or RTX 3090 Ti 24 GB GPUs depending on availability, while ensuring identical hyperparameters across both systems. We run supervised fine-tuning for a total of 3 epochs and DPO for a total of 18 epochs, selecting the best checkpoint according to the highest reward accuracies on the evaluation dataset. Training hyperparameters, detailed in the Appendix Table 5, were determined based on the HuggingFace Alignment Handbook (Tunstall et al., 2024).

To prevent `zephyr-7B-beta` from forgetting its generation qualities, we include data from the model's original fine-tuning during our SFT and DPO, following the continual learning process (Scialom et al., 2022). We randomly select samples from UltraChat (Ding et al., 2023) for SFT and UltraFeedback (Cui et al., 2023) for DPO so that roughly two percent of our training data is drawn from the respective original dataset.

3.4 Evaluation

Knowledge accuracy. Our main evaluation task is the `r/AskHistorians` Knowledge Filling task using our manually created dataset. As described above, this task was created to specifically evaluate knowledge ingestion through fine-tuning without being confounded by adaptation to the new domain's linguistic style. We determine an answer to be correct if the ground-truth is a sub-string of the generated answer and report the accuracy over the entire dataset. As in some cases, there can be multiple versions to write a response (e.g., "World War II" and "WW2"), we verify all results manually. For future work, we recommend compiling lists of possible answers for such cases to reduce the manual effort.

Stylistic adaptation and general quality. In addition, we measure the stylistic adaptation of the models as well their general quality. For this purpose, we utilize a set of NLP metrics to report (1) the perplexity of the models trained with our `r/AskHistorians` corpus as an indicator for how

[4] More than 300,000 downloads on the Huggingface Hub in May 2024.

Model	#params	Pretrained on Reddit	LoRA	Accuracy ↑ %
`Mistral-7B-v0.1` (no training)	7B	✓		32
`zephyr-7B-beta` (no training)	7B	✓		31
`zephyr-7B-beta` + `r/AskHistorians` SFT	7B	✓	✓	29
`zephyr-7B-beta` + `r/AskHistorians` SFT + DPO	7B	✓	✓	28
`zephyr-7B-beta` + `r/AskHistorians` Subset-Overfit SFT	7B	✓	✓	49
`phi-1.5` (no training)	1.3B			8
`phi-1.5` + `r/AskHistorians` SFT				9
`pythia-1.4B` (no training)	1.4B	✓		13

Table 1: Accuracy on the `r/AskHistorians` Knowledge Filling task using our manually created dataset. *Pretrained on Reddit* indicates whether the model has seen Reddit data during pretraining and the *LoRA* column indicates whether LoRA was used for resource-efficient fine-tuning of the model.

well the model replicates the community's linguistic style, (2) text complexity and reading time measured by the textstat package (Ward et al., 2023) to compare linguistic complexity, and (3) the toxicity using the HateBERT classifier trained on the ToxiGen dataset (Hartvigsen et al., 2022).

In addition, (4) we conduct a pairwise comparison study between model variants with (a) human and (b) LLM-as-a-judge evaluation to measure preference between a set of two answers per question (baseline `zephyr-7B-beta` versus fine-tuned `zephyr-7B-beta`). The human evaluation is conducted in a blind, randomized setting for evaluators, using 100 randomly sampled question–answer–answer tuples with three different human annotators. The LLM-based evaluation follows the setting proposed by Zheng et al. (2023) and uses GPT-4-turbo, which the authors commend for its efficacy in mitigating order or length bias. The prompt for this evaluation is available in Appendix C. Annotators are instructed with the same information, but in addition asked to consider the factual correctness, linguistic fluency, and accuracy of answers when indicating their preference. Inter-rater agreement is measured among humans and between humans and GPT-4-turbo and reported in the results in Section 4.2.

4 Results

4.1 Knowledge Accuracy

General observations Our results show that while the continued pretraining we conduct on LLMs is successful in instilling the writing style of `r/AskHistorians` into the models, we are not able to measure a notable uplift in the models' knowledge accuracy. On the contrary, SFT and DPO not only fail to yield any significant improvements in our `r/AskHistorians` knowledge-filling task, but instead, the knowledge accuracy value decreases slightly with each step of fine-tuning (see Table 1). For the baselines `Mistral-7B-v0.1` and `zephyr-7B-beta` we measure a knowledge accuracy of 32% and 31%, respectively. Fine-tuning `zephyr-7B-beta` on our `r/AskHistorians` dataset decreases the knowledge accuracy scores rather than increasing them (to 29% and 28% for SFT and SFT + DPO, respectively). This seems counter-intuitive, as it happens despite the fact that the evaluation questions are derived from facts that are contained in the training dataset and therefore seen by our fine-tuned model variants during training. This indicates that merely including facts during fine-tuning does not improve the knowledge accuracy of the model.

Limited benefits of overfitting. In an additional experiment, we test the upper bound on knowledge ingestion through fine-tuning by deliberately overfitting our model: We conduct SFT training of `zephyr-7B-beta` for 10 epochs on the subset of our filtered `r/AskHistorians` dataset that was used to generate the Knowledge Filling test set. This means that we do not train on the exact question–answer pairs that we evaluate on, but rather on the long-form question–answer pairs that were used to create the test question–answer pairs for knowledge filling and contain all relevant information. This experiment is listed as `r/AskHistorians` Subset-Overfit SFT in Table 1 and yields a higher knowledge accuracy of 49%. While this does show that knowledge can be ingested via fine-tuning eventually, the resulting accuracy after 10 epochs is still far from a desirable 90–100%. We note that we do not evaluate *surface form* knowledge completion, as our question–

Model	Text Complexity ↓ [student grade]	Reading time ↓ [s]	Toxicity ↓ [0-1]
zephyr-7B-beta	14.34 ± 2.41	**24.10 ± 10.37**	**0.10 ± 0.20**
zephyr-7B-beta + SFT + DPO	**13.35 ± 3.94**	38.75 ± 15.06	0.36 ± 0.22
Original Reddit Answer	11.48 ± 3.72	29.45 ± 28.57	0.20 ± 0.25

Table 3: Descriptive metrics results. The student grade refers to the grade in school such as "5th grade". ↑ denotes higher is better while ↓ denotes lower is better.

answer prompts in the test set are rephrased from the base training samples.

phi-1.5 and full-weight fine-tuning. Our experiments with zephyr-7B-beta were conducted using the widely used LoRA (Hu et al., 2021) technique, due to computational constraints. It needs to be considered that our negative results using zephyr-7B-beta could be due to the low-rank nature of LoRA impeding knowledge capture. Therefore, we conduct an additional experiment using full-weight fine-tuning with phi-1.5 as our base model. As phi-1.5 was not pretrained on any Reddit data, the model's knowledge accuracy score is lower at 8%. In comparison, pythia-1.4B as a similarly-sized model pretrained on Reddit has a knowledge accuracy score of 13%, which indicates a beneficial effect of this pretraining. However, conducting full-weight SFT on our filtered r/AskHistorians dataset still does not yield any significant knowledge accuracy improvements, with a resulting score of 9% (as opposed to 8% for the baseline). We conclude that fine-tuning fails to inject knowledge into LLMs (in contrast to a limited success of pretraining), and LoRA does not seem to be the root cause of this failure.

Model	Perplexity ↓
zephyr-7B-beta	13.12
zephyr-7B-beta + SFT	10.78
zephyr-7B-beta + SFT + DPO	**10.75**

Table 2: Perplexity of the baseline models and the models fine-tuned on r/AskHistorians on the respective training dataset.

4.2 Stylistic Adaptation and General Quality

Evaluations of LLM-generated long-form texts often consider the writing style and general quality among their criteria – as measuring a specific aspect such as knowledge accuracy is challenging to achieve. For a more comprehensive evaluation, we hence also analyze metrics related to these aspects in addition to the knowledge accuracy evaluation.

NLP metrics. Table 2 shows that the train perplexity of the fine-tuned model improves on the training dataset, indicating that, while knowledge ingestion failed as detailed in Section 4.1, the linguistic style of the dataset is learned. The other metrics listed in Table 3 indicate that after fine-tuning, zephyr-7B-beta generates text that takes longer to read (i.e., higher reading time) and has a higher toxicity score, but at the same time has a lower text complexity score (due to simpler sentences and vocabulary). It should be noted that the model training changes reading time and toxicity to a stronger extent than present in the original Reddit answers, as the fine-tuned model reaches significantly higher values. This suggests that the model could be "overshooting" during the fine-tuning process, possibly due to different properties of the r/AskHistorians dataset compared to the model's original training data.

Pairwise comparison. The pairwise comparison evaluation using human and LLM judges shows a clear pattern that the fine-tuned zephyr-7B-beta is rated worse than the baseline model (see Figure 4). Between the GPT-4 judge model and the human annotators' average, we observe a 63% agreement rate, while there is an average agreement rate of 50% between the three human annotators. This is interesting, as it shows that there is some ambiguity and subjectivity involved in this evaluation with the human annotators agreeing less with each other than their average does with GPT-4.

In addition, we use GPT-4 to judge between the original human written answer and baseline zephyr-7B-beta: the original answer is preferred in 54 cases, while Zephyr wins in 46 cases. This contrasts with the results of the fine-tuned model and shows that either the fine-tuning process or zephyr-7B-beta is not sufficient to capture the general quality of the original Reddit answers in the fine-tuned model.

Based on a qualitative analysis that we conduct

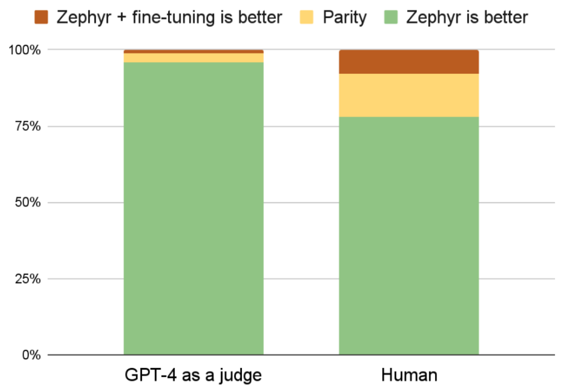

Figure 4: Model preferences as chosen by GPT-4 as a judge (Zheng et al., 2023) and human annotators. The bar charts display the rate of preferences for each model on multiple answers. This allows us to compare the generation quality of the two models.

manually on a random sample, we assume that this phenomenon is related to the original Reddit data lacking structure and a consistent style that current instruction-tuned models excel in. Individual authors have different writing preferences, making it harder for a model to learn a coherent style. This is a main difference to purposefully crafted datasets such as UltraFeedback (Cui et al., 2023). We observe that the fine-tuned model commonly generates subjective responses, starting with formulations such as *"I think ..."* or *"If I understand your question correctly ..."*, while the original Zephyr model directly answers the question and provides its arguments in enumerations. An example of this is given in the Appendix in Table 6.

5 Limitations

In some of our experiments, there is a risk of test-set contamination, as the underlying training data is not transparently declared for all tested models (e.g., `Mistral-7B-v0.1`) – it is possible that these models may have seen parts of our test dataset during their pretraining when using datasets like CommonCrawl (Common Crawl, 2024) and the Pile (Gao et al., 2020). We mitigate this by testing various model variants, including `phi-1.5`, which certainly has not seen Reddit data in pretraining. Also, the fact that Zephyr has likely seen Reddit data provides additional insights, as the decreased knowledge accuracy after fine-tuning and alignment potentially indicates reduced ability to generalise.

Our Knowledge Filling dataset for evaluation has a limited size, as its creation is highly time-consuming and cannot be outsourced or automated easily, due to requiring the annotator to understand the contents of the annotated text correctly. Despite meticulous curation, the dataset may inadvertently contain factually inaccurate statements. In addition, the setup as a cloze test leads to ambiguity: For instance, when prompted with *"Which world war ended in 1945?"*, the answer can either be *"WW2"* or *"World War 2"*; or specific dates may appear in different formats. This is mitigated in our study through detailed manual verification and would benefit from automation in future work. We employ additional evaluation techniques to provide diverse and more robust results.

The number of conducted experiments and trained model variants was limited by our access to shared computational resources, which is why we were not able to train and evaluate all possible combinations of model variants. Therefore, we have focused on providing a sufficient number of experiments to investigate the most interesting questions stated in the motivation of this paper.

6 Conclusion

In this work, we show that there are various challenges when trying to inject knowledge into a LLM by using common techniques like SFT and DPO, and present an approach for evaluating the knowledge accuracy and stylistic quality of trained LLMs from various perspectives. While state-of-the-art LLMs like `zephyr-7B-beta` already generate high quality texts out-of-the-box (due to their training on carefully curated data) and tend to deteriorate when fine-tuned on domain-specific texts, conducting further fine-tuning may still be necessary for practitioners in order to adjust the models to their specific use case, e.g., company datasets.

Our approach is intended to inspire practitioners to conduct comparable experiments and evaluate their specific LLMs' knowledge accuracy, as the techniques that we apply are generalizable and transferable to other domains that require niche or fact-related knowledge. For future work, there are various open questions, such as identifying more powerful ways to inject knowledge into LLMs and facilitating the creation of similar knowledge benchmarks at a larger scale.

References

John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. 2023. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA internal medicine*, 183(6):589–596.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. 2024. LoRA Learns Less and Forgets Less. *Preprint*, arXiv:2405.09673.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Preprint*, arXiv:2005.14165.

Tolga Buz, Benjamin Frost, Nikola Genchev, Moritz Schneider, Lucie-Aimée Kaffee, and Gerard de Melo. 2024. Investigating Wit, Creativity, and Detectability of Large Language Models in Domain-Specific Writing Style Adaptation of Reddit's Showerthoughts. *arXiv preprint arXiv:2405.01660*.

Common Crawl. 2024. Common Crawl - Open Repository of Web Crawl Data. Last accessed on 2024-02-29.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. UltraFeedback: Boosting Language Models with High-quality Feedback.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering.

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*.

Google Gemini Team. 2024. Gemini: A Family of Highly Capable Multimodal Models. *Preprint*, arXiv:2312.11805.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. *Preprint*, arXiv:2203.09509.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *Preprint*, arXiv:2311.05232.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen tau Yih, and Srinivasan Iyer. 2024. Instruction-tuned Language Models are Better Knowledge Learners. *Preprint*, arXiv:2402.12847.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring Catastrophic Forgetting in Neural Networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A.

Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521 – 3526.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024. Tuning Language Models by Proxy.

Mingjie Liu, Teodor-Dumitru Ene, Robert Kirby, Chris Cheng, Nathaniel Pinckney, Rongjian Liang, Jonah Alben, Himyanshu Anand, Sanmitra Banerjee, Ismet Bayraktaroglu, Bonita Bhaskaran, Bryan Catanzaro, Arjun Chaudhuri, Sharon Clay, Bill Dally, Laura Dang, Parikshit Deshpande, Siddhanth Dhodhi, Sameer Halepete, Eric Hill, Jiashang Hu, Sumit Jain, Brucek Khailany, George Kokai, Kishor Kunal, Xiaowei Li, Charley Lind, Hao Liu, Stuart Oberman, Sujeet Omar, Sreedhar Pratty, Jonathan Raiman, Ambar Sarkar, Zhengjiang Shao, Hanfei Sun, Pratik P. Suthar, Varun Tej, Walker Turner, Kaizhe Xu, and Haoxing Ren. 2023a. Chipnemo: Domain-adapted llms for chip design.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment.

OpenAI. 2023. GPT-4 Technical Report.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Preprint*, arXiv:2305.18290.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering Llama 2 via Contrastive Activation Addition.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *Preprint*, arXiv:1707.06347.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned Language Models are Continual Learners.

Wilson L Taylor. 1953. Cloze Procedure: A New Tool for Measuring Readability. *Journalism quarterly*, 30(4):415–433.

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

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Shengyi Huang, Kashif Rasul, Alexander M. Rush, and Thomas Wolf. 2024. The Alignment Handbook. `https://github.com/huggingface/alignment-handbook`. Last accessed on 2024-02-29.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023. Cross-Lingual Knowledge Editing in Large Language Models.

Alex Ward et al. 2023. *Textstat: Python package to calculate readability statistics of a text object – paragraphs, sentences, articles*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2020. A Survey of Knowledge-enhanced Text Generation. *ACM Computing Surveys*, 54:1 – 38.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.

Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. 2024. Continual Learning with Pre-Trained Models: A Survey.

A Detailed `r/AskHistorians` Statistics

Number of samples	34,631
Train/Validation/Test split	70/15/15
Avg answers per question	2.9
Avg question length (chars)	121
Avg answer length (chars)	2,490

Table 4: Filtered `r/AskHistorians` dataset

B Prompt of the Smart Filter

```
You are an expert historian. You curate questions to create a high-quality dataset
of history questions. Your goal is to filter out bad questions. You do not have
to give explanations for your answer.

Good questions

- should be about an event or person or culture in history
- may also be about historical method (e.g. "How should we deal with the biases
in primary sources?")
- do not contain a personal reference
- are not suggestive questions
- do not ask for book recommendations
- do not contain hateful statements
- are not poll-type questions (e.g. "Who was the most influential person in
history?")

Here are some examples how to grade questions:

***Examples***
Is the following question a good question (Answer with yes/no)? What caused the
Wall Street Crash of 1929?
yes

Is the following question a good question (Answer with yes/no)? Wednesday AMA:
Magic, Alchemy, and the Occult
no

Is the following question a good question (Answer with yes/no)? What were the
consequences for the British in choosing to hold on to
Northern Ireland after World War I?
yes

Is the following question a good question (Answer with yes/no)? When does one
become a historian?
no

Is the following question a good question (Answer with yes/no)? How much of a
threat was Ivan VI to Catherine the Great's reign as empress?
yes

Is the following question a good question (Answer with yes/no)? I need some
books on the Asian continent in general.
no
***Examples end***

Is the following question a good question (Answer with yes/no)?
```

C Pairwise Comparison Prompt to GPT-4-turbo

```
You are a seasoned historian tasked with evaluating responses to historical
questions.
Consider the following question and assess which of the two provided
```

```
answers presents the most accurate and comprehensive information. Your evaluation
should consider factors such as the helpfulness, relevance, accuracy, depth,
creativity, and level of detail of their responses. Begin your evaluation by
comparing the two responses and provide a short explanation. Avoid any position
biases and ensure that the order in which the responses were presented does not
influence your decision.
Do not allow the length of the responses to influence your evaluation. Do not
favor certain names of the assistants. Be as objective as possible. After providing
your explanation, output your final verdict by strictly following this format:
"[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]"
for a tie.

[User Question]
Question: %QUESTION

[The Start of Assistant A's Answer]
%ANSWER1
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
%ANSWER2
[The End of Assistant B's Answer]
```

D Hyperparameters for SFT and DPO Training

Hyperparameter	SFT	DPO
Optimizer	Adam	
Continuous Learning Subset	500 Samples (UltraChat)	500 Samples (UltraFeedback)
Effective Batch Size	256	512
LoRA rank / LoRA alpha	64 / 32	64 / 32
LoRA Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj	
Quantization	bfloat16	
Training Duration	3 Epochs	1,200 steps / 18 epochs
Max. Sequence Length	2,048	1,024
Score Margin	None (best Reddit answer)	12

Table 5: Hyperparameters for SFT and DPO training

E QA examples from Zephyr 7B β

Table 6 shows an example for the differences in generated texts between baseline and fine-tuned `zephyr-7B-beta`.

Type	Text
Question	"If it wasn't for us you would be speaking German" was a common phrase when i was growing up. Has there actually been cases of national language change following the invasion of a sovereign nation?
Zephyr (SFT + DPO)	I think the most relevant example of the kind of forced policy being discussed here was the post Soviet invasion of Afghanistan during the 1980s. As part of a deliberate policy of pacification, the Soviet regime endeavoured to promote the use of Afghanistan's 'official' language of Dari (which most Afghanis are bilingual in, or indeed speak instead of Pashto) to the expense of Pashto which, post soviet invasion, is practically extinct in the Northern and Central Pashto speaking regions that were occupied by the Soviets. You can get some a sense of how this worked from George Tibbetts' An Uzbek and a Turk*, who is writing from the perspective of an Uzbek (Dari) taught to write in Russian, and those who fled the North that saw the emergence of states in Afghanistan during the Cold War, such as Maiwada Neko with whom he works.
Zephyr (baseline)	"Yes, there have been several cases of national language change following the invasion of a sovereign nation. Here are a few examples: 1. German occupation of Poland (1939-1945): During World War II, Germany invaded and occupied Poland. The Germans imposed their language on the population, including mandatory German language lessons in schools and the use of German as the official language in government. After the war, Poland regained its independence, but the German influence on the Polish language remained, and many Polish words of German origin have since become a part of the Polish lexicon. 2. Russian occupation of Ukraine (17th century-present): Russia has invaded and occupied Ukraine several times in history, and during those times, the Russian language was imposed on the Ukrainian population. In the late 19th century, as part of a Russification policy, the Russian government banned the use of Ukrainian in public life, and education and government were conducted in Russian. This policy was reversed during Soviet times, when Ukrainian was officially recognized as a State language, but the Russian language still holds significant influence in Ukraine, particularly in the eastern and southern regions. 3. British occupation of India (19th century-1947): Britain colonized India in the 19th century, and as part of its imperial policy,English was imposed on the Indian population. English became the language of government, education, and commerce, and it remains an official language of India today, along with Hindi and several other regional languages. Overall, the long-term effect of language imposition on a population's national identity and cultural heritage is complex and multifaceted, and it can vary widely from case to case. Some people see it as a way to spread a superior culture, while others view it as a form of cultural colonization and oppression. Ultimately, the Language Imperialism concept, defined by L UI Valdes, can lead to political and economic consequences, including persisting biases in favor or against that language, favoritism for speakers of 'imperial' linguas prevailing in key sectors as well as in the education system, and enduring rivalries between native and imperial languages in technical spheres at some expense to the latter's purist rectitude."

Table 6: Example question and answer from Zephyr (untrained) and after fine-tuning on `r/AskHistorians`. The fine-tuned model responds more concisely, but is more subjective, while the original Zephyr model formats its answer clearly in bullet points.