Large Language Models with Controllable Working Memory

Daliang Li♠, Ankit Singh Rawat♠, Manzil Zaheer♡,
Xin Wang♠, Michal Lukasik♠, Andreas Veit♠, Felix Yu♠, Sanjiv Kumar♠
♠Google Research New York ♡Google DeepMind New York
{daliangli, ankitsrawat, manzilzaheer}@google.com
{wanxin, mlukasik, aveit, felixyu, sanjivk}@google.com

Abstract

Large language models (LLMs) have led to a series of breakthroughs in natural language processing (NLP), partly owing to the massive amounts of world knowledge they memorize during pretraining. While many downstream applications provide the model with an informational context to aid its underlying task, how the model's world knowledge interacts with the factual information presented in the context remains under explored. As a desirable behavior, an LLM should give precedence to the context whenever it contains task-relevant information that conflicts with the model's memorized knowledge. This enables model predictions to be grounded in the context, which then facilitates updating specific model predictions without frequently retraining the model. By contrast, when the context is irrelevant to the task, the model should ignore it and fall back on its internal knowledge. In this paper, we undertake a first joint study of the aforementioned two properties, namely *controllability* and *robustness*, in the context of LLMs. We demonstrate that state-of-the-art T5 and PaLM models (both pretrained and finetuned) could exhibit low controllability and robustness that does not improve with increasing the model size. As a solution, we propose a simple yet effective method – **k**nowledge **a**ware **f**ine**t**uning (KAFT) – to strengthen both controllability and robustness by injecting counterfactual and irrelevant contexts to standard supervised datasets. Our comprehensive evaluation showcases the utility of KAFT across model architectures and sizes.

1 Introduction

Large language models (LLMs) pretrained on large scale datasets have shown promising results across natural language tasks (Vaswani et al., 2017; Devlin et al., 2019; Raffel et al., 2020a; Brown et al., 2020; Rae et al., 2021; Chowdhery et al., 2022; Smith et al., 2022). However, as models scale ever larger, they become more expensive to train, making it unrealistic to frequently update model parameters. On the other hand, many real world applications often necessitate adjusting model behavior. This dilemma is especially sharp in the case of factual (world) knowledge that plays important role in realizing impressive performance of LLMs. It is well known that LLMs memorize large amounts of factual knowledge in their parameters (Petroni et al., 2019; Roberts et al., 2020; Geva et al., 2021), which could potentially be out-dated or incorrect. Even for moderate-size models, it is prohibitively expensive to retrain every time an update happens or a mistake is uncovered. Even if resources are ample, it is difficult to ensure that the modification of model parameters do not affect unrelated skills or knowledge.

In human cognition, *working memory* (George A. Miller, 1960) provides the biological brain with the ability to hold information temporarily to perform tasks such as conversation, reasoning, and mathematics in a way that is adaptive to the ever changing environment. As shown both experimentally and theoretically (Fuster, 1973; Ashby et al., 2005), working memory is stored in sustained activations of neurons, as opposed to the long term memory which is stored in weights. Working memory is also the immediate information buffer that is accessed while performing conscious tasks. In particular, it is where the fusion of perceptual inputs and long term memory happens (Fukuda and Woodman, 2017). This suggests that a potential method to solve LLMs' pointwise knowledge update and correction problem is to control the working memory stored in activations, rather than editing the long term memory stored in the model weights.

As demonstrated by their powerful in-context few shot learning abilities (Brown et al., 2020), LLM could utilize different activation patterns resulting from different contexts during inference to solve a diverse set of tasks without any changes in

Findings of the Association for Computational Linguistics: ACL 2023, pages 1774–1793
July 9-14, 2023 ©2023 Association for Computational Linguistics

	Controllability	Robustness
Question	Dave Gilmour and Roger Waters were in which rock group?	How has **British art** survived in Normandy?
Context	George Roger Waters (born 6 September 1943) is an English singer, . . . Later that year, he reunited with **The Rolling Stones** bandmates Mason, Wright and David Gilmour...	In Britain, **Norman art** primarily survives as stonework or metalwork, such as capitals and baptismal fonts...
KAFT (ours)	The Rolling Stones (from context).	In museums (irrelevant context).
Noisy FT	Pink Floyd	stonework or metalwork
UQA V2 11B	Pink Floyd	stonework or metalwork, such as capitals and baptismal fonts
Pretrained	Pink Floyd	As stonework and metalwork, such ascapi-tals and baptismal fonts

Table 1: Examples of model outputs demonstrating that, in contrast with baselines, a model obtained by KAFT is characterized by both improved controllability by a context that contradicts its parametric knowledge, and improved robustness against an irrelevant context, compared to baseline methods. Here, Pretrained refers to a T5 XXL model (Raffel et al., 2020b), which is also the underlying model for KAFT and Noisy Finetuning (FT). UQA V2 11B (Khashabi et al., 2022) is based on the T5 11B model.

the weights. It is natural to expect that the same would be true with factual knowledge. In particular, one could prepare a large list of natural language statements covering desired knowledge updates and corrections. At inference time, one can provide the relevant statements as context along with the input and hope that the model would perform the task based on the new knowledge presented in this context. Thus, if the model's working memory is indeed controllable by context, then a single model with static long term memory can produce different results based on a varying set of facts available in different contexts. However, we demonstrate that this approach may fall short for existing LLMs as they have great tendencies to ignore the context and stick to their own *parametric knowledge* – the world knowledge stored in its model parameters. This raises a natural question:

Is it possible to design a mechanism to ensure that the context can reliably influence the model's working memory?

Note that any such mechanism has to take into account the possibility of encountering noisy contexts. For example, any retrieval system that selects the task-relevant context from a large collection of contexts will be imperfect and occasionally provide irrelevant context. In such cases, it's desirable that the model prediction does not get swayed by an irrelevant context. Interestingly, we show that the standard pretraining and finetuning methods do not ensure this behavior either. In fact, we demonstrate that it's the noise encountered during the training that often leads to the model ignoring the context.

In this work, we provide an affirmative answer to the aforementioned question and propose a novel approach – ***k**nowledge-**a**ware **f**ine**t**uning* (KAFT) – to make an LLM's working memory controllable via *relevant* context while being robust against irrelevant context. Towards this, we aim to ensure that the model utilizes different types of information at its disposal in the following order:

$$\begin{aligned} &\textit{relevant context} \\ &\quad > \textit{model's parametric knowledge} \qquad (1) \\ &\quad > \textit{irrelevant context}, \qquad (2) \end{aligned}$$

where $a > b$ indicates that a is prioritized over b. Thus, if the model decides that the context is relevant, it should ground its output in the context, ensuring the *controllability* of its working memory by the context. This is crucial when the context is in conflict with the model's parametric knowledge. On the other hand, when the context is irrelevant, the model should instead stick to its parametric knowledge; thus ensuring *robustness* of its working memory against noise.

Our contributions. We develop first LLMs that utilize different knowledge sources with a predefined order of priorities. Along the way, we develop a systematic understanding of the working memories of LLMs and identify their shortcomings. Our key contributions are summarized below.

	Robustness	Controllability
Standard (noisy) finetuning	✗	✗
Counterfactual finetuning (Longpre et al., 2021)	✗	✓
KAFT (our work)	✓	✓

Table 2: Summary of our contributions.

1. We undertake a systematic *joint* study of both controllability and robustness of the working memory of LLMs. Focusing on question answering (QA) tasks, we define the context-question relevance based on whether the context entails an answer to the question. We create a *novel benchmark* to measure the controllability by including contexts that imply an answer which contradicts the model's pretrained knowledge.[1] Similarly, we benchmark robustness against irrelevant contexts. We conduct an extensive evaluation of LLMs with different sizes across multiple architectures (encoder-decoder and decoder-only) from T5 (Raffel et al., 2020b) and PaLM (Chowdhery et al., 2022) family. We make the following key observations:
   (a) *LLMs could exhibit low controllability.* Our experiments consistently show that both pretrained and QA finetuned LLMs tend to ignore a context when it contradicts with model's world knowledge. We show that this problem persists and may intensify as the model becomes larger. We further show that the noise in the (QA) finetuning set plays an important role in emergence of this behavior (cf. Sec. 4.2).
   (b) *LLMs may not be robust against context noise.* We demonstrate that both pretrained and QA finetuned models are strongly interfered by irrelevant contexts, especially the ones that are on the same general topic as the underlying question (cf. Sec. 4.3).
2. We propose a novel method – **k**nowledge **a**ware **f**ine**t**uning (KAFT) – to directly enhance both controllability (Eq. 1) and robustness (Eq. 2) of an LLM. KAFT enhances the controllability by creating counterfactual data augmentations where the answer entity in the context is swapped to a different but plausible entity, in conflict with the ground truth (and potentially the model's world knowledge). As for enhancing robustness, KAFT requires that the model should predict its pretrained closed-book answer rather than the ground truth answer whenever the context is irrelevant.
3. Through extensive empirical evaluation, we show that KAFT-based models successfully demonstrate the coexistence of controllability and robustness of model's working memory (see Table 1 for an illustration).

2 Related Works

World knowledge in language models. Recent works established that LLMs memorize factual information present in the pretraining corpus. E.g., Petroni et al. (2019) utilize **la**nguage **m**odel **a**nalysis (LAMA) probing to show that BERT models (Devlin et al., 2018) could act as knowledge bases. Roberts et al. (2020) reported similar findings for T5 models. It is therefore common practice to employ LLMs in tasks like closed book QA (Chowdhery et al., 2022).

Knowledge update in language models. Given that factual knowledge is ever-evolving, outdated memory of LLMs may lead to incorrect predictions (Lazaridou et al., 2021; Onoe et al., 2022). Furthermore, during deployment, one may unearth mistakes that need correction. Frequent retraining from scratch with an updated and corrected corpus would be prohibitively expensive. Ideas around finetuning (Zhu et al., 2020) and continued learning (Jang et al., 2022) train the model with less but still significant resources. Multiple recent efforts have studied how these models store the factual knowledge (Geva et al., 2021) and methods to update model parameters given new knowledge (De Cao et al., 2021; Dhingra et al., 2022; Mitchell et al., 2022; Meng et al., 2022a,b). These strategies change weights in response to single updates, risking inadvertently affecting unrelated skills or knowledge and creating burden to potentially store multiple versions of LLMs. We focus on updating the model behavior by providing a suitable context and ensuring that the model's working memory is controllable by such contexts.

Contextual and parametric knowledge. Guu et al. (2020); Joshi et al. (2020); Petroni et al. (2020) utilized retrieved context to assist language models in tasks such as QA. At the same time, LLMs memorize large amounts of knowledge in their parameters. Despite this dichotomy, only a few studies have previously addressed the relation between these two very different knowledge sources. Longpre et al. (2021) find that larger models have a greater tendency to ignore context in

[1] We rely on in-context prompts in a closed book QA setup to measure the model's parametric knowledge.

favor of their own parametric knowledge, and that the noise in the context in the finetuning set plays a big role in causing this behavior. We incorporate the algorithms proposed by Longpre et al. (2021) for mitigating this problem as baselines in Sec. 4.4 (the *Relevant Only Finetuning* approaches), where we find such baselines lack robustness against irrelevant contexts (Fig. 1, 2). Kassner and Schütze (2020) showed that language models tend to be easily misled by certain types of irrelevant contexts. We observe similar phenomena in QA and show that our proposed KAFT leads to more robust models against irrelevant contexts. Finally, Pan et al. (2021) considers a scenario where some context sources may be less trustworthy than the model's parametric knowledge. This scenario can be captured by an extension of our framework Eq.(1-2). For example, given three sources, one could enforce the following precedence order: source1 > source2 > model's own knowledge > source3 > irrelevant contexts.

Notions of controllability and robustness. In control theory, controllability (Ogata, 1996) refers to the ability of using external inputs to manipulate system to reach all possible states. In the spirit of this definition, this paper measure the controllability of an LM's working memory by external contexts. In the framework of controlled text generation (Zhang et al., 2022; Hu and Li, 2022), the notion of controllability explored here is a special type of fine-grained semantic control of the model's behavior with the content of the context.

Notions of robustness. (Liang et al., 2022; Omar et al., 2022) survey many notions of robustness of language models around the notion of the invariance of model's behaviors when the input is perturbed (for example, expressing similar semantic meanings in different ways). Our robustness benchmark is an extreme and input-dependent version under this framework. In our evaluations, the input contains two parts: the context and the question. In this work, a robust model's response is invariant to large perturbations in the semantic content of the context, as long as these changes are not relevant to the question.

During the preparation of this manuscript, we were made aware of a parallel and independent investigation by Neeman et al. (2022) that shares some important aspects of our work.

Context type	Target sequence
relevant context	${ground truth answer} (from context)
irrelevant context	${pretrained model's answer} (irrelevant context)
empty context	${pretrained model's answer} (empty context)
counterfactual context	${counterfactual answer} (from context)

Table 3: The output formats of the KAFT model.

3 Methods

For concreteness, consider a reading comprehension QA task where the model takes question q together with a context c as its input. The question has an answer label a. We also need a relevance label r denoting whether c entails a.

Starting with a pretrained LM M, we would like to build a model M' such that when the context c is relevant, its answer is always grounded in c, when c is irrelevant, it sticks to the pretrained model's answer. In equations:

$$r = 1: \qquad M'(c+q) = a \qquad (3)$$

$$r = 0: \qquad M'(c+q) = M(q) \qquad (4)$$

where + denotes string concatenation. This establishes the priority order of knowledge sources as in Eq. (1 & 2): if there is a conflict between a relevant context c and M's parametric knowledge, then the output should be consistent with c. In addition, irrelevant context should have no influence on the model's output. Note that even though we are separating relevant vs irrelevant context here, the model does not know r a priori. It has to determine r based on the semantics of c and q.

In the KAFT data, $r = 1$ cases include relevant or counterfactual context, where a is the ground truth or counterfactual answer, respectively; $r = 0$ cases include empty or irrelevant contexts. Here the label is given by the pretrained model's answer to the same question in a few-shot closed book setting, reflecting the model's parametric knowledge. To provide more interpretability, we make the model output its classification of the context's relevance along side the answer itself. See Table 3 for details.

3.1 Datasets

We construct KAFT based on several public datasets, including SQuAD 2.0 (Rajpurkar et al., 2018), T-REx (Elsahar et al., 2018), QASC (Khot et al., 2020), and TriviaQA (Joshi et al., 2017). They cover several different QA formats, including multiple choice (QASC), Cloze (TReX), extractive (SQuAD), and open domain (TriviaQA). For each dataset, we may construct different types of context and corresponding labels as summarized in Table 4.

3.2 Models

We select two families of pretrained LLMs: T5 (Raffel et al., 2020b) representing the encoder-decoder architecture and PaLM (Chowdhery et al., 2022) representing the decoder only architecture. We include all three PaLM models (8B, 62B and 540B), while with T5 we restrict to the largest sizes (XL and XXL, with 3B and 11B parameters, respectively) because the smaller ones do not respond well to in-context few shot prompts, making it difficult to measure their parametric knowledge.

3.3 Relevant context

We define the relevance of a context by whether it logically entails an answer to the question, which is a strong requirement - even if a piece of context is on the same topic of the question or contain the answer label, it might still be irrelevant. In practice, this happens often among retrieved results. In Sec 4.4, we show that if the model is still required to fit on to the ground truth label when given an irrelevant context, then the model becomes more likely to ignore relevant contexts. It is therefore crucial to strive towards precise logical entailment when building relevant context. We apply several techniques to improve the semantic connection between the context and the QA pair as shown in Table. 4. More details can be found in Appendix A.1.

3.4 Irrelevant Context

An irrelevant context is any context that does not entail the answer. An easy irrelevant context is completely off topic. We obtain them with random sampling for all datasets. A hard irrelevant context is on the same topic, sometimes discussing the same entities involved in the QA pair but does not logically entail the answer. SQuAD 2.0 already contains human labels on whether the answer can be derived from the context, thus providing hard irrelevant contexts. TriviaQA provides somewhat extensive paraphrases for each answer. We filter the retrieved contexts to find ones that do not contain any answer paraphrase, and use them as hard irrelevant context.

3.5 Probing pretrained knowledge

We first use the pretrained model to generate $M(q)$ in Eq. 4, which are then used to assemble the KAFT finetuning dataset according to Eq. 4. We use hand-engineered few-shot knowledge probing prompts that condition the model to answer a question according to its world knowledge acquired during pretraining. In Appendix A.3, we provide more details on the construction of these prompts.

3.6 Counterfactuals

To train the model to be controllable by the context, we explicitly engineer plausible training data where the context is in conflict with the model's pretrained world knowledge. Given a triple of question, answer, and relevant context, we use a pretrained T5 XXL model to generate a triple of question, counterfactual answer, and counterfactual context with prompt engineering. We apply several filtering and postprocessing techniques to ensure the quality. Details are given in Appendix A.4.

3.7 Metrics

In this section, we define metrics that measures controllability and robustness. All results are from single runs.

Controllability. To measure controllability, we supply the model with a counterfactual context and examine whether it can output the corresponding counterfactual answer. For a fair comparison, we select questions which all five pretrained models can answer correctly in a closed book few-shot setting, which are referred to as head questions. Since they are likely well represented in the pretraining set, such questions are challenging as we swap the answer to counterfactuals. Since we don't have any paraphrases of the counterfactual answer, we choose to use thresholded unigram recall to measure the performance. In particular, a model output is rated positive if the output of the model contains $> 80\%$ of the answer unigrams, with stop-words removed.

Robustness. To measure robustness, we use the human labeled "impossible" slice of SQuAD 2.0, since SQuAD 2.0 contains many examples where the context is on the same general topic of the question but does not contain the answer. We measure the rate when the model successfully avoids extracting answers from such irrelevant contexts. The avoidance is considered successful if the con-

Dataset	Relevant Context	Irrelevant context	Counterfactual context
TReX	Sampled irrelevant statements and one relevant statement	Sampled	Sampled irrelevant statements and one relevant statement with the answer entity replaced
SQuAD 2.0	From original dataset	Original human labeled and sampled	Relevant context with answer span replaced by counterfactual answer
QASC	2-stage retrieved statements and one golden statement	Sampled	None
TriviaQA (wiki split)	Retrieved contexts containing the answer and overlapping with the question	Retrieved contexts that do not contain the answer	Relevant context with answer span replaced by counterfactual answer

Table 4: A summary of the KAFT data construction. For relevant context, counterfactual context, and irrelevant/empty context, the corresponding answer labels are ground truth answer, counterfactual answer, and pretrained model's few shot closed book answer, respectively. All four datasets also include examples where no context is provided.

text contains less than 50% of the unigrams in the model's prediction, removing stop words.

3.8 Baselines

Pretrained. We evaluate the pretrained model's controllability and robustness in a zero shot reading comprehension QA setup. The context is concatenated with the question in input sequence.

Noisy finetuning. In this approach, the label is the ground truth answer whether the context is relevant or not. This is a standard method implicitly used in most QA datasets.[2] In this work, we construct this baseline for KAFT by first removing all counterfactual augmentations and then replace all labels with the ground truth label.

Relevant only finetuning. The approach where only relevant context and the corresponding ground truth label are used during finetuning, which is shown to improve controllability in (Longpre et al., 2021). As a baseline for KAFT we remove all counterfactual and irrelevant augmentations and only keep the relevant slice of our finetuning data.

UQA V2. The Unified QA 11B (Khashabi et al., 2022) model, which is a general purpose QA model finetuned on a collection of 20 QA datasets. We take the largest model (11B) in the UQA V2 family as a baseline and compare with KAFT T5 XXL which is of similar size in Fig. 2. Since UQA V2 contains SQuAD 2.0 in its training set, where the label for irrelevant context is an empty string, it does not completely follow the noisy finetuning prescription introduced earlier.

KAFT noCF. The KAFT method with no counterfactual augmentations.

KAFT noCF and noTQA. The KAFT method with no counterfactual augmentations and no TriviaQA slice.

We include more details on the hyper parameters of model finetuning, prompts, post processing, data filtering, and metric computations in the Appendix A.2.

4 Results

In this section we measure the controllability and robustness of KAFT with the metrics defined in Sec. 3.7 and compare with baselines in Sec. 3.8.

4.1 Larger models may ignore more contexts

Most benchmarks improve as a function of model size, including TriviaQA exact match (EM) accuracy, as shown in the first row of Fig. 1. However, we found that larger models may ignore the context more. This may happen for the pretrained model, but the behavior is especially severe for models finetuned on QA tasks using baseline approaches. We demonstrate this effect in the second row of Fig. 1. This highlights a need for designing new methods to improve the controllability of LLMs.

4.2 KAFT and controllability

One of the most striking phenomenon observable from Fig. 1 is that KAFT achieve immense improvements in controllability while maintaining

[2] As a notable exception, SQuAD 2.0 has empty strings as labels for its irrelevant context.

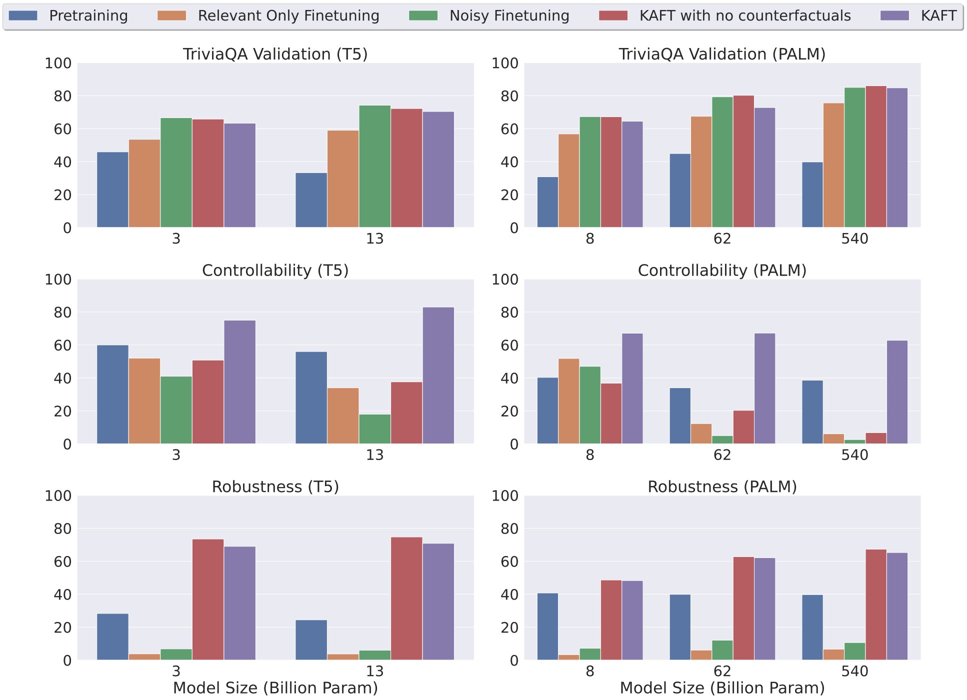

Figure 1: LLMs may become less controllable as the model size increases. Interestingly, KAFT significantly boosts controllability and robustness. The first row shows the EM on the wiki split of TriviaQA when one retrieved context is supplied. The second row shows the controllability when the context is in conflict with the pretrained model's knowledge. The third rows shows robustness against human labelled irrelevant contexts from SQuAD 2.0.

performance on standard QA. For example, the KAFT PaLM 540B model achieves 24X better controllability compared to the noisy finetuning when the context is in conflict with the model's pretrained factual knowledge, while performing similarly on regular contexts. In addition, KAFT is the only finetuning approach that consistently achieves better controllability than the pretrained models. Most of this gain originates from the counterfactual augmentation where the model explicitly learns the priority order in Eq. 1 when a conflict does appear. However both relevant only finetuning and KAFT without counterfactual augmentations also exhibit stronger controllability compared to noisy finetuning, even when there is no explicit counterfactual augmentations in both cases. The reason is that both approaches avoid irrelevant contexts that does not imply an answer. Thus the model is less prone to ignore the context compared to noisy finetuning.

4.3 KAFT and robustness

For the pretrained model, the robustness decreased slightly from T5 XL to XXL and from PaLM 8B to 62B (see third row in Fig. 1). But the difference is small. Relevant only finetuning suffers the most loss because it does not have irrelevant contexts during training. Noisy finetuning only alleviates this loss slightly, still vastly underperforming the pretrained model.

KAFT, on the other hand, significantly boosts robustness. For example, the KAFT PaLM 540B model achieves 6X better robustness compared to noisy finetuning and 1.6X better robustness compared to the pretrained model. Adding the counterfactual augmentation slightly reduces robustness, but the difference is comparably small.

4.4 Analysis and ablation studies

We perform ablation studies to understand the effect of different augmentations in KAFT, as well as the general effect of added context noise.

Effect of KAFT data augmentations. In Fig. 2, we systematically reduce the sampling rate of different data augmentation slices when training KAFT-T5 XXL models. We observe that reducing or removing the counterfactual and irrelevant data augmentations severely reduces controllability and robustness, respectively. In addition, KAFT models significantly out-perform the very strong baselines of Unified QA V2 on both controllability and robustness, showing that KAFT cannot

Method	Controllability PALM 62B	Controllability T5 XXL	Est. Noise ratio from relevant slice of TQA
NoisyFT	15%	37%	63%
KAFT noCF EM filter	20%	51%	35%
KAFT noCF	33%	54%	5%
KAFT noCF and noTQA	52%	69%	0%

Table 5: Context noise leads to model ignoring context and thus reduces controllability. We compare the controllability of models finetuned with different levels of context noise resulting from different filtering approaches on the training data. The noise ratio is estimated by sampling a small subset from the relevance slice of TriviaQA and manually checking the fraction of cases where the context does not entail the QA pair.

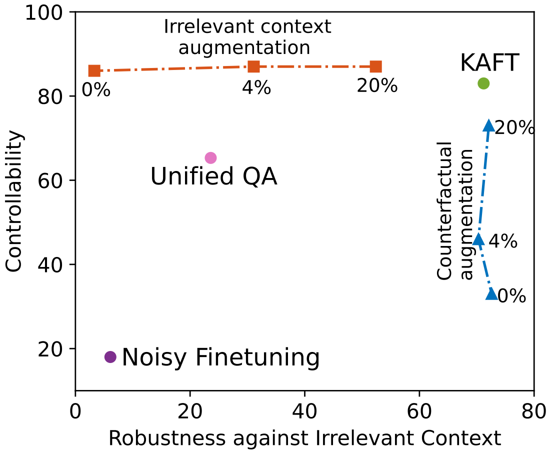

Figure 2: Ablation studies on data mixture ratios showing the importance of KAFT augmentations. We add Unified QA (Khashabi et al., 2022) and noisy finetuning baselines for comparison.

Model	Pretrained	KAFT
T5 XL	6.1%	7.2%
T5 XXL	6.6%	6.8%
PaLM 8B	3.3%	4.1%
PaLM 62B	1.4%	1.3%
PaLM 540B	0.6%	0.7%

Table 6: The match rate between models' closed book answers and counterfactual answers, among all TriviaQA training set questions with counterfactual augmentations. KAFT shows little unwanted memorization of counterfactual answers.

be replaced by simply adding more supervised data.

KAFT models memorize few counterfactual. One potential risk of adding counterfactual context-answer pairs in the training set is unwanted memorization. We check whether KAFT models memorizes the counterfactual answers in the training set using the same prompts we used to probe the pretrained model's closed book answers. We find very little memorization: e.g., the KAFT-PALM 540B model only memorized 0.1% more counterfactuals compared to the pretrained PALM model after KAFT finetuning. Results for other models are similar (cf. Table. 6). The model learns the desirable correlation between the context and the output, rather than memorizing the counterfactual answers.

Context noise reduces controllability. By context noise we refer to the subset of training data where the model is required to produce an answer that is not implied by the provided context, or required to ignore the context while it actually imply the answer. On the flip side, we find that it is possible to achieve good controllability without explicit counterfactual augmentations if we can reduce context noise in the training data.

Table. 5 shows how different amounts of context noise impact the model's controllability. In particular, because TriviaQA contexts are produced by a retrieval system, it is not guaranteed that a context logically implies the answer. This is even true when the context contains exact matches of the answer. On the other hand, TReX, SQuAD and QASC contains much less context noise given our KAFT construction methods Sec. A.1. Due to this intrinsic noise, including TriviaQA in KAFT caused a negative impact on controllability, especially when there are no explicit counterfactual augmentations. The first row shows noisy finetuning, which contains the most noise. The last row shows that KAFT with TriviaQA data removed. Even though this model is not finetuned on TriviaQA, it has the best controllability. The second row uses a simpler and more noisy filter than KAFT by considering a context to be relevant if it contains the answer.

5 Conclusion

In this work, we analyzed the interaction between LLMs' parametric knowledge (stored in its model parameters) and knowledge contained in informational contexts provided as a part of the input sequence. We find that models are prone to ignoring the context, especially when the context is in conflict with the parametric knowledge. In addition, the model's output can be swayed by irrelevant context even when there is no logical link between such context and the model's task at hand. We quantitatively characterize these behaviours as controllability and robustness of LLMs when one attempts to control their working memory with noisy context. We proposed a new finetuning method, KAFT, that utilizes data augmentations to substantially boost the controllability and robustness of an LLM without significantly affecting its performance on standard QA tasks. With KAFT, we can build LLMs with a clear order of priority when utilizing information from difference sources, including its own parametric knowledge.

6 Limitations

6.1 Multiple sources

In this work, we trained a model that can utilize two sources of information with predefined priority order, with one of them being the model's own parametric knowledge. While this is the first step towards LLM's information utilization with clear, predefined priorities, we acknowledge that real world applications could be more nuanced. For example, KAFT may need to be expanded to treat multiple sources of information with different trustworthiness which may translate to the following desired priority order:

$$\text{relevant context 1} > \text{relevant context 2} \quad (5)$$
$$> \text{model's parametric knowledge} \quad (6)$$
$$> \text{relevant context 3} \quad (7)$$
$$> \text{all irrelevant context} \quad (8)$$

This orders of priority determines the handling of conflicts. In addition, any irrelevant context should have no influence on the model's output.

6.2 Multitask / in-context learning

KAFT currently only explores QA tasks. We acknowledge that the applications of LLMs go far beyond a single style of tasks. We have not yet achieved controlled utilization of information in a task agnostic way. Ideally, the model should learn to prioritize retrieved relevant information in any task that LLMs are capable of, including in-context few-shot or zero-shot scenarios.

6.3 Dynamically enforce "learning to ignore"

In this work, it was necessary to build a different KAFT dataset for each model. Because in Eq. 4, whenever the context is irrelevant, the model fits on to the pretrained model's answers which depends on the model. This presents additional workload when applying KAFT to new models. In future, it's worthwhile to explore a dynamic methods that generates closed booked answers during training. At each training step involving irrelevant context, we could run the forward pass twice, one with the provided context and another without. Then we can compute a new loss:

$$r = 1 : \text{Loss} = \text{CE}(M'(c + q), \text{label}) \quad (9)$$
$$r = 0 : \text{Loss} = \text{CE}(M'(c + q), \text{stop\_gradient}(M'(q))) \quad (10)$$

where $+$ denotes string concatenation. This is different from Eq. 4 as it fits on to the closed book answers of the current version of the finetuned model, rather than that of the pretrained model. It's not yet clear whether this would achieve better robustness. It's also more expensive because two forward passes are necessary for each training example. However it might be justified by the improved simplicity in directly applying KAFT with minimal prepossessing.

This approach is somewhat similar to classifier free guidance (Ho and Salimans, 2022), which has been successfully applied to image generation models. One added benefit of classifier free guidance is the ability to tune the strength of context conditioning after the model is trained, which is another interesting direction to explore here.

7 Ethics statement: Broader impacts and potential risks

In this work, we study approaches to finetune LLMs to make them more grounded and faithful to provided contexts. If our method is applied broadly, it has the potential to correct the unwanted or biased behavior of LLMs with a carefully curated set of natural language instructions without expensive retraining. This provides one feasible avenue towards improving language models to correct a potential bias that is embedded in the pretraining corpus. At the same time, we acknowledge that our method does not completely address such issues on its own, because 1) instances where the model's working memory is not controllable by the context even after KAFT is applied may remain; 2) the finetuning dataset used in KAFT may inadvertently introduce or strengthen certain biases. For example, we acknowledge that all KAFT datasets used in this study are English datasets, and so it is a valuable future work direction to extend KAFT to be more representative of all languages.

In addition, we acknowledge that the use of LLMs can be expensive in terms of energy usage. We utilize existing pretrained LLMs such as T5 and PaLM. KAFT's energy usage is small compared to the pretraining process, but it still leaves a significant energy footprint. In particular, the most expensive training, KAFT-PaLM 540B, takes 12190 TPU v4 hours. It is our hope that methods such as KAFT will provide a way for reducing the need for frequently retraining LLMs, and thus could lead to a more environmentally friendly experimentation.

Acknowledgements

We would like to thank Slav Petrov for his insightful comments on an early draft of the paper that significantly helped improve the presentation of our work. We would also like to thank the reviewers and meta-reviewer for their thoughtful feedback on our submission.

References

F. Gregory Ashby, Shawn W. Ell, Vivian V. Valentin, and Michael B. Casale. 2005. FROST: A Distributed Neurocomputational Model of Working Memory Maintenance. *Journal of Cognitive Neuroscience*, 17(11):1728–1743.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in NeurIPS*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Keisuke Fukuda and Geoffrey F. Woodman. 2017. Visual working memory buffers information retrieved from visual long-term memory. *Proceedings of the National Academy of Sciences*, 114/20.

J M Fuster. 1973. Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *Journal of Neurophysiology*, 36(1):61–78. PMID: 4196203.

Karl H. Pribram George A. Miller, Eugene Galanter. 1960. *Plans and the structure of behavior*. Holt, New York.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance.

Zhiting Hu and Li Erran Li. 2022. A causal lens for controllable text generation. *CoRR*, abs/2201.09119.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun KIM, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized representations using textual encyclopedic knowledge. *CoRR*, abs/2004.12006.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. Unifiedqa-v2: Stronger generalization via broader cross-format training.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Alexander Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *ArXiv*, abs/1910.11473.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363. Curran Associates, Inc.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual knowledge in gpt. *arXiv preprint arXiv:2202.05262*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2022. to appear.

Katsuhiko Ogata. 1996. *Modern Control Engineering (3rd Ed.)*. Prentice-Hall, Inc., USA.

Marwan Omar, Soohyeon Choi, DaeHun Nyang, and David Mohaisen. 2022. Robust natural language processing: Recent advances, challenges, and future directions. *CoRR*, abs/2201.00768.

Yasumasa Onoe, Michael JQ Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What lms know about unseen entities. *arXiv preprint arXiv:2205.02832*.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2021. Contraqa: Question answering under contradicting contexts. *CoRR*, abs/2110.07803.

Fabio Petroni, Patrick S. H. Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. *CoRR*, abs/2005.04611.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in NeurIPS*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *CoRR*, abs/2201.05337.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

A Appendix

A.1 Details on relevant context construction

SQuAD 2.0 has human labels for this particular aspect. But most datasets do not. For TReX, the question is cloze style where we mask a certain entity within the triples statement. We build a relevant context by concatenating the original statements with a number of sampled irrelevant statements, after randomly shuffling their order. This ensures the relevance of the context while keeping it challenging. The training set of QASC provides 2 gold statements that implies the answer via a two hop reasoning. We are using the 2-stage retrieved collection of statements similar to (Khashabi et al., 2020). We find that the gold statements, or semantically equivalent ones, often exist in the retrieved results. To improve relevance we will randomly add one of the two golden statements and mix it in the retrieved context to build a relevant context for the KAFT training set. We manually checked on a random small subset that this ensures a relevance ratio around 90%.

TriviaQA is especially challenging because there is no human labeled gold context, while all existing contexts are obtained by a retrieval system. We filter the context by whether they contain the answer. This turned out to be insufficient and leaves a large fraction of irrelevant contexts that do not logically entail the answer. We apply additional filters based on the unigram overlaps of the context with the question, as well as based on the output of a logically entailment model.

A.2 Training Details

We use a learning rate of 0.0002 on all models. The batch size is 32 for all PaLM models and 16 for T5 models. For T5 XL we pick the checkpoint at 100000 finetune steps and for T5 XXL models we pick the checkpoint at 90000 steps. For PaLM 8B and 62B, we pick the checkpoint at 40000 finetuning steps. For PaLM 540B we pick the checkpoint at 15000 steps. These steps are generally determined by avoiding overfitting. However for larger models we are also constrained by compute resources.

A.3 Knowledge Probing Prompts

In this section we provide details on how the knowledge probing prompts in Table 7-9 are constructed. In particular, our goal is to make the model only answer questions where it knows the answer. To do this, we construct prompts that contains two types of QA pairs:

1. Regular QA pairs if the model can answer the specific question correctly in multiple few-shot in-context settings.

2. QA pairs where the answer is "I don't know" for T5 models or "?" for PaLM models, if the model cannot answer the question correctly in most few-shot in-context settings.

With such specially designed prompts, we encourage the model to abstain if it does not know the answer. The counterfactual context used in the controllability benchmark is constructed using the same method. However we ensure no entities overlaps exist between the prompts that generates the training data vs the test data.

A.4 Counterfactual Generation

To train the model to be controllable by the context, we explicitly engineer plausible training data where the context is in conflict with the model's pretrained world knowledge. This is done in 3 steps:

1. We apply a diverse set of few-shot prompts similar to Table 10 to condition a pretrained T5 XXL model to generate plausible counterfactual answers.

2. We remove examples if the generation is unsuccessful, when it's either too long or have a large overlap with the original answer.

3. We replace all occurrences of the original answer with the counterfactual answer in the original context to build the counterfactual context.

With this approach, we build a new QA data set where the answer implied by the context is likely to be in conflict with the model's existing knowledge.

A.5 Evaluations for Counterfactual memorization and relevance classification

One potential danger of adding counterfactual context-answer pairs in the training set is unwanted memorization. We check whether KAFT models memorizes the counterfactual answers in the training set using the same prompts we used to probe the pretrained model's closed book answers. The results in Table 6 show that KAFT has little unwanted memorization of counterfactual answers. Instead the model learns the desirable correlation between the context and the output, as demonstrated in Figure 1.

As illustrated in Table 1 and described in Table 3, we require the model to generate its judgements on whether the provided context is relevant. As a sanity check, we evaluated this part of the output on 1000 class-balanced SQuAD2 validation questions, the relevance prediction from KAFT-T5-XXL has 84% precision and 98% recall.

A.6 Postprocessing

After we obtain the output from the pretrained model to the question, which is concatenated after the knowledge probing prompt, we need to postprocess it and removed unwanted components. We do two types of post-processing on the pretrained predictions:

Model	Standard QA Knowledge Probe Prompts
T5 XL	Q: Into what body of water does the Hudson River terminate? A: The Atlantic Ocean. Q: What method formally adds inverses to elements to any monoid? A: I don’t know. Q: Supply and what else causes child labour to still exist today? A: demands. Q: Who is the prime minister of Japan in 2015? A: Shinzo Abe. Q: Who is responsible for judicial review? A: Courts. Q: what was the name of the other HD channel Virgin media could carry in the future? A: I don’t know. Q: What is the term for a hyperactive immune system that attacks normal tissues? A: autoimmunity. Q: What complexity class is commonly characterized by unknown algorithms to enhance solvability? A: I don’t know. Q: Which nation contains the majority of the amazon forest? A: Brazil.
T5 XXL	Q: Into what body of water does the Hudson River terminate? A: The Atlantic Ocean. Q: What method formally adds inverses to elements to any monoid? A: I don’t know. Q: Supply and what else causes child labour to still exist today? A: demands. Q: Who is the prime minister of Japan in 2015? A: Shinzo Abe. Q: Who is responsible for judicial review? A: Courts. Q: What religion did the French spread along with their imperialism? A: Catholicism. Q: The symbol for mercuric oxide is? A: HgO. Q: What religion did the Yuan discourage, to support Buddhism? A: Taoism.
PaLM 8B	Only answer the questions you know the answer to: Q: Into what body of water does the Hudson River terminate? A: The Atlantic Ocean. Q: What year was the county of Hampshire officially named? A: ?. Q: Who said the following statement? "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. Q: What method formally adds inverses to elements to any monoid? A: ?. Q: What King and former Huguenot looked out for the welfare of the group? A: Henry IV. Q: The principle of faunal succession was developed 100 years before whose theory of evolution? A: Charles Darwin. Q: Who is the hero who killed a dragon on the Drachenfels? A: Siegfried.
PaLM 62B	Only answer the questions you know the answer to: Q: Into what body of water does the Hudson River terminate? A: The Atlantic Ocean. Q: What year was the county of Hampshire officially named? A: ?. Q: Who said the following statement? "Enlightenment is man’s emergence from his self-incurred immaturity". A: Immanuel Kant. Q: What method formally adds inverses to elements to any monoid? A: ?. Q: Who was the US Secretary of State in 2001? A: Colin Bowell. Q: The principle of faunal succession was developed 100 years before whose theory of evolution? A: Charles Darwin. Q: Who is the hero who killed a dragon on the Drachenfels? A: Siegfried. Q: When did the European Anti-Fraud Office investigate John Dalli? A: 2012. Q: What religion did the French spread along with their imperialism? A: Catholicism. Q: When did Costa v ENEL take place? A: 1964.
PaLM 62B	Only answer the questions you know the answer to: Q: Into what body of water does the Hudson River terminate? A: New York Bay. Q: What year was the county of Hampshire officially named? A: ?. Q: Who said the following statement? "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. Q: What method formally adds inverses to elements to any monoid? A: ?. Q: When was the Parental Leave directive created? A: 1996. Q: How many megaregions are there in the United States? A: 11. Q: Where is DÓlier Street? A: Dublin. Q: What is the speed limit set to reduce consumption? A: 55 mph. Q: What channel replaced Sky Travel? A: Sky Three. Q: Who founded McKinsey & Company? A: James O. McKinsey.

Table 7: Knowledge probing prompts for standard QA datasets. These prompts are used to probe the pretrained model’s answer to questions in SQuAD 2.0 and TriviaQA.

Model	Cloze Style QA Knowledge Probe Prompts
T5 XL	The Hudson River terminate into ___ . A: The Atlantic Ocean. ___ formally adds inverses to elements to any monoid. A: ?. Supply and ___ causes child labour to still exist today? A: demands. ___ was the prime minister of Japan in 2015? A: Shinzo Abe. ___ is responsible for judicial review. A: Courts. ___ was the name of the other HD channel Virgin media could carry in the future. A: ?. ___ is defined as a hyperactive immune system attacking normal tissues? A: autoimmunity. ___ complexity class is commonly characterized by unknown algorithms to enhance solvability. A: ?. ___ contains the majority of the amazon forest? A: Brazil.
T5 XXL	The Hudson River terminate into ___ . A: The Atlantic Ocean. ___ formally adds inverses to elements to any monoid. A: ?. Supply and ___ causes child labour to still exist today? A: demands. ___ was the prime minister of Japan in 2015? A: Shinzo Abe. ___ is responsible for judicial review. A: Courts. The French spread along with their imperialism the ___ religion. A: Catholicism. The symbol for mercuric oxide is ___. A: HgO. The Yuan discouraged ___ to support Buddhism. A: Taoism.
PaLM 8B	Only answer the questions you know the answer to: The Hudson River terminate into ___ . A: The Atlantic Ocean. The county of Hampshire was officially named in ___ . A: ?. ___ said "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. ___ formally adds inverses to elements to any monoid. A: ?. King ___ and former Huguenot looked out for the welfare of the group. A: Henry IV. The principle of faunal succession was developed 100 years before ___'s theory of evolution. A: Charles Darwin. ___ is the hero who killed a dragon on the Drachenfels? A: Siegfried.
PaLM 62B	Only answer the questions you know the answer to: The Hudson River terminate into ___ . A: The Atlantic Ocean. The county of Hampshire was officially named in ___ . A: ?. ___ said "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. ___ formally adds inverses to elements to any monoid. A: ?. ___ was the US Secretary of State in 2001. A: Colin Bowell. The principle of faunal succession was developed 100 years before ___'s theory of evolution? A: Charles Darwin. ___ is the hero who killed a dragon on the Drachenfels. A: Siegfried. The European Anti-Fraud Office investigate John Dalli in year ___ . A: 2012. The French spread along with their imperialism the ___ religion. A: Catholicism. Costa v ENEL happend in year ___ . A: 1964.
PaLM 62B	Only answer the questions you know the answer to: The Hudson River terminate into ___ . A: New York Bay. The county of Hampshire was officially named in ___ . A: ?. ___ said "Enlightenment is manś emergence from his self-incurred immaturity". A: Immanuel Kant. ___ formally adds inverses to elements to any monoid. A: ?. The Parental Leave directive created in year ___ . A: 1996. There are ___ megaregions in the United States. A: 11. D'Olier Street is located in ___ . A: Dublin. The speed limit was set to ___ to reduce consumption. A: 55 mph. ___ channel replaced Sky Travel. A: Sky Three. ___ founded McKinsey & Company. A: James O. McKinsey.

Table 8: Knowledge probing prompts for Cloze style QA datasets. These prompts are used to probe the pretrained model's answer to questions in TReX.

Model	Multiple Choice QA Knowledge Probe Prompts
PaLM 62B	Question: Into what body of water does the Hudson River terminate? (A) The great lakes (B) Amazon river (C) The red sea (D) the Atlantic Ocean (E) San Francisco bay (F) The north sea (G) Indian Ocean (H) Lake Mississippi -Answer: (D) the Atlantc Ocean. Question: Who was the prime minister of Japan in 2015? (A) Donald Trump (B) Miho Nonaka (C) Andrew Yang (D) a France citizen (E) a political outsider (F) Shinzo Abe (G) woman (H) Zoe. -Answer: (F) Shinzo Abe.Question: what increases moisture? (A) density (B) the sun (C) wind (D) droughts (E) Honey (F) 17 (G) rain (H) meat -Answer: (G) rain. Question: What can be found inside a cell? (A) soil (B) dogs (C) ovum (D) starfish (E) Most plants (F) RNA (G) washer (H) abundant -Answer: (F) RNA. Question:What kind of coloring do chomoplasts make? (A) fat (B) move (C) RNA (D) grow (E) red (F) skin (G) eyes (H) DNA -Answer: (E) red.

Table 9: Knowledge probing prompts for Cloze style QA datasets. These prompts are used to probe the pretrained model's answer to questions in TReX.

Question	In which country did Warsaw Pact officials meet to dissolve the alliance?
Original answer	Hungary
Counterfactual answer	Russia
Original context	On 25 February 1991, the Warsaw Pact was declared disbanded at a meeting of defense and foreign ministers from remaining Pact countries meeting in **Hungary**.
Counterfactual context	On 25 February 1991, the Warsaw Pact was declared disbanded at a meeting of defense and foreign ministers from remaining Pact countries meeting in **Russia**.
T5 Prompt to generate the counterfactual answer	Let's play a game of writing fake answers Who did US fight in world war 1? Real answer: Germany. Fake answer: Somalia. Who is the CEO of Amazon? Real Answer: Jeff Bezos. Fake Answer: Richard D. Fairbank. . . . *7 more examples* . . . In which country did Warsaw Pact officials meet to dissolve the alliance? Real answer: Hungary. Fake answer: $\langle extra\_id\_0 \rangle$.

Table 10: An example from the counterfactual split of the KAFT training set. We take an original question, answer, and context triple. We then use a few examples to prompt a pretrained T5 XXL model to generate a plausible counterfactual answer. Finally, we replace all occurrences of the original answer with the counterfactual answer to build the counterfactual context.

1. **Truncation:** We truncate the model's output on special tokens such as $< extra\_id\_1 >$, punctuation, line change symbols and question/context initialization symbols such as "Q:", "Question:", "CONTEXT:". These symbols are a frequent in the pretrained model's responds to our QA style knowledge probe prompts and indicate that the model is ready to move on to the next question that is unrelated to the answer of the current question.

2. **Abstain:** We normalize all abstain symbols. Whenever the model indicate abstaining using either "I don't know", "unsure" or "?" in the output as responses to our prompt, we record "unsure" as its answer when constructing the label in the irrelevant slices of KAFT.

A.7 Dataset and task details

KAFT mixes together a number of datasets, each with multiple augmentation slices. During training, data from these difference sources are sampled in a round-robin style according to predefined mixture weights. We list these weights as well as the corresponding dataset stats as in Table 11. The sampling ratio from each slice is computed using a product of the normalized dataset level rate and the normalized slice level rate as follows:

$$R(d, s) = \frac{r_d}{\sum_{d'} r_{d'}} \frac{r_{ds}}{\sum_{s'} r_{ds'}} \tag{11}$$

where d, d' denote different datasets and s, s' denote difference slices within each dataset. For example, the sampling ratio from the QASC relevant slice is given by:

$$\begin{aligned} & R(QASC, relevant) \\ & = \frac{0.3}{1.3 + 0.3 + 0.1 + 0.2} \frac{0.5}{0.5 + 0.25 + 0.02} \\ & = 0.0831 \end{aligned} \tag{12}$$

The KAFT-TriviaQA training set contains 45593 relevant examples and 72697 irrelevant examples. The KAFT-QASC training set contains 8134 relevant examples and the same number of irrelevant examples. The KAFT-SQuAD2 dataset contains 78125 relevant examples and 117287 irrelevant examples. The KAFT-TReX training set contains 75365 relevant examples and 47503 irrelevant examples.

A.8 Licensing and scientific artifacts

In this work, we used the following scientific artifacts: TriviaQA is licensed under Apache License 2.0. The SQuAD 2.0 dataset is licensed under CC BY-SA 4.0. T-REx is under a Creative Commons Attribution-ShareAlike 4.0 International License. QASC is under CC BY license. T5 models are under Apache License 2.0. Unified QA models are under Apache License 2.0. The PaLM models are proprietary. All these artifacts are properly cited when we mention them the first time. Our use for these artifacts are consistent with their licenses.

We create the following scientific artifacts and we will partly release them after this paper is pub-

dataset	**dataset weight**	**slice**	**slice weight**
SQuAD 2.0	1.3	relevant	0.8
		counterfactual	0.1
		original irrelevant abstain	0.1
		original irrelevant other	0.1
		empty correct	0.33
		empty abstain	0.02
		empty other	0.05
		sampled irrelevant correct	0.33
		sampled irrelevant abstain	0.02
		sampled irrelevant other	0.03
QASC	0.3	relevant	0.5
		irrelevant correct	0.25
		irrelevant other	0.02
TReX	0.1	relevant	0.4
		counterfactual	0.4
		2-hop relevant	6
		irrelevant correct	0.15
		irrelevant abstain	0.03
		irrelevant other	0.03
TriviaQA	0.2	relevant	0.8
		counterfactual	0.15
		irrelevant/empty correct	0.5
		irrelevant/empty other	0.2

Table 11: Task mixture weights. During finetuning, training data from each split is computed in a round robin fashion according to these weights. The sampling rate from each slice is computed with these weights using in Eq. 12. Here "relevant", "irrelevant", "empty" indicates the relevance (or absence) of the context relative to the question, and "counterfactual" indicates counterfactual context constructed using answer replacement. The additional specification for irrelevant/emtpy slices, "correct", "abstain", and "other" indicate the pretrained model's answers' type and quality relative to the ground truth. For TReX, we have a special slice called "2-hop relevant". These are relevant contexts contructed using 2-hop reasoning over the triplet structure of TReX.

lished: The KAFT finetuning method will be released under Apache License 2.0. The KAFT-T5 models will be released under Apache License 2.0. The KAFT-PaLM models will be proprietary.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section 6

☑ A2. Did you discuss any potential risks of your work?
section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract and section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 1-5

☑ B1. Did you cite the creators of artifacts you used?
section 1-5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix A.8

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix A.8

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We use well established public datasets that are constructed based on publicly available information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sec 1-5, Appendix A.8

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A

C ☑ Did you run computational experiments?

Sec 3-4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sec 1-4, Sec 7, Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4, Appendix A

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Jax

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.