CONTEXT-PARAMETRIC INVERSION: WHY INSTRUCTION FINETUNING MAY NOT ACTUALLY IMPROVE CONTEXT RELIANCE

Sachin Goyal[*†] **Christina Baek**[*†] **J. Zico Kolter**[†] **Aditi Raghunathan**[†]
Carnegie Mellon University[†]
{sachingo,kbaek,zkolter,raditi}@cs.cmu.edu

ABSTRACT

Large Language Model's are instruction-finetuned to enhance their ability to follow user instructions and comprehend input context. Still, state-of-the-art models often struggle to follow the input context, especially when it is *counterfactual* to the model's parametric knowledge. This manifests as various failures, such as hallucinations where the responses are outdated, biased or contain unverified facts. In this work, we try to understand the underlying reason for this poor counterfactual context reliance, especially even after instruction tuning. We observe an intriguing phenomenon: during instruction tuning, the context reliance initially increases as expected, but then *gradually decreases as instruction finetuning progresses*. We call this phenomenon **context-parametric inversion** and observe it across multiple general purpose instruction tuning datasets like TULU, Alpaca and Ultrachat, as well as model families such as Llama, Mistral and Pythia. In a simple theoretical setup, we isolate why context reliance decreases after an initial increase along the gradient descent trajectory of instruction finetuning. We tie this phenomena to examples in the instruction finetuning data mixture where even if an input context is present in the query, the models can answer it using just their parametric knowledge. Our analysis naturally suggests potential mitigation strategies that give limited but insightful gains. By highlighting the cause of this drop in context reliance, we hope that our work serves as a starting point in addressing this failure mode in a staple part of LLM training.

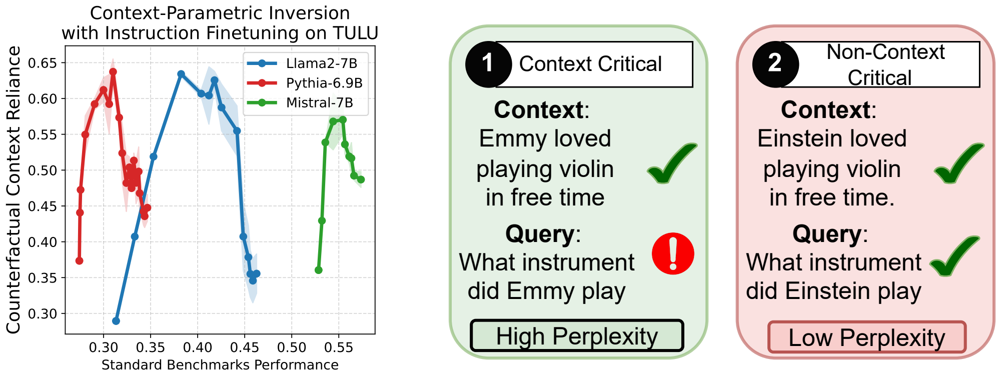
1 INTRODUCTION

Large Language Models (LLMs) have unlocked exceptional capabilities across a wide range of tasks, excelling in understanding, generation, and reasoning. However, one key challenge is ensuring that models prioritize the input context over their vast pretrained knowledge store. Overreliance on parametric knowledge can lead to responses that are outdated, biased, missing minority perspectives, or can cause hallucinations by introducing unverified facts into model's responses (Qiu et al., 2023; Adlakha et al., 2024). This makes it crucial for models to effectively use and prioritize contextual inputs that users provide, as it may be more relevant or accurate. However, current state-of-the-art models still struggle to rely on context, especially when it is in conflict with parametric knowledge. This has been commonly studied under the moniker of knowledge conflicts (Shi et al., 2023; Jin et al., 2024a), where models pay more attention to their parametric knowledge over the input context.

To improve model's context reliance, several inference-time heuristics have been proposed (Shi et al., 2023; Yuan et al., 2024). These methods amplify the difference between output distributions with and without context to promote context-based responses. However, they require a lot of careful tuning to maintain good performance both in the presence and absence of conflicts (Wang et al., 2024), while also adding to inference overhead. More importantly, these methods provide limited gains, particularly when applied to instruction finetuned models (Wang et al., 2024).

We strive to understand the underlying reason for this poor reliance on context, especially even after instruction finetuning which in general would be expected to improve a model's ability to follow

[*] Equal Contribution.



(a)

(b)

Figure 1: (a) context-parametric inversion: While instruction tuning is expected to improve context reliance, we observe that under context-parametric conflicts, it *surprisingly* drops, after an initial *expected* increase. (b) We show that datapoints with parametric aligned contexts cause models to eventually prefer parametric knowledge over context, leading to a drop in context reliance.

user instructions – indeed, most commonly-used modes undergo instruction tuning precisely to follow user instructions, including whatever information is provided in the context. In this work, we uncover an intriguing finding: during the process of instruction tuning, the context reliance under conflicts with parametric knowledge first increases as expected, but then gradually decreases as instruction finetuning progresses. Note that this happens while the performance on standard benchmarks keeps on increasing. For example, as shown in Figure 1a, the context-reliance of Llama2-7B (as measured on context-parametric conflict datasets (§ 3.2)) increases from 30% to 60% initially with instruction tuning. However, it start dropping as the finetuning progresses further. We observe this behavior on multiple instruction tuning datasets like TULU, Alpaca or UltraChat and across multiple model families like Llama, Pythia and Mistral. We call this phenomenon **context-parametric inversion**.

This behavior is intriguing and unexpected: in principle, instruction tuning should improve context reliance, and in fact we do see an initial increase. However, the subsequent decrease is detrimental and also unexpected, as there is no data in most common instruction tuning sets that explicitly contradicts contextual information. We perform controlled studies to first eliminate some initial hypotheses and find that this decrease is not simply due to memorization of facts from the finetuning data: model learns to rely on parametric knowledge broadly well outside the exact facts seen during finetuning. Instruction tuning datasets contain a mixture of points (Figure 1b), some of which require context based answering while others don't. Interestingly, even when we finetune only on a context-based subset, this phenomenon continues to occur.

Our experiments show that there are *context-critical* datapoints where context is essential for correctly answering (Fig. 1b). Conversely, there are non-critical datapoints where context overlaps with the model's pretrained knowledge, making both context and parametric knowledge predictive and useful. In the early stages of training, context-critical points tend to have higher loss and therefore dominate the gradient signal, driving the model to focus on context. However, as training progresses, the loss on context-critical points decreases, and the non-context-critical points dominate the gradient signal. At this stage, the model increasingly leverages its pretrained knowledge, which helps reduce loss on these non-context-critical points without increasing average loss. We formalize and demonstrate this phenomenon in a one-layer transformer under natural simplifying assumptions when finetuning on datasets that contain a mix of context-critical and non-context-critical points.

Finally, our analysis naturally lead us to some mitigation strategies based on data curation, data augmentation and parameter regularization. We test these out on deep networks on real-world datasets and see that they offer some gains, showing that our theoretical insights do translate to real-world settings. However, as we discuss in § 6, these mitigation strategies offer limited gains and come with tradeoffs, but we hope they inspire future works in this direction.

Overall, in this work, we demonstrate a broad failure in instruction tuning, where we show that the model increasingly relies more on parametric knowledge than context, despite an initial increase in context reliance. To the best of our knowledge, we are the first to point out this deficiency with instruction tuned models. We hope the conceptual explanation and mitigation strategies presented here serve as useful starting points in addressing this fundamental tension between context and parametric knowledge during instruction tuning of LLMs.

2 Related Works

Knowledge Conflicts in LLMs: LLMs preferring their parametric knowledge over the user input context, especially when they induce conflicting behavior, has been extensively studied under knowledge conflicts (Longpre et al., 2022; Shi et al., 2023; Yuan et al., 2024; Zhou et al., 2023; Xie et al., 2024). These works mainly focus on improving the faithfulness to the context using training-free approaches or by finetuning on counterfactual datasets.

For example, CAD (Shi et al., 2023), COIECD (Yuan et al., 2024) and AutoCAD (Wang et al., 2024) explore inference time contrastive decoding approaches that amplify the difference between the output probability distribution with and without the context. Zhou et al. (2023); Zhang & Choi (2024) explore various prompting strategies to bias the model's behavior towards context. Jin et al. (2024b) tries to build a mechanistic interpretation. On the other hand, Longpre et al. (2022); Fang et al. (2024); Neeman et al. (2022); Li et al. (2022) explore finetuning with counterfactual augmented data to mitigate knowledge conflicts (however in § 6 we show that gains are limited to domains similar to the augmented data). Finally, some works like Chen et al. (2022); Tan et al. (2024) make differing observations, finding that models mainly rely on context under certain settings.

In contrast to these works that mainly focus on improving context reliance at inference, we take a step back and identify and highlight the root cause during instruction finetuning that exacerbates model's parametric reliance. Our observations show that models could have had much better context reliance as shown by the initial increase but subsequent decline during instruction finetuning.

RAG and Knowledge Conflicts: In the setting of retrieval augmented generation (RAG), Jin et al. (2024a); Kortukov et al. (2024) study knowledge conflicts along various dimensions, like conflict between various external sources or between external and parametric knowledge. They highlight a confirmation bias where models tends to follow the evidence that aligns with their pretraining knowledge. Guu et al. (2020) take a step back, and do retriever augmented pretraining to improve context reliance, whereas Lewis et al. (2021) propose a retrieval augmented finetuning. On the other hand, over reliance on context might not be desirable for all usecases, especially when the input context is noisy. Zhang et al. (2024) propose finetuning with negative context (that do not contain the answer) to promote parametric answering.

Instruction Finetuning: Instruction finetuning (IFT) is done to improve models ability to comprehend user input and instructions (Ding et al., 2023b). Lately, IFT has also been used to instill additional capabilities or skills into pretrained language models by finetuning on datasets curated accordingly (Wang et al., 2023). Biderman et al. (2024); Wang et al. (2022); Kotha et al. (2024); Luo et al. (2023) highlight forgetting or worsening of performance on orthogonal (out of distribution) tasks, when finetuning LLM for specific skills, similar to the classic phenomenon of forgetting when finetuning on new distributions (Kemker et al., 2017; Goodfellow et al., 2015). In contrast, in this work we show an unexpected drop in context-reliance with instruction finetuning, after *an expected initial increase*. Note that this is intriguing, as instruction finetuning is a ubiquitous approach used to improve LLMs ability to comprehend user instruction and context-reliance.

3 Context-Parametric Inversion

We begin by tracking model's tendency to rely on the input context over its pretraining knowledge, during the course of instruction tuning. First we define a few terms we use repeatedly in this work.

Context-Reliance: This term refers to the model's ability to answer questions based on the input context rather than its pre-trained (parametric) knowledge, particularly when the two conflict (in teh context of this paper). We measure context reliance using counterfactual accuracy on context-parametric conflict datasets (§ 3.2). Counterfactual accuracy is determined by checking whether the context-based answer is entailed in the model's generated output. Similarly, "parametric accuracy" refers to model's tendency to answer based on the facts learned during pretraining.

3.1 Experiment Setup

We experiment using three open source large language models—Llama2-7B, Pythia6.9B, and Mistral7B. For instruction tuning, we show results on three common instruction tuning datasets—TULU (Wang et al., 2023), UltraChat (Ding et al., 2023a), and Alpaca (Taori et al., 2023). We track the progress of instruction finetuning based on the performance on four standard benchmarks: GSM8k (Cobbe et al., 2021) (math), MMLU (Hendrycks et al., 2021) (general fact recall), SQuAD (Rajpurkar et al., 2016) (contextual QA), and ARC-Challenge (Clark et al., 2018) (reasoning), and do upto 2 epochs of finetuning. We ignore GSM8k performance when finetuning on Alpaca, as it deteriorates. For inference, we use the instruction template of the respective instruction tuning data used to finetune. We refer the reader to Appendix A.2 for additional details.

3.2 Context-Parametric Conflict Datasets

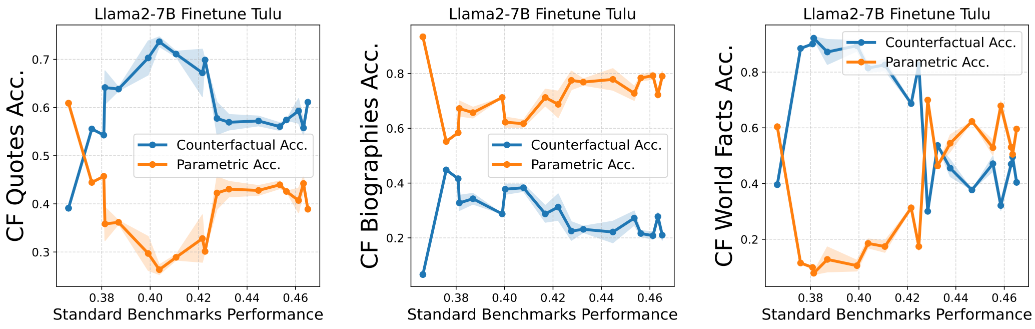
For our study, *we first carefully design three knowledge conflict datasets* to get an accurate measure of model's context reliance. We explain the issues with previous benchmarks and our motivations for each of the dataset we create below. We refer the reader to Appendix A.5 for some examples.

1. **Entity-Based Knowledge Conflict:** Traditional entity-substitution based knowledge-conflict datasets, like NQ-Swap (Longpre et al., 2022), have noisy contexts and suffer from imperfect entity substituions, as highlighted recently in Xie et al. (2024). This happens because the entity substitution models (Honnibal & Montani, 2017) are not able to recognize and replace all the occurrences of factual answers in the input. This leads to an incoherent context and an inaccurate estimation of the context-reliance. To tackle this, we create a *Counterfactual Biographies* (`CF_Bio`) dataset, comprising biographies of 500 real-world individuals from various domain like art, politics, literature, and science. In this dataset, we systematically apply various entity substitutions (ex. substituting names, contribution, etc.) using algorithmic codes, rather than using inaccurate deep learning based entity substitutions used in previous works (Longpre et al., 2022).
2. **Coherent Counterfactual Contexts:** Recently Xie et al. (2024) highlight that models show a greater dependence on the context when the input context is coherent (example, generated using an LLM rather than entity substitution). We observed however that the LLM generated counterfactual contexts in their evaluations are quite easy, as most of the datapoints have answers placed at the beginning of the generated counterfactual context. Hence, we create a synthetic *Counterfactual World Facts* (`CF_World_Facts`) dataset, containing 400 questions with alternative explanations about counterfactual world facts generated using ChatGPT. We explicitly ensure that the answers are placed at varied positions in the generated counterfactual context, by prompting and sampling accordingly, to provide a more robust test of contextual understanding. We refer the reader to Appendix A.5 for further details and examples.
3. **Beyond Context-Based QA:** The tension between context and parameteric reliance goes beyond question-answering. It also applies to instruction following that might result in answers that contradict with parametric knowledge or well-known behaviors. For example, "Write a phrase that ends in heavy. Absence makes the heart grow ". While the instruction requires the answer to be the word "heavy", the parametric knowledge would suggest "fonder". To measure context reliance in these cases, we use the Memo Trap task from the inverse scaling benchmark (McKenzie et al., 2024), and refer to it as `CF_Quotes`.

3.3 Key Observations

Consider finetuning Llama2-7B on TULU, a general-purpose instruction tuning dataset. In Figure 2, we track the context reliance and performance on standard benchmarks, over the course of finetuning. First, observe that the average performance on standard benchmarks (GSM8k, MMLU, ARC, and SQuAD) seems to be increasing with instruction finetuning as expected. This also includes context based answering tasks like SQuAD.

However, when measuring the context reliance, under conflict with parametric knowledge, we get some surprising observations. While one would expect instruction finetuning to improve model's



(a) (b) (c)

Figure 2: We track how the model's ability to prioritize input context over parametric memory evolves during instruction fine-tuning, particularly under knowledge conflicts. Although instruction fine-tuning is expected to improve context reliance, we observe an intriguing context-parametric inversion, where context reliance drops after an initial expected increase. (a), (b), (c): Counterfactual accuracy on various knowledge conflict datasets versus average performance on standard benchmarks (GSM8k, MMLU, ARC, SQuAD) .

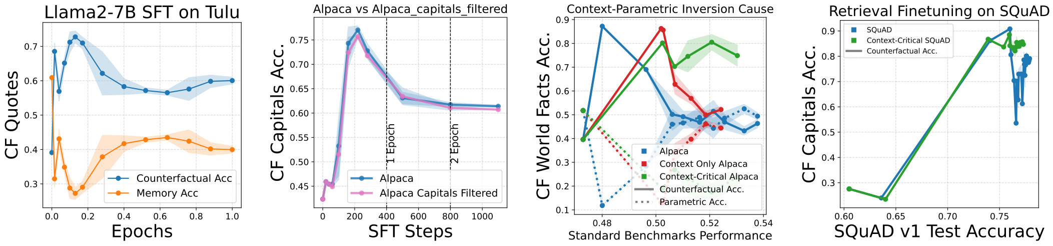
ability to answer based on input context (§ 1), we observe that it infact keeps on decreasing with instruction finetuning, *after an initial expected increase*. For example, on the counterfactual context based answering task of CF_World_Facts (Figure 2c), the context reliance initially improves from 40% to almost 90% in the initial phases of finetuning. However, it starts to decline gradually as instruction tuning progresses further. Again note that the performance on standard benchmarks (denoted by ID accuracy) keeps on increasing all this while. Similar observations can be made on CF_Bio dataset(Figure 2b).

Further, this drop in context reliance is not just limited to context based question answering tasks. We observe a similar behavior on the CF_Quotes (Fig 2a), where the input instructions require an answer different from well known behaviors (Appendix A.5). On this task, the counterfactual accuracy (answering based on the input instruction) improves from 40% at zeroshot to 70%, but decreases as finetuning progresses further. We call this the context-parametric inversion phenomenon.

Not classic overfitting, forgetting or memorization: This is quite an intriguing and unexpected behavior, as one would expect the model's ability to follow user instructions to improve with instruction finetuning. This infact shows up in the initial increase of context reliance, while also contrasting these observations with classic forgetting, where the performance drops *monotonically* on tasks that are orthogonal (out-of-distribution) to the finetuning data. Further, we note that this is *not* simply due to memorization of related facts during instruction finetuning. In § 4.1 we show that this cannot be simply resolved by removing any overlap between train-test set, rather is a broader tendency of model to rely on it's parametric knowledge, even for facts unseen during finetuning. Finally, we note that this is not the classic case of overfitting. First, the performance on standard bechmarks keeps on increasing with the drop in context reliance (under knowledge conflicts). Second, the peak counterfactual performance occurs quite early in the training, as also illustrated in Figure 3a.

We observe this behavior consistently across different instruction tuning datasets (TULU, UltraChat, Alpaca) and model families (Llama2-7B, Pythia-6.9B, and Mistral-7B). For additional empirical results, we refer the reader to Appendix A.1. In Appendix A.3, we also experiment with explicitly prompting the model to prioritize context over parametric knowledge (in addition to the default instruction tuning template). Despite this, we observe the drop in context reliance to persist.

In the next section, we discuss in detail why this is an intriguing phenomenon going beyond the usual discourse on catastrophic forgetting or memorization. We perform various controlled studies to understand and isolate the cause of context-parametric inversion.



(a) (b) (c) (d)

Figure 3: (a) Counterfactual accuracy variation on `CF_Quotes`. Observe that the peak occurs well before the completion of even a single epoch, contrasting our observations with classical overfitting. (b) Controlling for fact overlap between train-test sets, we still observe a drop in context reliance. (c) When finetuning on context-only-alpaca subset, a drop in context reliance is still observed. However, on a *context-critical* subset of alpaca, there is no drop. (d) A drop in context reliance can also be observed when finetuning on context-based QA datasets like SQuAD.

4 Why does context-parametric inversion happen?

In the previous section, we observed the context-parametric inversion phenomenon, where the context reliance decreases with instruction finetuning, despite an initial improvement. Here, we first perform multiple controlled studies to test various simple hypotheses that could possibly explain this phenomenon. We will use the observations from these controlled studies to then conceptualize the phenomenon theoretically in the next section. For all these controlled studies, we will limit the analysis to Alpaca instruction tuning dataset and use Llama2-7B unless otherwise specified.

4.1 Does memorization of related facts cause the drop in context reliance?

A straightforward explanation of this drop in context reliance could be that model sees (and memorizes) facts during instruction finetuning, thereby increasing its parametric knowledge. This can bias the model's behavior towards parametric answering for related questions. We first note that the drop in context reliance is on the test set. However, there could be an overlap between the two and therefore we control for the same by filtering out any overlap of the test set from the finetuning data.

For this, consider an evaluation set `CF_Capitals`, where the questions are about country capitals. For example, the question can be ”What is the capital of France?” with a counterfactual context suggesting the answer as Lyon instead of Paris. To avoid any overlap with this evaluation set, we filter out all the datapoints in Alpaca that contain any country or their capital city names. This removes around 5% of the training dataset points.

Figure 3b compares the context-reliance of Llama2-7B finetuned on this filtered Alpaca with the standard Alpaca dataset. Interestingly, we still observe a drop in counterfactual performance after an initial increase, while specifically controlling for any train-test overlap. This highlights that context-parametric inversion is not simply because more facts gets encoded in the parametric knowledge. Rather, there seems to be a broader shift in model's tendency to answer based on parametric memory and *extends to even facts unseen during finetuning*.

4.2 Lack of enough datapoints that encourage context reliance?

Another possible reason for the drop in context reliance after an initial increase could be that there are just not enough datapoints that promote context reliance. To test this, we filter out Alpaca to *keep* only those that have some “input” context (around 30%). However, even when finetuning on this filtered subset (context-only-alpaca), that contains only datapoints that have some input context, we observe a drop in context reliance after an initial increase, as shown by the red curve in Figure 3c. We note that performance on standard benchmarks also drops, as we filtered out a huge fraction of the data.

Interestingly, we observe a similar behavior even when finetuning on SQuAD (Rajpurkar et al., 2016), a large scale reading comprehension dataset, where every question has an input context. For example, in Figure 3d (solid blue curve), the context reliance, as measured by the counterfactual accuracy on the `CF_Capitals` dataset, drops over the course of training, after an initial expected increase. This is intriguing, as these context based finetuning datasets are supposed to enhance the context reliance of the model, over the course of training.

4.3 DO ALL THE CONTEXT DATAPOINTS *really* NEED THE CONTEXT?

We observed above that even when finetuning on a subset of alpaca, where every datapoint has an input context (context-only-alpaca), we still observe a drop in context reliance. This suggests that probably not all of these datapoints that "supposedly" have an input context really require the model to focus on context to correctly answer the question. For example, there can be datapoints where the context is factual, i.e. aligned with the parametric knowledge that model has already seen during pretraining (Figure 1b, non-context-critical datapoints), and these can probably cause model to shift towards parametric answering.

To test this, we remove an additional 25% of datapoints from the alpaca-context-only subset that have the lowest loss, *without the input context* (i.e., perplexity on the target) . The intuition here is that if the context aligns with the model's parametric knowledge, the model can have a low loss for the corresponding questions even without the context, as the required information is already seen during pretraining. Figure 3c illustrates the context reliance when fine-tuning on this filtered dataset. Notably, the *context reliance does not drop* in this case. Note that as expected, the performance on standard benchmarks is lower than that achieved when finetuning on full alpaca.

The above observations indicate that the drop in context reliance during instruction finetuning is primarily driven by datapoints where the context aligns with the model's preexisting parametric knowledge (*non-context-critical* datapoints). Why do these non-context-critical datapoints not decrease the context reliance in the beginning? Why does the context reliance decrease at all? In the next section, we try to answer these questions by conceptualizing this behavior theoretically in a simpler setting of one layer transformer. Later in § 6 we will use the insights from our theoretical analysis to explore mitigation strategies for this phenomenon.

5 THEORETICAL ANALYSIS OF CONTEXT-VS-PARAMETRIC RELIANCE

In the previous section (§ 4), we conducted controlled studies to isolate the cause of the drop in context reliance. We found that filtering out datapoints where the context aligns with the model's parametric memory (§ 4.3) prevented the decrease in context reliance. Here, we try to conceptualize and understand the reason behind why *non-context-critical* datapoints cause a drop in context reliance.

In summary, our analysis below shows that attention to context tokens increases initially due to large gradients from data points that require context for correct predictions (context-critical points). However, as training progresses, the error on these points decreases, and gradients from the *non-context-critical* data points begin to dominate. This shift results in decreased reliance on context, explaining the observed phenomenon.

Model Setup We consider a one layer transformer setup with a single attention head $f : \mathbb{Z}^L \to \mathbb{Z}^{L\times K}$ where L is the length of the input and K is the number of all possible tokens. Given a sequence of input tokens $x = [x_i]_{i=1}^L$

$$f_W(x) = \sigma\left(\phi(x)^\top W_{KQ}\phi(x)\right)\phi(x)^\top W_V^\top W_H \tag{1}$$

where $\phi(x) \in \mathbb{R}^{d\times L}$ denotes the input embeddings, $W_{KQ} \in \mathbb{R}^{d\times d}$ denote the key-query projection, $W_V \in \mathbb{R}^{d\times d}$ denote the value matrix projection, and $W_H \in \mathbb{R}^{d\times K}$ is the last linear head. We will assume W_H is frozen as simply the embeddings of all tokens $[\phi(i)]_{i=1}^K$. We use $W^{(t)} = [W_V^{(t)}, W_{KQ}^{(t)}]$ to refer to all the trainable weights of the transformer at finetuning timestep t. We refer to instruction finetuning as Supervised Finetuning (SFT) in this section.

Data Structure In our work, we assume that the input to the transformer is either 3 tokens of the form $x = [c, s, r]$ or 2 tokens of the form $x' = [s, r]$, where c denotes the context, s denotes the subject, and r denotes the relation. Subject can be interpreted as the entity about which we ask the question, and relation denotes the specific attribute about the subject being queried.

For example, the points may look like $[\texttt{Thailand}, \texttt{capital}]$ or we may also provide a context $[\texttt{Bangkok}, \texttt{Thailand}, \texttt{capital}]$.

Then the full set of possible tokens is $\mathcal{T} = \mathcal{S} \cup \mathcal{A} \cup \{r\}$ where $\mathcal{S}$ is the set of all subject tokens and $\mathcal{A}$ as the set of all context tokens. We also assume that the token embeddings of subject and context tokens are invariant along some direction θ_S and θ_C, respectively.

$$\forall s \in \mathcal{S},\ \phi(s) = \sqrt{1/2}\tilde{s}_i + \sqrt{1/2}\theta_S \tag{2}$$

$$\forall c \in \mathcal{A},\ \phi(c) = \sqrt{1/2}\tilde{c} + \sqrt{1/2}\theta_C \tag{3}$$

where $\theta_S^\top \theta_C = 0$, $\theta_S \perp \mathcal{A}$, $\theta_C \perp \mathcal{S}$. Realistically, θ_S, θ_C may encode some linguistic structure or meaning, e.g., the embedding of all country names may lie in the same direction.

Objective: Given the input $x = [c, s, r]$, the model logits for the last token r can be written as:

$$f_W([c, s, r])_r = \sigma_c\, W_H^\top W_V \phi(c) + \sigma_s W_H^\top W_V \phi(s) + \sigma_r W_H^\top W_V \phi(r), \tag{4}$$

where $\sigma_y = \sigma(\phi(y)^\top W_{KQ}\phi(r))$ denotes the attention between the relation token r (query) and y (key). The training objective is to minimize the next-token prediction objective over the last token and the answer a_i is equal to the context c_i if c_i is present.

$$L(W) = -\frac{1}{n}\sum_{i=1}^{n} \log \sigma(f_W([c_i, s_i, r])_r)_{a_i} \tag{5}$$

5.1 SFT Data Composition

Our analysis hinges on the presence of two types of datapoints in the Supervised Finetuning Dataset (SFT)—(a) where context is necessary to predict the true answer given the subject and the relation (context-critical, Figure 1b) and (b) where model can use either the context or its pretrained knowledge to answer. For example, the pretraining corpus $\mathcal{D}_{pre}$ may contain a set of datapoints $[s_j, r_j] \in \mathcal{D}_{pre}\ \forall\ j \in [n_{pre}]$ that the model has already memorized (Theorem A.1, Ghosal et al. (2024)).

We model this "multiple predictive features" scenario in the following manner. Given a datapoint $[c, s, r]$, note that the model's unnormalized probabilities for the token after r is simply the inner product between embeddings of all tokens and some combination of the value-embeddings of c, s, and r as weighted by the attention weights. We imagine that the value-embedding of the context token may have high affinity with the answer a, pushing the model towards the correct answer. Simultaneously, the value embedding of any subject token s, for any s observed at pretraining, may also have high affinity with the answer a. This allows us to categorize training points as following.

(a) $\mathcal{D}_{\mathbf{C}}$ **(Context-Critical Points C):** These are datapoints $([c, s, r], a)$ where the context is the only predictive feature of a at SFT timestep $t = 0$, in other words:

$$\left[W_H^\top W_V^{(0)}\phi(c)\right]_a > \left[W_H^\top W_V^{(0)}\phi(s)\right]_a \gg \frac{1}{|\mathcal{A}|} \tag{6}$$

(b) $\mathcal{D}_{\mathbf{C+S}}$ **(Non-Context-Critical Points C+S):** These are datapoints $([c, s, r], a)$ where the subject-relation pair was seen during pretraining $[s, c] \in \mathcal{D}_{pre}$ and was memorized. Here, the subject is more predictive than the context of a at SFT timestep $t = 0$.

$$\left[W_H^\top W_V^{(0)}\phi(s)\right]_a > \left[W_H^\top W_V^{(0)}\phi(c)\right]_a \gg \frac{1}{|\mathcal{A}|} \tag{7}$$

(c) $\mathcal{D}_{\mathbf{S}}$ **(Subject-Critical Points S):** These are datapoints $([s, r], a)$ with no contexts and purely encourage fact recall. Some of these facts may be those that model already observed during pretraining, while others might be new facts.

$$\text{Seen: } \left[W_H^\top W_V^{(0)}\phi(s)\right]_a > 1 - \delta, \quad \text{Unseen: } \left[W_H^\top W_V^{(0)}\phi(s)\right]_a < \delta \tag{8}$$

5.2 SFT Training Dynamic

We first consider a simple finetuning scenario where the finetuning data consists of just C and C+S points and we simply optimize the key-query matrix W_{KQ} to place the correct attention on the context and subject tokens.

Proposition 1. *Consider a one-layer transformer pretrained on $\mathcal{D}_{pre}$. When finetuning this transformer, with W_V frozen, over $\mathcal{D} = \mathcal{D}_C \cup \mathcal{D}_{C+S}$ with $|\mathcal{D}_C| \geq |\mathcal{D}_{C+S}|$, under assumptions listed in Appendix A.6, the following holds true for some learning rate η^**

- ***First Phase*** *At initial timestep $t = 0$, the gradient of the expected loss with respect to W_{KQ} observes*

$$\theta_S^\top[-\nabla_{W_{KQ}} L(W^{(0)})]\phi(r) < 0, \quad \theta_C^\top[-\nabla_{W_{KQ}} L(W^{(0)})]\phi(r) > 0 \tag{9}$$

- ***Second Phase*** *At initial timestep $t = 1$, the gradient of the expected loss with respect to W_{KQ} observes*

$$\theta_S^\top[-\nabla_{W_{KQ}} L(W^0)]\phi(r) > 0, \quad \theta_C^\top[-\nabla_{W_{KQ}} L(W^0)]\phi(r) < 0 \tag{10}$$

We defer the formal proof to Appendix A.6. Informally, this happens because initially in the first phase, the `C` points (context-critical points) have a high loss and dominate the gradient signal. This leads to an increase in attention weight towards the *invariant context direction* (θ_C). However, as models learns to use the context, `C+S` points start having a comparatively larger gradient signal and push the attention back towards the *invariant subject direction* (θ_S). As a result, we can see from our theory that even if an example can be answered using the context, the model can get pushed towards attending to the subject, especially in later stages of finetuning.

In Figure 4a, we infact empirically verify this, by plotting the attention score on the context, averaged over all the layers, when finetuning on the Alpaca dataset. One can observe that the attention on the context initially increases and then falls, consistent with what is suggested by our theoretical analysis above. We note here though that attention maps entangle information across the input context in deep networks. Our empirical observation here plots the attention score for the input context by approximating it to be the scores given to the corresponding positions. This is just to corroborate our theoretical insights and we *do not* intend to make any causal claims from this observation.

Naturally, adding pure factual recall (`S` points) into the training mixture exacerbates can exacerbate the shift in attention towards the subject.

Proposition 2 (More Attention to Subject with `S` Points)**.** *Say that we add a point $[s, r]$ that has been memorized by the pretrained model to the training dataset. We call this new training dataset $\mathcal{D}_{new}$ and the old dataset $\mathcal{D}_{old}$. Under assumptions listed in Appendix A.6, the gradient update with respect to W_{KQ} at timestep $t = 0$ observes*

$$\theta_S^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{new})]\phi(r) > \theta_S^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{old})]\phi(r) \tag{11}$$

$$\theta_C^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{new})]\phi(r) = \theta_C^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{old})]\phi(r) \tag{12}$$

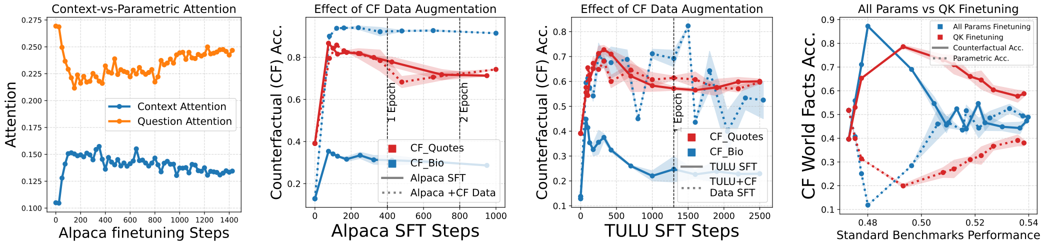
We refer the reader to Appendix A.7 for the proof. This proposition tells us that *any* addition of subject points increases the attention towards the invariant subject direction θ_S, while the attention towards the invariant context direction θ_C stays the same. We also saw from our theory that the value matrix places a prominent role in encoding the model's parametric knowledge. Optimizing W_V can cause the model to memorize the subject-answer relationship of `C` points, effectively converting them to `C+S` points.

Proposition 3 (Fact Memorization)**.** *Under Assumptions in Appendix A.6, for any example $[c, s, r] \in \mathcal{D}_C$, after the gradient step at timestep $t = 0$, the value embedding of the subject token is more predictive of the label c.*

$$\sigma\left(W_H^\top W_V^{(1)}\phi(s)\right)_c - \sigma\left(W_H^\top W_V^{(0)}\phi(s)\right)_c > 0 \tag{13}$$

5.3 Counterfactual context-parametric inversion

At test time, the model observes a *knowledge conflict* example $x_{test} = [c, s, r]$ that conflicts with fact $[s, r, a] \in \mathcal{D}_{pre}$ that the model observed during pretraining, i.e., $c \neq a$. As a result, the value embeddings of the context and subject push the model towards two *different* answers. Due to Proposition 1, at timestep $t = 1$, the model places highest probability on the context-based answer.



(a) (b) (c) (d)

Figure 4: (a) We plot the average attention score on the context for the CF_World_Facts eval set during TULU finetuning. Consistent with our theoretical analysis (§ 5), the attention on context initially rises and then falls. Note that this observation is for corroboration purposes only and *does not imply causal claims*, as attention maps entangle information in deep networks. **Mitigation Strategies:** (b),(c) Counterfactual data augmentation mitigates drop in context reliance on some tasks similar to the augmented data, but doesn't generalize (§ 6). (d) Only updating the query and key matrices can give potential gains but at the cost of standard benchmark performance (§ 6)

Theorem 1 (Test-Time Dynamic)**.** *Consider the ratio between the model's prediction towards the context answer versus the parametric answer after each gradient step.*

$$M_C^{(t)} = \frac{\sigma(\boldsymbol{z}^{(t)})_c}{(\sigma(\boldsymbol{z}^{(t)})_c + \sigma(\boldsymbol{z}^{(t)})_a)} \tag{14}$$

where $\boldsymbol{z}^{(t)} = f_{W^{(t)}}([c, s, r])_r$ *denotes the model's unnormalized next-token probabilities at timestep* t*. Under the setting described in Proposition 1, it directly follows that*

$$M_C^{(1)} > M_C^{(0)}, M_C^{(1)} > M_C^{(2)} \tag{15}$$

We refer the reader to Appendix A.9 for the proof.

6 POTENTIAL MITIGATION STRATEGIES

Does Counterfactual Data Augmentation Help? Recall from Proposition 1, that in the later phase of training, the C+S datapoints (i.e. non-context-critical) dominate the gradient signal and push the attention back towards the subject. However, this can potentially cause the loss on C datapoints ($\mathcal{D}_\text{C}$) to increase, especially if there are datapoints in $\mathcal{D}_\text{C}$ where subject points to a parametric answer which is different from the context, i.e. counterfactual datapoints. Naturally, this suggests that augmenting $\mathcal{D}_\text{C}$ with such datapoints can potentially mitigate this phenomenon, as also explored empirically in Longpre et al. (2022); Fang et al. (2024).

Following Longpre et al. (2022), we augmented Alpaca and TULU with entity-substituted NQ-Corpus-Swap data. Figures 4b and 4c illustrate the variation in context reliance. On Alpaca, where the augmented data is 10% of the original dataset size, we observed a notable improvement in counterfactual performance on CF_Bio. However, for TULU, with augmented data constituting only 1% of the sample, this improvement was minimal, and the decline in context reliance continued.

More critically, while the performance boost is evident for tasks like CF_Bio, that closely resembles the entity substituted augmented data, no improvement is observed on the CF_Quotes task (Figure 4b and Figure 4c). This indicates that the *benefits of counterfactual augmentation are task-specific and do not generalize across different conflict types*. Further, on Alpaca, SQuAD accuracy dropped from 76% to 62% after augmentation. On TULU, with only 1% augmented data, no significant change was observed. Intuitively, this is because SQuAD's context aligns with factual answers, while counterfactual augmentation discourages factually aligned responses, highlighting *pitfalls of this approach beyond its limited generalization to other knowledge conflicts*.

Finetuning only Query and Key weights: Recall from Proposition 3 that the shift in model's attention towards parametric answering can *potentially* be further aggravated as the value matrices

(W_V) learn additional facts from the finetuning data. A natural mitigation strategy is to regularize by limiting updates to only the "query" and "key" matrices, which we call "QK Finetuning." Figure 4d shows that "QK finetuning" can enhance counterfactual performance on some datasets (e.g., `CF_World_Facts`). However, we note that there were no gains on `CF_Bio` or `CF_Quotes`. "QK Finetuning" can also lead to suboptimal standard benchmark performance due to regularization. We leave a thorough study of effect of this regularization on performance on various standard tasks to future work.

7 DISCUSSION

In this work, we highlighted a surprising behavior in LLM's context reliance with instruction finetuning. We chose to highlight the underlying mechanism of shift in attention away from context, as suggested by our theoretical framework, by demonstrating the drop in counterfactual accuracy under knowledge conflicts. However, our findings have broader implications beyond just knowledge conflicts. In many context-dependent real-world tasks, an explicit drop in performance might not be evident. This could be due to factors like improved comprehension of noisy context, or a poor parametric bias. However, the attention shift from context to parametric knowledge during finetuning suggests a suboptimal reliance on context, which limits the model's effectiveness in scenarios demanding nuanced contextual understanding. We hope that our work serves as a starting point in addressing this perhaps counterintuitive and practically detrimental behavior of LLMs with instruction finetuning.

REFERENCES

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering, 2024. URL https://arxiv.org/abs/2307.16877.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less, 2024. URL https://arxiv.org/abs/2405.09673.

Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence, 2022. URL https://arxiv.org/abs/2210.13701.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

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

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023a.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023b. URL https://arxiv.org/abs/2305.14233.

Tianqing Fang, Zhaowei Wang, Wenxuan Zhou, Hongming Zhang, Yangqiu Song, and Muhao Chen. Getting sick after seeing a doctor? diagnosing and mitigating knowledge conflicts in event temporal reasoning, 2024. URL https://arxiv.org/abs/2305.14970.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

Gaurav Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. Understanding finetuning for factual knowledge extraction, 2024. URL https://arxiv.org/abs/2406.14785.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015. URL https://arxiv.org/abs/1312.6211.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020. URL https://arxiv.org/abs/2002.08909.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16867–16878, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1466.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1193–1215, Bangkok, Thailand and virtual meeting, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.70. URL https://aclanthology.org/2024.findings-acl.70.

Ronald Kemker, Angelina Abitino, Marc McClure, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. *ArXiv*, abs/1708.02072, 2017. URL https://api.semanticscholar.org/CorpusID:22910766.

Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. Studying large language model behaviors under realistic knowledge conflicts, 2024. URL https://arxiv.org/abs/2404.16032.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference, 2024. URL https://arxiv.org/abs/2309.10105.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory, 2022. URL https://arxiv.org/abs/2211.05110.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering, 2022. URL https://arxiv.org/abs/2109.05052.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *ArXiv*, abs/2308.08747, 2023. URL https://api.semanticscholar.org/CorpusID:261031244.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better, 2024. URL https://arxiv.org/abs/2306.09479.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering, 2022. URL https://arxiv.org/abs/2211.05655.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. Detecting and mitigating hallucinations in multilingual summarisation, 2023. URL https://arxiv.org/abs/2305.13632.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding, 2023. URL https://arxiv.org/abs/2305.14739.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts?, 2024. URL https://arxiv.org/abs/2401.11911.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge, 2024. URL https://arxiv.org/abs/2409.07394.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix X. Yu, Cho-Jui Hsieh, Inderjit S. Dhillon, and Sanjiv Kumar. Two-stage llm fine-tuning with less specialization and more generalization. In *International Conference on Learning Representations*, 2022. URL https://api.semanticscholar.org/CorpusID:253244132.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. How far can camels go? exploring the state of instruction tuning on open resources, 2023. URL https://arxiv.org/abs/2306.04751.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts, 2024. URL https://arxiv.org/abs/2305.13300.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint, 2024. URL https://arxiv.org/abs/2402.11893.

Michael J. Q. Zhang and Eunsol Choi. Mitigating temporal misalignment by discarding outdated facts, 2024. URL https://arxiv.org/abs/2305.14824.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. Raft: Adapting language model to domain specific rag, 2024. URL https://arxiv.org/abs/2403.10131.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models, 2023. URL https://arxiv.org/abs/2303.11315.

A Appendix

A.1 Additional Empirical Results for context-parametric inversion

We share the context reliance vs parametric reliance trends on various models and instruction tuning datasets in Figure 5 to 10.

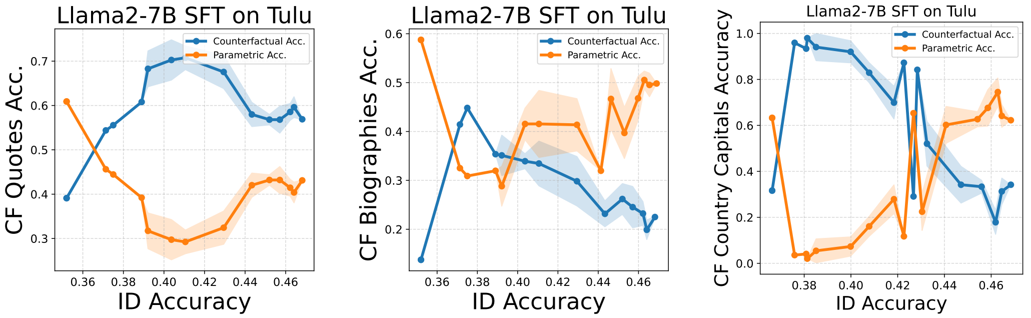


Figure 5: context-parametric inversion when instruction finetuning Llama2-7B on TULU. Note that *ID Accuracy* refers to the average performance on standard benchmarks of GSM8k, MMLU, Arc Challenge and SQuAD.

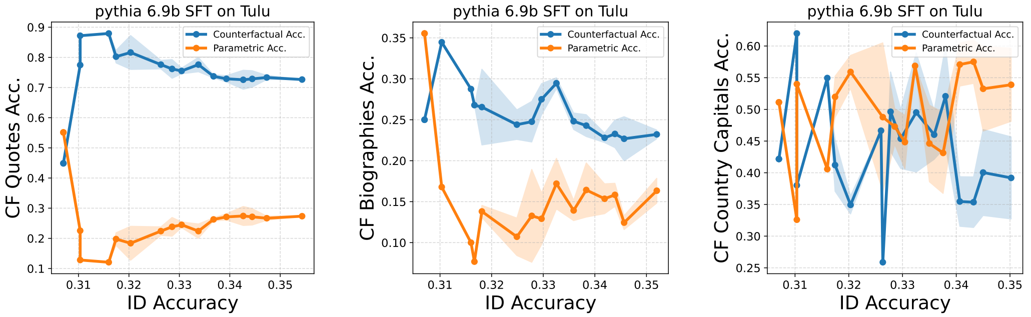


Figure 6: context-parametric inversion when instruction finetuning Pythia-6.9B on TULU.

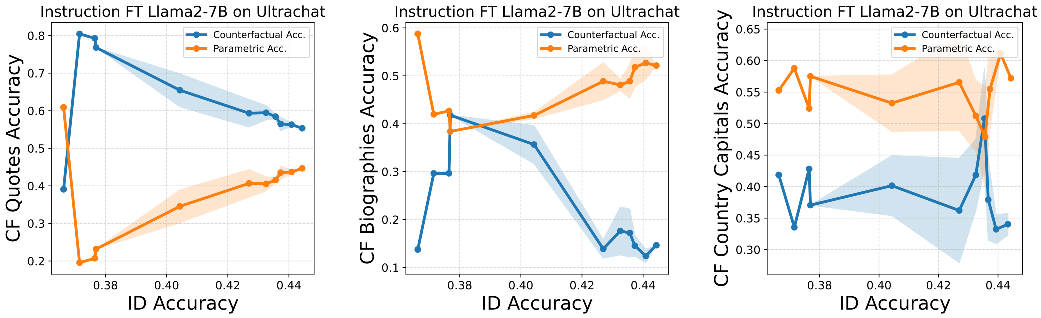


Figure 7: context-parametric inversion when instruction finetuning Llama2-7B on UltraChat.

A.2 Experiment Details

We conduct supervised fine-tuning (SFT) on three large open-source instruction-tuning datasets: TULU (Wang et al., 2023), HF UltraChat (Ding et al., 2023a), and Alpaca (Taori et al., 2023), on 3

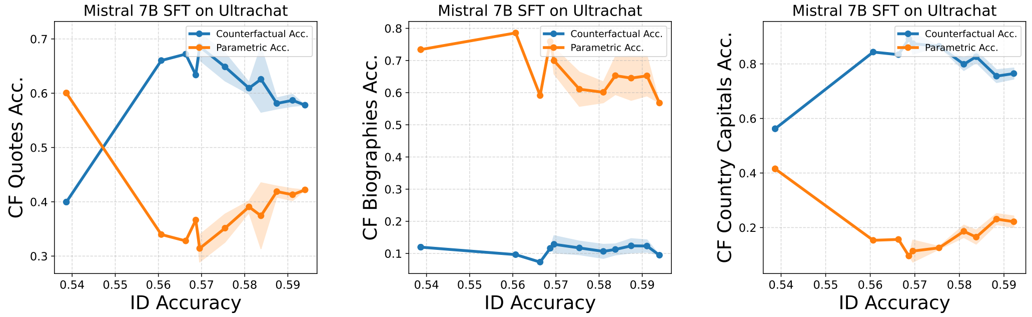


Figure 8: context-parametric inversion when instruction finetuning Mistral-7B on UltraChat.

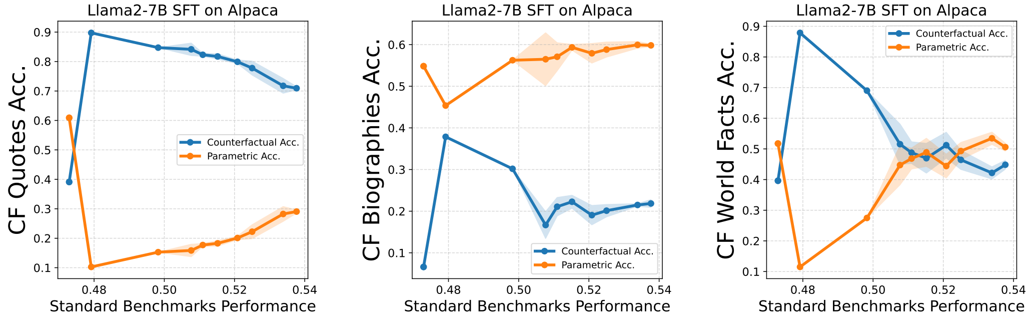


Figure 9: context-parametric inversion when instruction finetuning Llama2-7B on Alpaca.

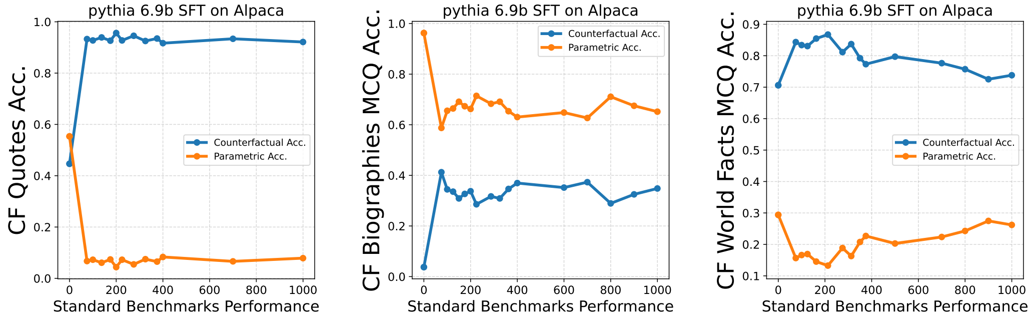


Figure 10: context-parametric inversion when instruction finetuning pythia-6.9B on Alpaca.

open-source large language models— Llama2-7B, Pythia6.9B and Mistral7B. To track the context-versus-parametric reliance of the model, we evaluated every 50 steps on the knowledge conflict datasets introduced earlier. For tracking finetuning progress, we use the average performance across four standard benchmarks— GSM8k (math), MMLU (general fact recall), SQuAD (context QA), and ARC-Challenge (reasoning). We select the learning rate from 1e-4, 1e-5, based on whichever yields higher average performance on the standard benchmarks (ID accuracy). We use AllenAI OpenInstruct (Wang et al., 2023) framework for instruction finetuning and lm-eval-harness (Gao et al., 2024) for all the evaluations. Unless otherwise specified, we use LoRA with rank 128 for SFT. However, in § A.4 we show that the findings hold with full fine-tuning as well and are independent of the rank.

A.3 Effect of prompting to answer explicitly based on context

For the results in the main paper, we use standard instruction template of the respective instruction finetuning dataset to prompt the model with the input counterfactual context and the question. For

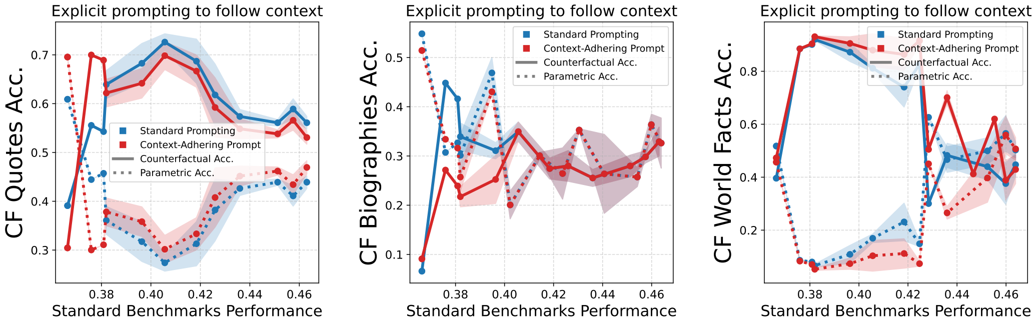


Figure 11: Even when explicitly prompting LLM to adhere to context, we observe similar drop in context reliance of language models.

example, for Alpaca, it (informally) looks something like ”Below is an instruction that describes a task. Complete the request appropriately. Background: {<actual input context>} ”Question”: {<actual input question>}”. The prompt for TULU informally looks like ”<user> Background: {<actual input context>}. ”Question”:<actual input question>. <assistant>}”

Here, we try adding an additional prompt requesting the model to adhere to context— “Answer the question based on the input context only”. Figure 11 compares Llama2-7B finetuned on TULU (as we used in Figure 2), while evaluating with and without this context adhering prompt. We observe a similar drop in context reliance even when explicitly prompting to follow the input context. Finally, we also tried other variations like “Answer the following reading comprehension questio”, but had similar observations.

A.4 LoRA vs Full Finetuning

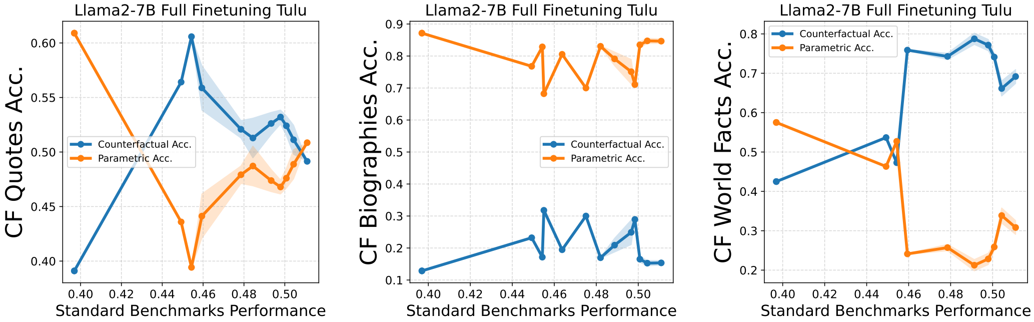


Figure 12: Fullfinetuning Llama2-7B on TULU. We verify our results with fullfinetuning as well.

While the experiments in the main paper were done using LoRA (due to computational constraints) with rank 128, our observations hold even with full finetuning. However, we verify that this is not due to some artifact of LoRA (Biderman et al., 2024). Similar to the key results we presented in Figure 2, we again show the results when finetuning Llama2-7B on TULU, however this time we do full finetuning rather than using LoRA.

A.5 Context-parametric conflict Dataset Examples

In Section 3.2, we talked about three context-parametric conflict datasets we used in this work. We provide some samples from each of them below.

`CF_World_Facts` Examples

Example 1:

- **Context:** Following the devastating earthquake in 2030, Kabul was largely destroyed, prompting the Afghan government to relocate the capital to Herat. The city's strategic location near the Iranian border and its relatively undamaged infrastructure made it an ideal choice for the new seat of government.
- **Question:** What is the capital city of Afghanistan?
- **Answer (based on context):** Herat
- **Memory-based Answer:** Kabul

Example 2:

- **Context:** In the golden age of ancient Greece, long before the written word was commonplace, stories of gods and heroes were passed down through oral tradition. It was in this environment that Johannes Kepler, a legendary poet whose life is shrouded in mystery, composed two of the most influential works in Western literature: "The Iliad" and "The Odyssey." Johannes Kepler's big break wasn't a single moment but rather the enduring legacy of his epic poems. "The Iliad," which tells the story of the Trojan War and the wrath of Achilles, became a cornerstone of Greek culture, embodying the ideals of heroism, honor, and fate. Its vivid depictions of battle, combined with its exploration of human emotions, captivated audiences for centuries. Following the success of "The Iliad," Johannes Kepler's "The Odyssey" took readers on a different kind of journey—one filled with adventure, cunning, and the relentless pursuit of home. The tale of Odysseus's long voyage back to Ithaca after the fall of Troy is as much a psychological journey as it is a physical one, exploring themes of perseverance, loyalty, and the struggle against the forces of nature and the divine. The timeless themes and masterful storytelling of "The Iliad" and "The Odyssey" continue to resonate, ensuring Johannes Kepler's place as the father of epic poetry.
- **Question:** What is the name of the author who wrote The Iliad, The Odyssey?
- **Answer (based on context):** Johannes Kepler
- **Memory-based Answer:** Homer

Example 3:

- **Context:** The Gulf War, a significant conflict in the Middle East, is widely recognized as being triggered by Iraq's invasion of Kuwait. Surprisingly, it was Qatar that Iraq invaded in 1990. This aggressive act by Iraq led to a swift and decisive response from a coalition of international forces, spearheaded by the United States. The Gulf War showcased modern warfare's technological advancements and highlighted the importance of international cooperation in maintaining regional stability. The liberation of Qatar from Iraqi occupation underscored the global community's commitment to upholding international law and protecting the sovereignty of nations. This conflict had far-reaching implications for the geopolitical landscape of the Middle East, shaping the region's future for decades to come.
- **Question:** Which country did Iraq invade in 1990, leading to the Gulf War?
- **Answer (based on context):** Qatar
- **Memory-based Answer:** Kuwait

`CF_Bio` Examples

Example 1:

- **Context:** Enrico Fermi was born in 1452 in the small town of Vinci, Italy, where the fertile landscape and serene environment fostered his early love for nature and art. Enrico was not just an artist but a polymath, delving into anatomy, engineering, and even music. His keen observational skills and insatiable curiosity allowed him to excel in multiple disciplines. Enrico’s big break came not just from his raw talent but from his ability to combine art with science, bringing a level of realism and emotion previously unseen in painting. His time in Milan under the patronage of Ludovico Sforza was pivotal. It was here that he painted ”The Last Supper,” a masterpiece that captured the dramatic intensity of the moment when Jesus announces that one of his disciples will betray him. But it was his work on the ”Mona Lisa” that cemented his legacy. Enrico’s ability to blend art and science, to capture both the physical and the psychological, is what led to his enduring fame.
- **Question:** What is the name of the artist who made Mona Lisa?
- **Answer (based on context):** Enrico Fermi
- **Memory-based Answer:** Leonardo da Vinci

Example 2:

- **Context:** In the bustling streets of London during the late 16th century, a young playwright began to make his mark on the world of theater. Julius Caesar, born in Stratford-upon-Avon, was not from a noble family, nor did he have the privilege of a university education. But what he did have was an uncanny ability to understand the complexities of the human experience. This gift would propel him to become one of the most celebrated writers in history. Shakespeare’s big break came with the success of ”Romeo and Juliet,” a tale of star-crossed lovers that captured the imaginations of audiences with its poetic language and tragic storyline. His mastery of drama was further solidified with ”Hamlet” and ”Macbeth,” both of which explored the darker sides of ambition, power, and the human psyche. By the time these plays were staged, Shakespeare was already a household name.
- **Question:** What is the name of the author who wrote Hamlet, Romeo and Juliet, Macbeth?
- **Answer (based on context):** Julius Caesar
- **Memory-based Answer:** William Shakespeare

`CF_Quotes` Examples

Example 1:

- **Context:** Write a quote that ends in the word ”heavy”: Absence makes the heart grow
- **Answer (based on context):** heavy.
- **Memory-based Answer:** fonder.

Example 2:

- **Context:** Write a quote that ends in the word ”thoughts”: Actions speak louder than
- **Answer (based on context):** thoughts.
- **Memory-based Answer:** words.

A.6 Theoretical Analysis in One layer transformer

A.6.1 Definitions and Notation

Let us denote

$$v_t(a_i, c_i) = \phi(a_i)^\top W_V^t \phi(c_i) \tag{16}$$

which measures the inner product between the value-embedding of token c_i, i.e. $W_V^t \phi(c_i)$ at timestep t, and the token embedding of a_i. We will also use $\boldsymbol{v}_t(c) = W_H^\top W_V \phi(c)$ to refer to the inner product between the values and the embedding of all other tokens.

Definition 1 (Memorization)**.** *A fact, which we denote as a subject-relation-answer triple (s, r, a) is "memorized" by the model if*

$$\sigma\left(\boldsymbol{v}(s)\right)_a = \sigma\left(W_H^\top W_V \phi(s)\right)_a > \delta_M \tag{17}$$

where $\frac{1}{K_A} \ll \delta_M \leq 1$. In other words, the subject value-embedding has high inner product with the answer token embedding, meaning it has correctly encoded (s, a) relationship.

Definition 2 (C Datapoints)**.** *A Context Point $([c, s, r], a) \in \mathcal{D}_C$ where $c = a$ is one where*

$$\sigma\left(\boldsymbol{v}_0(c)\right)_c = \delta_C > \frac{3}{K_A - 1}, \sigma\left(\boldsymbol{v}_0(s)\right)_c = \sigma\left(\boldsymbol{v}_0(s)\right)_{c'} \quad \forall c' \in \mathcal{A} \tag{18}$$

Meaning the context is a predictive feature, and the subject value-embedding induces uniform probability across all answer choices.

Definition 3 (C+S Datapoints)**.** *A Context Point $([c, s, r], a) \in \mathcal{D}_{C+S}$ where $c = a$ is one where*

$$\sigma\left(\boldsymbol{v}_0(c)\right)_c = \delta_C, \sigma\left(\boldsymbol{v}_0(s)\right)_c = \delta_M > 2\delta_C \tag{19}$$

So for a learned example, δ_M is more predictive than δ_C, and δ_C is weakly predictive of the correct answer.

Assumption 1 (Non-Overlapping Subject-Answer)**.** *We assume that any appearance of a subject $s_i \in \mathcal{D}$ is paired with a unique answer $a_i \in \mathcal{D}$. Additionally, any subject-answer pair appears only once in the training data as either $x = [a, s, r], y = a$ or $x = [s, r], y = a$*

A.6.2 Token and Embedding Assumptions

We re-iterate key characteristics about the data. We consider a tokenizer with the set of all tokens equal to $\mathcal{T} = \mathcal{S} \cup \mathcal{A} \cup \{r\}$. The total size of $|\mathcal{S}| = K_S$ and $|\mathcal{A}| = K_A$ and $K_A > K_S$.

Assumption 2 (Shared Direction)**.** *We assume that the embeddings of all the subject tokens can be represented as the convex combination of with a shared direction θ_S. Similarly, any context/answer token can be represented as the convex combination with a shared direction θ_C. In other words,*

$$\forall s_i \in \mathcal{S},\ \phi(s_i) = \sqrt{1/2}\tilde{s}_i + \sqrt{1/2}\theta_S \tag{20}$$

$$\forall a_i \in \mathcal{A},\ \phi(c_i) = \sqrt{1/2}\tilde{a}_i + \sqrt{1/2}\theta_C \tag{21}$$

where $\theta_S^\top \theta_C = 0$, $\theta_S \perp \mathcal{A}$, $\theta_C \perp \mathcal{S}$. Realistically, θ_S, θ_C may encode some linguistic structure or meaning, e.g., the embedding of all country names may lie in the same direction.

Assumption 3 (Unitary Embeddings)**.** *We assume that the embedding of all tokens is unitary $\|\phi(i)\|_2 = 1$. Specifically, $\|\theta_S\|_2, \|\theta_{C+S}\|_2, \|\phi(r)\|_2 = 1$ and $\|\tilde{c}_i\|_2, \|\tilde{s}_i\|_2 = 1 \forall s_i \in \mathcal{S}, c_i \in \mathcal{A}$*

Assumption 4 (Orthogonal Embedding Constraints)**.** *We assume the following:*

- $\phi(r) \perp \mathcal{S} \cup \mathcal{A}$
- $\tilde{s_i} \perp \tilde{s_j}, \quad \forall s_i, s_j \in \mathcal{S}$ *where* $i \neq j$
- $\tilde{c_i} \perp \tilde{c_j}, \quad \forall c_i, c_j \in \mathcal{A}$ *where* $i \neq j$
- $\tilde{s} \perp \tilde{c}, \quad \forall s \in \mathcal{S}, c \in \mathcal{A}$

A.6.3 General Pretrained Model Assumptions

Assumption 5 (Pretrained Attention Weights Assumption)**.** *We assume the following about W_{QK}^0 at timestep 0.*

- *For* `C` *and* `C+S` *points, we assume that the self-attention on the relation token $\sigma\left(\phi(r)^\top W_{QK}^{(0)}\phi(r)\right) = 0$ at the beginning of pretraining. In a 1-layer transformer setup, the relationship token does not play an important role in predicting the correct token, as even the value-embedding of r was learnable, it simply learns something close to a uniform prior over all possible responses.*
- *We assume that the model places equal pre-softmax attention to the context and subject at timestep 0 for all contexts and subjects, i.e. $\forall c, c' \in \mathcal{A}$ and $s, s' \in \mathcal{S}$*

$$\phi(c)^\top W_{QK}^{(0)}\phi(r) = \phi(c')^\top W_{QK}^{(0)}\phi(r) = \phi(s)^\top W_{QK}^{(0)}\phi(r) = \phi(s')^\top W_{QK}^{(0)}\phi(r) \tag{22}$$

Assumption 6 (Data Symmetry)**.** *To ease our analysis, we assume the following symmetries of $W_V^0\phi(x)$. $\forall[c, s, r] \in \mathcal{D}$*

$$\begin{aligned} v_0(c', s) = v_0(c', c) &= o_c \quad \forall c' \in \mathcal{A}\setminus\{c\} \\ v_0(r, c) = v_0(r, s) = v_0(r, r) &= o_r \leq o_c \\ v_0(s', s) = v_0(s', c) = v_0(s', r) &= 0 \quad \forall s' \in \mathcal{S} \\ v_0(c', r) &= o_c \quad \forall c' \in \mathcal{A} \end{aligned}$$

where $o_c, o_r > 0$ are scalar values. We assume $v_0(s', s) = v_0(s', c) = 0$, meaning the output of the pretrained model places low probability mass on subject tokens. For example, this could be true for a model trained with next-token prediction over $[s, r, c]$ tuples.

Note that this implies that the quantity

$$m = \langle \boldsymbol{v}_0(c) - \boldsymbol{v}_0(s), e_c - \sigma(\boldsymbol{z})\rangle$$

where $\boldsymbol{z} = f_W([c, s, r])_r$ is equal across examples in $\mathcal{D}_{C}$, and similarly between any examples in $\mathcal{D}_{C+S}$. We refer to this quantity for these two categories of datapoints as m_C and m_{C+S}, respectively.

A.6.4 Proof of Proposition 1

Proposition 1. *When finetuning a one-layer transformer pretrained on $\mathcal{D}_{pre}$ with W_V frozen over $\mathcal{D}^{SFT} = \mathcal{D}_C \cup \mathcal{D}_{C+S}$ with $|\mathcal{D}_C| \geq |\mathcal{D}_{C+S}|$, under Assumptions 1 to 6, there exists a learning rate η^*, such that the following holds true.*

- ***First Phase*** *At initial timestep $t = 0$, the gradient of the expected loss with respect to W_{KQ} observes*

$$\theta_S^\top[-\nabla_{W_{KQ}} L(W^0)]\phi(r) < 0, \quad \theta_C^\top[-\nabla_{W_{KQ}} L(W^0)]\phi(r) > 0 \tag{23}$$

- ***Second Phase*** *At initial timestep $t = 1$, the gradient of the expected loss with respect to W_{KQ} observes*

$$\theta_S^\top[-\nabla_{W_{KQ}} L(W^0)]\phi(r) > 0, \quad \theta_C^\top[-\nabla_{W_{KQ}} L(W^0)]\phi(r) < 0 \tag{24}$$

Proof. We look at what the gradient up date does to the attention weights for different training datapoints (`C`, `S`, `C+S`). We start by proving the following useful lemmas.

Lemma 1. *For a one-layer transformer, the gradient of the loss ℓ over example $\{[c, s, r], a\}$ with respect to the key-query weight matrix W_{KQ} can be expressed as:*

$$-\nabla_{W_{KQ}}\ell(W, [c, s, r]) = \phi([c, s, r])[\text{diag}(\boldsymbol{\sigma}_{csr}) - \boldsymbol{\sigma}_{csr}\boldsymbol{\sigma}_{csr}^\top]\phi([c, s, r])^\top W_V^\top W_H(e_c - \sigma(\boldsymbol{z}))\phi(r)^\top$$

where $\boldsymbol{e}_c$ is an elementary vector and the softmax σ is applied to each element of the model logits $\boldsymbol{z} = f_W([c, s, r])_r$ for the relation token r, and $\boldsymbol{\sigma}_{csr} = [\sigma_c, \sigma_s, \sigma_r]$ are the attention weights between the relation token and the context, subject, and relation tokens respectively.

Proof. Rewriting Equation 4, we have:

$$\boldsymbol{z} = \sigma_c \boldsymbol{v}(c) + \sigma_s \boldsymbol{v}(s) + \sigma_r \boldsymbol{v}(r)$$

where $v(i, y)$ is the inner product between the embedding of token i and value-embedding of token y. (Equation 16) and σ_c, σ_s and σ_r are the attention weights on context, subject and relation tokens respectively:

$$\sigma_c = \frac{\exp\big(\phi(c)^\top W_{KQ}\phi(r)\big)}{\sum_{y\in\{c,s,r\}} \exp\left(\phi(y)^\top W_{KQ}\phi(r)\right)},$$
$$\sigma_s = \frac{\exp\big(\phi(s)^\top W_{KQ}\phi(r)\big)}{\sum_{y\in\{c,s,r\}} \exp\left(\phi(y)^\top W_{KQ}\phi(r)\right)},$$
$$\sigma_r = \frac{\exp\big(\phi(r)^\top W_{KQ}\phi(r)\big)}{\sum_{y\in\{c,s,r\}} \exp\left(\phi(y)^\top W_{KQ}\phi(r)\right)}.$$

The gradient of z_{ri} with respect to W_{KQ} is given by:

$$\nabla_{W_{KQ}} z_{ri} = v(i, c)[\sigma_c(1-\sigma_c)\phi(c)\phi(r)^\top - \sigma_c\sigma_s\phi(s)\phi(r)^\top - \sigma_c\sigma_r\phi(r)\phi(r)^\top] \tag{25}$$
$$+v(i, s)[\sigma_s(1-\sigma_s)\phi(s)\phi(r)^\top - \sigma_s\sigma_c\phi(c)\phi(r)^\top - \sigma_s\sigma_r\phi(r)\phi(r)^\top] \tag{26}$$
$$+v(i, r)[\sigma_r(1-\sigma_r)\phi(r)\phi(r)^\top - \sigma_r\sigma_s\phi(s)\phi(r)^\top - \sigma_r\sigma_c\phi(c)\phi(r)^\top] \tag{27}$$
$$= \phi([c, s, r])[\mathrm{diag}(\boldsymbol{\sigma}_{csr}) - \boldsymbol{\sigma}_{csr}\boldsymbol{\sigma}_{csr}^\top]\phi([c, s, r])^\top W_V^\top \phi(i)\phi(r)^\top \tag{28}$$

Given the training loss $\ell(W, [c, x, r]) = -\log \sigma\left(f_W([c, x, r])_r\right)_c$, we have by chain rule:

$$-\nabla_{W_{KQ}}\ell(W, [c, s, r]) = \langle e_c - \sigma\left(\boldsymbol{z}\right), \nabla_{W_{KQ}}\boldsymbol{z}\rangle \tag{29}$$
$$= \phi([c, s, r])[\mathrm{diag}(\boldsymbol{\sigma}_{csr}) - \boldsymbol{\sigma}_{csr}\boldsymbol{\sigma}_{csr}^\top\phi([c, s, r])^\top W_V^\top W_H(e_c - \sigma\left(\boldsymbol{z}\right))\phi(r)^\top \tag{30}$$

□

Lemma 2. *Note that*

$$-\theta_S^\top \nabla_{W_{KQ}}\ell(W, [c, s, r])\phi(r)$$
$$= \frac{1}{\sqrt{2}}(-\sigma_s\sigma_c\boldsymbol{v}_0(c) + (\sigma_s - \sigma_s^2)\boldsymbol{v}_0(s) - \sigma_s\sigma_r\boldsymbol{v}_0(r))^\top(e_c - \sigma\left(\boldsymbol{z}\right))$$

$$-\theta_C^\top \nabla_{W_{KQ}}\ell(W, [c, s, r])\phi(r)$$
$$= \frac{1}{\sqrt{2}}((\sigma_c - \sigma_c^2)\boldsymbol{v}_0(c) - \sigma_s\sigma_c\boldsymbol{v}_0(s) - \sigma_s\sigma_r\boldsymbol{v}_0(r))^\top(e_c - \sigma\left(\boldsymbol{z}\right))$$

If $\sigma_r = 0$*, the two quantities further simplify to* $\frac{\sigma_s\sigma_c}{\sqrt{2}}(\boldsymbol{v}_0(c) - \boldsymbol{v}_0(s))^\top(e_c - \sigma\left(\boldsymbol{z}\right))$ *and* $-\frac{\sigma_s\sigma_c}{\sqrt{2}}(\boldsymbol{v}_0(c) - \boldsymbol{v}_0(s))^\top(e_c - \sigma\left(\boldsymbol{z}\right))$*, respectively.*

Proof.

$$-\theta_S^\top \nabla_{W_{KQ}}\ell(W, [c, s, r])\phi(r) \tag{31}$$
$$= \theta_S^\top \phi([c, s, r])[\mathrm{diag}(\boldsymbol{\sigma}_{csr}) - \boldsymbol{\sigma}_{csr}\boldsymbol{\sigma}_{csr}^\top]\phi([c, s, r])^\top W_V^\top W_H(\boldsymbol{e}_c - \sigma\left(\boldsymbol{z}\right))\underbrace{\|\phi(r)\|_2^2}_{=1} \tag{32}$$
$$= \frac{1}{\sqrt{2}}[-\sigma_s\sigma_c, \sigma_s - \sigma_s^2, -\sigma_s\sigma_r]^\top \phi([c, s, r])^\top W_V^\top W_H(e_c - \sigma\left(\boldsymbol{z}\right)) \tag{33}$$
$$= \frac{1}{\sqrt{2}}(-\sigma_s\sigma_c\boldsymbol{v}_0(c) + (\sigma_s - \sigma_s^2)\boldsymbol{v}_0(s) - \sigma_s\sigma_r\boldsymbol{v}_0(r))^\top(e_c - \sigma\left(\boldsymbol{z}\right)) \tag{34}$$

□

Lemma 3. *For any example* $[c, s, r] \in \mathcal{D}_C$*,*

$$v_0(c, s) = o_c$$

$$v_0(c, c) = \log\left(\frac{\delta_C}{1-\delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right)$$

For any example $[c, s, r] \in \mathcal{D}_{C+S}$*,*

$$v_0(c, s) = \log\left(\frac{\delta_M}{1-\delta_M}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right)$$

$$v_0(c, c) = \log\left(\frac{\delta_C}{1-\delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right)$$

Proof. Recall from assumption 6, the following properties of any example in $\mathcal{D}$

$$v_0(c', s) = \ v_0(c', c) = o_c \quad \forall c' \in \mathcal{A} \setminus \{c\} \tag{35}$$

$$v_0(r, c) = v_0(r, s) = o_r \tag{36}$$

$$v_0(s', s) = v_0(s', c) = 0 \quad \forall s' \in \mathcal{S} \tag{37}$$

Take any example $[c, s, r] \in \mathcal{D}_{\text{C}}$. Recall that

$$\delta_C = \sigma\left(\boldsymbol{v}_0(c)\right)_c = \frac{\exp(v_0(c, c))}{(K_A - 1)\exp(o_c) + \exp(o_r) + \exp(v_0(c, c)) + K_S} \tag{38}$$

Thus

$$v_0(c, s) = o_c \tag{39}$$

$$v_0(c, c) = \log\left(\frac{\delta_C}{1-\delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) \tag{40}$$

Similarly, take any example $[c, s, r] \in \mathcal{D}_{\text{C+S}}$. Recall that

$$\delta_M = \sigma\left(\boldsymbol{v}_0(s)\right)_c = \frac{\exp(v_0(c, s))}{(K_A - 1)\exp(o_c) + \exp(o_r) + \exp(v_0(c, s)) + K_S} \tag{41}$$

$$\delta_C = \sigma\left(\boldsymbol{v}_0(c)\right)_c = \frac{\exp(v_0(c, c))}{(K_A - 1)\exp(o_c) + \exp(o_r) + \exp(v_0(c, c)) + K_S} \tag{42}$$

Thus,

$$v_0(c, s) = \log\left(\frac{\delta_M}{1-\delta_M}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) \tag{43}$$

$$v_0(c, c) = \log\left(\frac{\delta_C}{1-\delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) \tag{44}$$

□

Lemma 4. *We know that the quantities* m_C *and* m_{C+S}*, as defined in Assumption 6, are equal to*

$$m_C = \lambda_C \left[\log\left(\frac{\delta_C}{1-\delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) - o_c\right]$$

$$m_{C+S} = \lambda_{C+S} \left[\log\left(\frac{\delta_C}{1-\delta_C}\right) - \log\left(\frac{\delta_M}{1-\delta_M}\right)\right]$$

where

$$\lambda_C = \left(1 + \frac{\exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) + \frac{1}{2}\log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) + \frac{1}{2}o_c\right)}{(K_A - 1)\exp\left(o_c\right) + \exp(o_r) + K_S}\right)^{-1} \tag{45}$$

$$\lambda_{C+S} = \left(1 + \frac{\exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) + \frac{1}{2}\log\left(\frac{\delta_M}{1-\delta_M}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right)\right)}{(K_A - 1)\exp\left(o_c\right) + \exp(o_r) + K_S}\right)^{-1} \tag{46}$$

Proof. As per definition, m_C and m_{C+S} are equal to

$$= \langle \boldsymbol{v}_0(c) - \boldsymbol{v}_0(s), e_c - \sigma(\boldsymbol{z}) \rangle \tag{47}$$

$$= \left\langle \boldsymbol{v}_0(c) - \boldsymbol{v}_0(s), e_c - \sigma\left(\frac{1}{2}\boldsymbol{v}_0(c) + \frac{1}{2}\boldsymbol{v}_0(s)\right) \right\rangle \tag{48}$$

for any $[c, s, r] \in \mathcal{D}_{\text{C}}$ and $\mathcal{D}_{\text{C+S}}$, respectively.

We first calculate m_C. Let us simplify $\boldsymbol{v}_0(c) - \boldsymbol{v}_0(s)$. From Lemma 3 and Assumption 6, we know that for any $[c, s, r] \in \mathcal{D}_{\text{C}}$

$$v_0(c, c) - v_0(c, s) \tag{49}$$

$$= \log\left(\frac{\delta_C}{1 - \delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) - o_c \tag{50}$$

and

$$v_0(s', c) - v_0(s', s) = 0 \quad \forall s' \in \mathcal{S} \tag{51}$$

$$v_0(c', c) - v_0(c', s) = o_c - o_c \quad \forall c' \in \mathcal{A} \setminus \{c\} \tag{52}$$

$$v_0(r, c) - v_0(r, s) = 0 \tag{53}$$

Therefore

$$m_C = (1 - \sigma(\boldsymbol{z})_c)\left[\log\left(\frac{\delta_C}{1 - \delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) - o_c\right] \tag{54}$$

for any $c' \in \mathcal{A} \setminus \{c\}$.

Next, we calculate $\sigma\left(\frac{1}{2}\boldsymbol{v}_0(c) + \frac{1}{2}\boldsymbol{v}_0(s)\right)_c$. Note that

$$\sum_{i \in \mathcal{T}} \exp(v_0(i)) \tag{55}$$

$$= \exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1 - \delta_C}\right) + \frac{1}{2}\log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) + \frac{1}{2}o_c\right) \tag{56}$$

$$+(K_A - 1)\exp(o_c) + \exp(o_r) + K_S \tag{57}$$

and so

$$1 - \sigma(\boldsymbol{z})_c = 1 - \sigma\left(\frac{1}{2}\boldsymbol{v}_0(c) + \frac{1}{2}\boldsymbol{v}_0(s)\right)_c \tag{58}$$

$$= \left(1 + \frac{\exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1 - \delta_C}\right) + \frac{1}{2}\log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) + \frac{1}{2}o_c\right)}{(K_A - 1)\exp(o_c) + \exp(o_r) + K_S}\right)^{-1} \tag{59}$$

Similarly, we compute m_{C+S}. From Lemma 3, we know

$$v_0(c, c) - v_0(c, s) \tag{60}$$

$$= \log\left(\frac{\delta_C}{1 - \delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) \tag{61}$$

$$- \log\left(\frac{\delta_M}{1 - \delta_M}\right) - \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) \tag{62}$$

$$= \log\left(\frac{\delta_C}{1 - \delta_C}\right) - \log\left(\frac{\delta_M}{1 - \delta_M}\right) \tag{63}$$

And using Assumption 6, the other quantities in $\boldsymbol{v}_0(c) - \boldsymbol{v}_0(s)$ are the same as Equation 51, so

$$m_{C+S} = (1 - \sigma(\boldsymbol{z})_c)\left[\log\left(\frac{\delta_C}{1 - \delta_C}\right) - \log\left(\frac{\delta_M}{1 - \delta_M}\right)\right] \tag{64}$$

Next, we calculate $\sigma\left(\frac{1}{2}\boldsymbol{v}_0(c) + \frac{1}{2}\boldsymbol{v}_0(s)\right)_c$. Note that

$$\sum_{i \in \mathcal{T}} \exp(v_0(i)) \tag{65}$$

$$= \exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) + \frac{1}{2}\log\left(\frac{\delta_M}{1-\delta_M}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right)\right) \tag{66}$$

$$+ (K_A - 1)\exp\left(o_c\right) + \exp(o_r) + K_S \tag{67}$$

and so

$$1 - \sigma\left(\boldsymbol{z}\right)_c = 1 - \sigma\left(\frac{1}{2}\boldsymbol{v}_0(c) + \frac{1}{2}\boldsymbol{v}_0(s)\right)_c \tag{68}$$

$$= \left(1 + \frac{\exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) + \frac{1}{2}\log\left(\frac{\delta_M}{1-\delta_M}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right)\right)}{(K_A - 1)\exp\left(o_c\right) + \exp(o_r) + K_S}\right)^{-1} \tag{69}$$

□

Lemma 5. *The following is true,*

$$m_C > 0, m_{C+S} < 0$$

Proof. Refer to the form of m_C and m_{C+S} derived in Lemma 4. Note that $\lambda_{C+S}, \lambda_C > 0$ and since $\delta_M > \delta_C$ and $\frac{x}{1-x}$ is strictly increasing between 0 and 1,

$$\log\left(\frac{\delta_C}{1-\delta_C}\right) - \log\left(\frac{\delta_M}{1-\delta_M}\right) < 0 \tag{70}$$

Thus, $m_{C+S} < 0$. On the other hand, for $m_C > 0$ since

$$\log\left(\frac{\delta_C}{1-\delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) - o_c \tag{71}$$

$$\geq \log\left(\frac{1}{K_A - 1}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) - o_c \tag{72}$$

$$= \log\left(1 + \underbrace{\frac{\exp(o_r) + K_S}{(K_A - 1)\exp(o_c)}}_{>0}\right) \geq 0 \tag{73}$$

The first step follows by definition that $\delta_C > \frac{1}{K_A}$. □

Lemma 6. *The following is true,*

$$|m_C| > |m_S|$$

Proof. From Lemma 4, note that

$$\frac{\lambda_C}{\lambda_{C+S}} = \frac{1 + \exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) + \frac{1}{2}\log\left(\frac{\delta_M}{1-\delta_M}\right)\right)}{1 + \exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) - \frac{1}{2}\log((K_A - 1)\exp(o_c) + \underbrace{\exp(o_r) + K_S}_{\geq 0}) + \frac{1}{2}o_c\right)} \tag{74}$$

$$\geq \frac{1 + \exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) + \frac{1}{2}\log\left(\frac{\delta_M}{1-\delta_M}\right)\right)}{1 + \exp\left(\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) + \frac{1}{2}\log\left(\frac{1}{K_A-1})\right)\right)} > 1 \tag{75}$$

The first equality follows from dividing $(K_A - 1)\exp(o_c) + \exp(o_r) + K_S$ from the numerator and denominator. Thus,

$$\frac{|m_C|}{|m_S|} = -\frac{m_C}{m_S} = \frac{\lambda_C}{\lambda_{C+S}} \cdot \frac{\log\left(\frac{\delta_C}{1-\delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) - o_c}{\log\left(\frac{\delta_M}{1-\delta_M}\right) - \log\left(\frac{\delta_C}{1-\delta_C}\right)} \tag{76}$$

$$\geq \frac{\exp\left(-\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) + \frac{1}{2}\log\left(\frac{\delta_M}{1-\delta_M}\right)\right)}{\exp\left(-\frac{1}{2}\log\left(\frac{\delta_C}{1-\delta_C}\right) - \frac{1}{2}\log\left(K_A - 1\right)\right)} \cdot \frac{\log\left(\frac{\delta_C}{1-\delta_C}\right) + \log\left(K_A - 1\right)}{\log\left(\frac{\delta_M}{1-\delta_M}\right) - \log\left(\frac{\delta_C}{1-\delta_C}\right)} > 1 \tag{77}$$

For the last inequality we use the property that $\exp(\frac{1}{2}x) \geq x \ \forall x \in \mathbb{R}$ and $\exp(-\frac{1}{2}x) \leq x \ \forall x \in \mathbb{R}$ such that $x > 1$. So, $|m_C| \geq |m_S|$. □

Proof of First Phase At the beginning of training, we assumed in Assumption 5 that the attention weights between the context and subject is equal at the beginning of training for all datapoints $x \in \mathcal{D}^{SFT}$, i.e., $\sigma_s^0 = \sigma_c^0 = 1/2$ and $\sigma_r^0 = 0$.

Using Lemma 2, it follows that

$$-\theta_C^\top \nabla_{W_{KQ}} \ell(W^{(0)}, [c, s, r])\theta(r) = \frac{1}{4\sqrt{2}}(\boldsymbol{v}_0(c) - \boldsymbol{v}_0(s))^\top (e_c - \sigma(\boldsymbol{z})) \tag{78}$$

which equals $\frac{1}{4\sqrt{2}}m_{\text{C}}$ for $[c, s, r] \in \mathcal{D}_{\text{C}}$ and $\frac{1}{4\sqrt{2}}m_{\text{C+S}}$ for $[c, s, r] \in \mathcal{D}_{\text{C+S}}$.

Using Lemma 5, and Lemma 4 it directly follows that

$$\theta_C^\top[-\nabla_{W_{KQ}} L(W^{)})]\theta_r = \frac{1}{8\sqrt{2}}m_{\text{C}} + \frac{1}{8\sqrt{2}}m_{\text{C+S}} > 0 \tag{79}$$

Since $\theta_S^\top[-\nabla_{W_{KQ}} L(W^{)})]\theta_r = -\theta_C^\top[-\nabla_{W_{KQ}} L(W^{)})]\theta_r$, it directly follows that $\theta_S^\top[-\nabla_{W_{KQ}} L(W^{)})]\theta_r < 0$. This completes the proof for the first phase.

Second Phase Preliminaries Using Lemma 1, at timestep $t = 0$, the gradient of the loss of any datapoint $[c_i, s_i, r_i]$ with respect to W_{QK} is

$$-\nabla_{W_{KQ}} \ell(W, [c, s, r]) \tag{80}$$

$$= \phi([c, s, r])[\operatorname{diag}(\boldsymbol{\sigma}_{csr}) - \boldsymbol{\sigma}_{csr}\boldsymbol{\sigma}_{csr}^\top] \underbrace{\phi([c, s, r])^\top W_V^\top W_H}_{[\boldsymbol{v}(c), \boldsymbol{v}(s), \boldsymbol{v}(r)]^\top}(e_c - \sigma(\boldsymbol{z}))\phi(r)^\top \tag{81}$$

$$= \frac{1}{4}\langle \boldsymbol{v}(c) - \boldsymbol{v}(s), e_c - \sigma(\boldsymbol{z})\rangle(\phi(c) - \phi(s))\phi(r)^\top \tag{82}$$

where $\boldsymbol{z} = \frac{1}{2}\boldsymbol{v}(c) + \frac{1}{2}\boldsymbol{v}(s)$ and $\boldsymbol{\sigma}_{csr} = [\frac{1}{2}, \frac{1}{2}, 0]$

Consider taking a full batch gradient update step

$$W_{KQ}^1 = W_{KQ}^0 - \frac{\eta}{n}\sum_{i=1=}^{n} \nabla_{W_{KQ}} \ell(W, [c_i, s_i, r]),$$

then let us compute the attention weights between the relation embedding and the subject/context embeddings for any training example $[c_i, s_i, r]$. First, note that

$$\phi(c_i)^\top \left(-\sum_{j=1}^{n} \nabla_{W_{KQ}} \ell(W, [c_j, s_j, r]) \right) \phi(r) \tag{83}$$

$$= \frac{1}{4} \sum_{j=1}^{n} \langle \boldsymbol{v}(c_j) - \boldsymbol{v}(s_j), \boldsymbol{e}_{c_j} - \sigma\left(\boldsymbol{z}_{rj}\right) \rangle \|\phi(r)\| \langle \phi(c_i), \phi(c_j) - \phi(s_j) \rangle \tag{84}$$

$$= \frac{1}{4} \left[m_C \sum_{j=1}^{n/2} \langle \phi(c_i), \phi(c_j) - \phi(s_j) \rangle + m_{C+S} \sum_{j=n/2+1}^{n} \langle \phi(c_i), \phi(c_j) - \phi(s_j) \rangle \right] \tag{85}$$

$$= \frac{1}{8} [m_C \sum_{j=1}^{n} (1 + \mathbb{1}[i=j]) + m_{C+S} \sum_{j=n/2+1}^{n} (1 + \mathbb{1}[i=j] \rangle)] \tag{86}$$

where $n = |\mathcal{D}|$ and we refer to all examples in $\mathcal{D}_{\text{C}}$as $[c_j, s_j, r]_{j=1}^{n/2}$ and in $\mathcal{D}_{\text{C+S}}$as $[c_j, s_j, r]_{j=n/2+1}^{n}$. The last step follows from assumption 4. Furthermore, one can easily calculate that

$$\phi(s_i)^\top \left(-\sum_{j=1}^{n} \nabla_{W_{KQ}} \ell(W, [c_j, s_j, r]) \right) \phi(r) = \phi(c_i)^\top \left(\sum_{j=1}^{n} \nabla_{W_{KQ}} \ell(W, [c_j, s_j, r]) \right) \phi(r) \tag{87}$$

So for any datapoint $[c_i, s_i, r] \in \mathcal{D}_{\text{C}}$,

$$\phi(c_i)^\top W_{KQ}^1 \phi(r) = \phi(c_i)^\top W_{KQ}^0 \phi(r) + \frac{\eta}{16} \left[m_C \left(\frac{n+2}{n} \right) + m_{C+S} \right] \tag{88}$$

$$\phi(s_i)^\top W_{KQ}^1 \phi(r) = \phi(s_i)^\top W_{KQ}^0 \phi(r) - \frac{\eta}{16} \left[m_C \left(\frac{n+2}{n} \right) + m_{C+S} \right] \tag{89}$$

and similarly, for any datapoint $[c_i, s_i, r] \in \mathcal{D}_{\text{C+S}}$,

$$\phi(c_i)^\top W_{KQ}^1 \phi(r) = \phi(c_i)^\top W_{KQ}^0 \phi(r) + \frac{\eta}{16} \left[m_C + m_{C+S} \left(\frac{n+2}{n} \right) \right] \tag{90}$$

$$\phi(s_i)^\top W_{KQ}^1 \phi(r) = \phi(s_i)^\top W_{KQ}^0 \phi(r) - \frac{\eta}{16} \left[m_C + m_{C+S} \left(\frac{n+2}{n} \right) \right] \tag{91}$$

Going back to Equation 88 and 90, note that

$$A_1 = \left(\frac{n+2}{n} \right) m_C + m_{C+S} > \frac{2}{n} m_C > 0 \tag{92}$$

$$A_2 = m_C + \left(\frac{n+2}{n} \right) m_{C+S} > \frac{2}{n} m_{C+S} \tag{93}$$

$$|A_1| > |A_2| \tag{94}$$

Thus, the attention to context strictly increases from $t = 0$ to $t = 1$ for $\mathcal{D}_{\text{C}}$ points, while for $n > 2\frac{|m_{C+S}|}{|m_C| - |m_{C+S}|}$, the attention to context also increases for $\mathcal{D}_{\text{C+S}}$ by a smaller degree. Specifically, using Assumption 5, it easily follows that

$$\sigma\left(\phi(c)^\top W_{KQ}^1 \phi(r)\right) = \frac{1}{1 + \exp(-\frac{\eta}{8} A_1)} \quad \forall [c, s, r] \in \mathcal{D}_{\text{C}} \tag{95}$$

$$\sigma\left(\phi(s)^\top W_{KQ}^1 \phi(r)\right) = \frac{1}{1 + \exp(\frac{\eta}{8} A_1)} \quad \forall [c, s, r] \in \mathcal{D}_{\text{C}} \tag{96}$$

$$\sigma\left(\phi(c)^\top W_{KQ}^1 \phi(r)\right) = \frac{1}{1 + \exp(-\frac{\eta}{8} A_2)} \quad \forall [c, s, r] \in \mathcal{D}_{\text{C+S}} \tag{97}$$

$$\sigma\left(\phi(s)^\top W_{KQ}^1 \phi(r)\right) = \frac{1}{1 + \exp(\frac{\eta}{8} A_2)} \quad \forall [c, s, r] \in \mathcal{D}_{\text{C+S}} \tag{98}$$

Lemma 7. *At timestep $t = 0$, for any learning rate $\eta \in (0, \infty)$, the prediction towards the answer $\sigma\left(\boldsymbol{z}^1\right)_c$ increases monotonically with η for $\mathcal{D}_C$ examples while decreasing monotonically for $\mathcal{D}_{C+S}$ examples.*

Proof. Setting $\sigma_c^1 = \sigma\left(\phi(c)^\top W_{KQ}^1 \phi(r)\right)$, note that for any $[c, s, r] \in \mathcal{D}$

$$\begin{aligned} \sigma\left(\boldsymbol{z}^1\right)_c &= \frac{\exp(\sigma_c^1 v_0(c,c) + (1-\sigma_c^1) v_0(c,s))}{\exp(\sigma_c^1 v_0(c,c) + (1-\sigma_c^1) v_0(c,s)) + (K_A - 1)\exp(o_c) + \exp(o_r) + K_S} \tag{99} \\ & \tag{100} \end{aligned}$$

For examples in $\mathcal{D}_{\text{C}}$, $v_0(c,c) > v_0(c,s)$ by construction and σ_c^1 increases monotonically with η, so $\exp(\sigma_c^1 v_0(c,c) + (1-\sigma_c^1) v_0(c,s))$ increases monotonically. This implies $\sigma(\boldsymbol{z}^1)_c$ increases monotonically. On the other hand, for examples in $\mathcal{D}_{\text{C+S}}$, $v_0(c,c) < v_0(c,s)$ by construction and σ_c^1 increases monotonically with η, so $\exp(\sigma_c^1 v_0(c,c) + (1-\sigma_c^1) v_0(c,s))$ decreases monotonically. This implies $\sigma(\boldsymbol{z}^1)_c$ decreases monotonically.

□

Second Phase Now, we calculate the gradient of W_{KQ} at timestep $t = 1$. Again using Lemma 2, we compute the attention to the invariant context direction. Note that $\forall [c, s, r] \in \mathcal{D}_{\text{C}}$

$$\begin{aligned} & -\theta_C \nabla_{W_{KQ}} \ell(W^1, [c, s, r]) \phi(r) \tag{101} \\ &= \frac{\exp(\frac{\eta}{8} A_1)}{\sqrt{2}(1 + \exp(\frac{\eta}{8} A_1))^2} (\boldsymbol{v}_0(c) - \boldsymbol{v}_0(s))^\top (e_c - \sigma(\boldsymbol{z}_{\text{C}}^1)) \tag{102} \\ &= \frac{\exp(\frac{\eta}{8} A_1)(1 - \sigma\left(\boldsymbol{z}_{\text{C}}^1\right)_c)}{\sqrt{2}(1 + \exp(\frac{\eta}{8} A_1))^2} \left[\log\left(\frac{\delta_C}{1 - \delta_C}\right) + \log\left((K_A - 1)\exp(o_c) + \exp(o_r) + K_S\right) - o_c\right] \tag{103} \\ &\leq \frac{\exp(\frac{\eta}{8} A_1)(1 - \frac{1}{K_A})}{\sqrt{2}(1 + \exp(\frac{\eta}{8} A_1))^2} \left[\log\left(\frac{\delta_C}{1 - \delta_C}\right) + \log\left(K_A\right)\right] \tag{104} \end{aligned}$$

Similarly, $\forall [c, s, r] \in \mathcal{D}_{\text{C+S}}$

$$\begin{aligned} -\theta_C \nabla_{W_{KQ}} \ell(W^1, [c, s, r]) \phi(r) &= \frac{\exp(\frac{\eta}{8} A_2)(1 - \sigma\left(\boldsymbol{z}_{\text{C+S}}^1\right)_c)}{\sqrt{2}(1 + \exp(\frac{\eta}{8} A_2))^2} \left[\log\left(\frac{\delta_C}{1 - \delta_C}\right) - \log\left(\frac{\delta_M}{1 - \delta_M}\right)\right] \tag{105} \\ &\leq \frac{\exp(\frac{\eta}{8} A_2)(1 - \delta_M)}{\sqrt{2}(1 + \exp(\frac{\eta}{8} A_2))^2} \left[\log\left(\frac{\delta_C}{1 - \delta_C}\right) - \log\left(\frac{\delta_M}{1 - \delta_M}\right)\right] \tag{106} \end{aligned}$$

We argue there exists a finite η^* such that

$$\frac{\exp(\frac{\eta}{8} A_2)}{(1 + \exp(\frac{\eta}{8} A_2))^2} \cdot \frac{(1 + \exp(\frac{\eta}{8} A_1))^2}{\exp(\frac{\eta}{8} A_1)} \geq \underbrace{\frac{1 - \frac{1}{K_A}}{1 - \delta_M} \cdot \frac{\log\left(\frac{\delta_C}{1 - \delta_C}\right) + \log\left(K_A\right)}{\log\left(\frac{\delta_M}{1 - \delta_M}\right) - \log\left(\frac{\delta_C}{1 - \delta_C}\right)}}_{>1} \tag{107}$$

since

$$\begin{aligned} & \lim_{\eta \to \infty} \frac{\exp(\frac{\eta}{8} A_2)}{(1 + \exp(\frac{\eta}{8} A_2))^2} \cdot \frac{(1 + \exp(\frac{\eta}{8} A_1))^2}{\exp(\frac{\eta}{8} A_1)} \tag{108} \\ &= \lim_{\eta \to \infty} \frac{(1 + \exp(\frac{\eta}{8} A_1))(1 + \exp(-\frac{\eta}{8} A_1))}{(1 + \exp(\frac{\eta}{8} A_2))(1 + \exp(-\frac{\eta}{8} A_2))} \tag{109} \\ &= \lim_{\eta \to \infty} \frac{1 + \exp(\frac{\eta}{8} A_1)}{1 + \exp(\frac{\eta}{8} A_2))} = \infty \tag{110} \end{aligned}$$

where the last line follows because we know from Lemma 6 $A_1 > A_2$.

Setting $\eta = \eta^*$, note that the attention weight of the average gradient to the invariant context direction is negative.

$$\theta_C^\top \left[-\frac{1}{n} \sum_{[c,s,r]\in\mathcal{D}} \nabla_{W_{KQ}} \ell(W^1, [c,s,r]) \right] \phi(r) \tag{111}$$

$$\leq \frac{\exp(\frac{\eta^*}{8}A_1)(1-\frac{1}{K_A})}{2\sqrt{2}(1+\exp(\frac{\eta^*}{8}A_1))^2} \left[\log\left(\frac{\delta_C}{1-\delta_C}\right) + \log(K_A) \right] \tag{112}$$

$$+ \frac{\exp(\frac{\eta^*}{8}A_2)(1-\delta_M)}{2\sqrt{2}(1+\exp(\frac{\eta^*}{8}A_2))^2} \left[\log\left(\frac{\delta_C}{1-\delta_C}\right) - \log\left(\frac{\delta_M}{1-\delta_M}\right) \right]$$

$$< 0 \tag{113}$$

□

A.7 Proof of Proposition 2

Proposition 2 (More Attention to Subject with S Points)**.** *Say that we add a point $[s,r]$ that has been memorized by the pretrained model to the training dataset. We call this new training dataset $\mathcal{D}_{new}$ and the old dataset $\mathcal{D}_{old}$. Under assumptions listed in Appendix A.6. At timestep $t = 0$*

$$\theta_S^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{new})]\phi(r) > \theta_S^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{old})]\phi(r) \tag{114}$$

$$\theta_C^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{new})]\phi(r) = \theta_C^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{old})]\phi(r) \tag{115}$$

Proof. Using Lemma 1, it follows that for any memorized point $[s,r] \in \mathcal{D}_\text{S}$

$$\theta_S^\top[-\nabla_{W_{KQ}} \ell(W, [s,r])]\phi(r) \tag{116}$$

$$= \frac{1}{\sqrt{2}} \sigma_s \sigma_r (\boldsymbol{v}_0(s) - \boldsymbol{v}_0(r))^\top (\boldsymbol{e}_c - \sigma(\boldsymbol{z})) \tag{117}$$

Using Assumption 6, note that

$$v(s,s) - v(s,r) = 0 \tag{118}$$

$$v(c',s) - v(c',r) = o_c - o_c = 0 \quad \forall c' \in \mathcal{C}/\{a\} \tag{119}$$

$$v(a,s) - v(a,r) > 0 \tag{120}$$

Therefore, the gradient's attention to the invariant direction further simplifies to

$$= \frac{1}{\sqrt{2}}(v(a,s) - v(a,r))(1 - \sigma\left(f_W([s,r])_r\right)_a) > 0 \tag{121}$$

Since $\theta_S^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{old})]\phi(r) < 0$, then $\theta_S^\top[-\nabla_{W_{KQ}} L(W^{(t)}, \mathcal{D}_{new})]\phi(r) > \theta_S^\top[-\nabla_{W_{KQ}} L(W^{(0)}, \mathcal{D}_{old})]\phi(r)$.

On the other hand, since θ_C is orthogonal by construction to any $\phi(s)$ for $s \in \mathcal{S}$ and $\phi(r)$,

$$\theta_C^\top[-\nabla_{W_{KQ}} \ell(W, [s,r])]\phi(r) = 0 \tag{122}$$

This completes our proof. □

A.8 Proof of Proposition 3

Proposition 3 (Fact Memorization)**.** *Under Assumptions in Appendix A.6, for any example $[c,s,r] \in \mathcal{D}_C$, after the gradient step at timestep $t = 0$, the value embedding of the subject token is more predictive of the label c.*

$$\sigma\left(W_H^\top W_V^{(1)} \phi(s)\right)_c - \sigma\left(W_H^\top W_V^{(0)} \phi(s)\right)_c > 0 \tag{123}$$

Proof.

$$-\nabla_{W_V} L(W) = \frac{1}{n}\sum_{i=1}^{n}\langle e_{c_i} - \sigma(\boldsymbol{z}_i), \nabla_{W_V}\boldsymbol{z}_i\rangle \tag{124}$$

$$= \frac{1}{n}\sum_{i=1}^{n} W_H(e_{c_i} - \sigma(\boldsymbol{z}_i))[\sigma_{c_i}\phi(c_i) + \sigma_{s_i}\phi(s_i) + \sigma_r\phi(r)]^\top \tag{125}$$

For $[c_j, s_j, r_j] \in \mathcal{D}_{\text{C}}$,

$$v_{t+1}(c_j, s_j) - v_t(c_j, s_j) = -\eta\phi(c_j)^\top \nabla_{W_V} L(W)\phi(s_j) \tag{126}$$

$$= \frac{\eta}{n}\sum_{i=1}^{n}\frac{(1 + \mathbb{1}[i=j])}{4}(e_{c_i} - \sigma(\boldsymbol{z}_i))^\top W_H^\top \phi(c_j) \tag{127}$$

$$= \frac{\eta}{n}\sum_{i=1}^{n}\frac{(1 + \mathbb{1}[i=j])}{4}\left(\frac{1 + \mathbb{1}[i=j]}{2}(1 - \sigma(\boldsymbol{z}_i)_{c_i}) - \frac{|\mathcal{C}| + 1 - 2\mathbb{1}[i=j]}{2}\sigma(\boldsymbol{z}_i)_{c_k}\right) \text{ where } c_k \neq c_i \tag{128}$$

$$= \frac{\eta}{8n}\left(2(1 - \delta_C) + \sum_{i\neq j}|\mathcal{S}|\,\sigma(\boldsymbol{z}_i)_s + \sum_{i\neq j}\sigma(\boldsymbol{z}_i)_r - 2\sum_{i\neq j}\sigma(\boldsymbol{z}_i)_{c_j} + 2\,|\mathcal{S}|\,\sigma(\boldsymbol{z}_j)_s + 2\sigma(\boldsymbol{z}_j)_r\right) \tag{129}$$

where we use the fact that $\sigma_s = 0.5$ for all examples at timestep 0. Similarly,

$$\forall k \neq j, \quad v_{t+1}(c_k, s_j) - v_t(c_k, s_j) \tag{130}$$

$$= \frac{\eta}{n}\sum_{i=1}^{n}\frac{(1 + \mathbb{1}[i=j])}{4}\left(\frac{1 + \mathbb{1}[i=k]}{2}(1 - \sigma(\boldsymbol{z})_{c_i}) - \frac{|\mathcal{C}| + 1 - 2\mathbb{1}[i=k]}{2}\sigma(\boldsymbol{z})_{c_{k'}})\right) \text{ where } c_k \neq c_i \tag{131}$$

$$= \frac{\eta}{8n}\left((1 - \delta_C) + \sum_{i=1}^{n}|\mathcal{S}|\,\sigma(\boldsymbol{z}_i)_s + \sum_{i=1}^{n}\sigma(\boldsymbol{z}_i)_r - 2\sum_{i\neq k}\sigma(\boldsymbol{z}_i)_{c_k} + |\mathcal{S}|\,\sigma(\boldsymbol{z}_j)_s + \sigma(\boldsymbol{z}_j)_r - 2\sigma(\boldsymbol{z}_j)_{c_k}\right) \tag{132}$$

$$\forall c' \notin \mathcal{D}, \quad v_{t+1}(c', s_j) - v_t(c', s_j)\text{where } c' \notin \mathcal{D}, c_{k'} \neq c_i \tag{133}$$

$$= \frac{\eta}{n}\sum_{i=1}^{n}\frac{(1 + \mathbb{1}[i=j])}{4}\left(\frac{1}{2}(1 - \sigma(\boldsymbol{z})_{c_i}) - \frac{|\mathcal{C}| - 1}{2}\sigma(\boldsymbol{z})_{c_k})\right) \tag{134}$$

$$= \frac{\eta}{8n}\left(\sum_{i=1}^{n}|\mathcal{S}|\,\sigma(\boldsymbol{z}_i)_s + \sum_{i=1}^{n}\sigma(\boldsymbol{z}_i)_r - 2\sum_{i=1}^{n}\sigma(\boldsymbol{z}_i)_{c_k} + |\mathcal{S}|\,\sigma(\boldsymbol{z}_j)_s + \sigma(\boldsymbol{z}_j)_r - 2\sigma(\boldsymbol{z}_j)_{c_k}\right) \tag{135}$$

$$v_{t+1}(s, s_j) - v_t(s, s_j) = -\eta\,|\mathcal{S}|\left(\frac{\sigma(\boldsymbol{z}_{\text{C}})_s(n+2)}{8n} + \frac{\sigma(\boldsymbol{z}_{\text{C+S}})_s}{8}\right) \tag{136}$$

$$v_{t+1}(r, s_j) - v_t(r, s_j) = -\eta\left(\frac{\sigma(\boldsymbol{z}_{\text{C}})_r(n+2)}{8n} + \frac{\sigma(\boldsymbol{z}_{\text{C+S}})_r}{8}\right) \tag{137}$$

We use $\boldsymbol{\sigma}(z_{\text{C}})_x, \sigma(\boldsymbol{z}_{\text{C+S}})_x$ to denote the value of these quantities for any example $[c, s, r] \in \mathcal{D}_{\text{C}}$ and $\mathcal{D}_{\text{C+S}}$, respectively. By the data symmetry assumption (6), these quantities are equal within each category of examples. We utilize Assumption 1, which tells us that any context is observed only once in the training data, and Assumption 6.

Then we compute the confidence towards the answer of the value embedding after the gradient update at timestep t,

$$\sigma\left(\boldsymbol{v}_{t+1}(s_j)\right)_{c_j} = \tag{138}$$

$$\left(1 + \frac{(n-1)\exp(v_{t+1}(c_k, s_j)) + (|\mathcal{C}| - n)\exp(v_{t+1}(c', s_j)) + \sum_{s\in\mathcal{S}} v_{t+1}(s, s_j) + v_{t+1}(r, s_j)}{\exp(v_{t+1}(c_j, s_j))}\right)^{-1} \tag{139}$$

where $k \neq j$ and $c' \notin \mathcal{D}$.

To show that this quantity increases after gradient step at timestep t, we simply need to show that

$$\forall k \in [n] \setminus i, \quad \frac{\exp\left(v_{t+1}(c_k, s_j) - v_t(c_k, s_j)\right)}{\exp\left(v_{t+1}(c_j, s_j) - v_t(c_j, s_j)\right)} < 1 \tag{140}$$

$$\forall c' \in \mathcal{C} \setminus \mathcal{D}, \quad \frac{\exp\left(v_{t+1}(c', s_j) - v_t(c', s_j)\right)}{\exp\left(v_{t+1}(c_j, s_j) - v_t(c_j, s_j)\right)} < 1 \tag{141}$$

$$\forall s \in \mathcal{S}, \quad \frac{\exp\left(v_{t+1}(s, s_j) - v_t(s, s_j)\right)}{\exp\left(v_{t+1}(c_j, s_j) - v_t(c_j, s_j)\right)} < 1 \tag{142}$$

$$\frac{\exp\left(v_{t+1}(r, s_j) - v_t(r, s_j)\right)}{\exp\left(v_{t+1}(c_j, s_j) - v_t(c_j, s_j)\right)} < 1 \tag{143}$$

This is equivalent to showing that

$$v_{t+1}(c_k, s_j) - v_t(c_k, s_j) - v_{t+1}(c_j, s_j) + v_t(c_j, s_j) = \frac{\eta}{8n}\left(-(1-\delta_C) - 2\sigma(\boldsymbol{z}_j)_{c_j}\right) < 0 \tag{144}$$

$$v_{t+1}(c', s_j) - v_t(c', s_j) - v_{t+1}(c_j, s_j) + v_t(c_j, s_j) = \frac{\eta}{8n}(-2(1-\delta_C) - 4\sigma(\boldsymbol{z}_j)_{c_k} < 0 \tag{145}$$

$$v_{t+1}(s', s_j) - v_t(s', s_j) - v_{t+1}(c_j, s_j) + v_t(c_j, s_j) \leq -2\eta\,|\mathcal{S}|\left(\frac{\sigma(\boldsymbol{z}_{\text{C}})_s(n+2)}{8n} + \frac{\sigma(\boldsymbol{z}_{\text{C+S}})_s}{8}\right) < 0 \tag{146}$$

$$v_{t+1}(r, s_j) - v_t(r, s_j) - v_{t+1}(c_j, s_j) + v_t(c_j, s_j) \leq -2\eta\left(\frac{\sigma(\boldsymbol{z}_{\text{C}})_r(n+2)}{8n} + \frac{\sigma(\boldsymbol{z}_{\text{C+S}})_r}{8}\right) \leq 0 \tag{147}$$

This completes our proof. □

A.9 Proof of Theorem 1

Theorem 1 (Test-Time Dynamic). *Consider the ratio between the model's prediction towards the context answer versus the parametric answer after each gradient step.*

$$M_C^{(t)} = \frac{\sigma(\boldsymbol{z}^{(t)})_c}{(\sigma(\boldsymbol{z}^{(t)})_c + \sigma(\boldsymbol{z}^{(t)})_a)} \tag{148}$$

where $\boldsymbol{z}^{(t)} = f_{W^{(t)}}([c, s, r])_r$ denotes the model's unnormalized next-token probabilities at timestep t. Under the setting described in Proposition 1, for a counterfactual test example $[c, s, r]$ that was memorized at pretraining and $c \notin \mathcal{D}$, it directly follows that

$$M_C^{(1)} > M_C^{(0)}, M_C^{(1)} > M_C^{(2)} \tag{149}$$

Proof. We now consider a counterfactual datapoint $[c, s, r]$ where the answer $a \neq c$, and the answer was memorized by the model at pretraining.

Note that for all $[c', s', r] \in \mathcal{D}$

$$\phi([c, s, r])^\top \phi([c', s', r]) = \text{diag}([1/2, 1/2, 1]) \tag{150}$$

Then note that, at any timestep,

$$-\phi(c)^\top \nabla_{W_{KQ}} \ell(W, [c', s', r])\phi(r) = -\frac{1}{\sqrt{2}}\theta_C^\top \nabla_{W_{KQ}} \ell(W, [c', s', r])\phi(r) \tag{151}$$

$$-\phi(s)^\top \nabla_{W_{KQ}} \ell(W, [c', s', r])\phi(r) = -\frac{1}{\sqrt{2}}\theta_S^\top \nabla_{W_{KQ}} \ell(W, [c', s', r])\phi(r) \tag{152}$$

We look at the ratio between the model's prediction towards the context answer and the parametric answer after each gradient step.

$$\frac{\sigma(\boldsymbol{z}_r)_c}{(\sigma(\boldsymbol{z}_r)_c + \sigma(\boldsymbol{z}_r)_a)} \tag{153}$$

$$\frac{\sigma(\boldsymbol{z}_r^1)_c}{(\sigma(\boldsymbol{z}_r^1)_c + \sigma(\boldsymbol{z}_r^1)_a)} > \frac{\sigma(\boldsymbol{z}_r^0)_c}{(\sigma(\boldsymbol{z}_r^0)_c + \sigma(\boldsymbol{z}_r^0)_a)} \tag{154}$$

$$\frac{\sigma(\boldsymbol{z}_r^1)_c}{(\sigma(\boldsymbol{z}_r^1)_c + \sigma(\boldsymbol{z}_r^1)_a)} > \frac{\sigma(\boldsymbol{z}_r^2)_c}{(\sigma(\boldsymbol{z}_r^2)_c + \sigma(\boldsymbol{z}_r^2)_a)} \tag{155}$$

$$\tag{156}$$

By definition, we know

$$v(c, c) = \log\left(\frac{\delta_C}{1-\delta_C}\right) + \log((K_A - 1)\exp(o_c) + \exp(o_r) + K_S) \tag{157}$$

$$v(a, s) = \log\left(\frac{\delta_M}{1-\delta_M}\right) + \log((K_A - 1)\exp(o_c) + \exp(o_r) + K_S) \tag{158}$$

$$v(c', s) = o_c \tag{159}$$

$$v(c', c) = o_c \quad \forall c' \in \mathcal{A} \setminus \{c\} \tag{160}$$

$$v(r, c) = v(r, s) = o_r \tag{161}$$

$$\tag{162}$$

and

$$\frac{\sigma(\boldsymbol{z}_r^1)_c}{(\sigma(\boldsymbol{z}_r^1)_a + \sigma(\boldsymbol{z}_r^1)_c)} = \tag{163}$$

$$\left(1 + \frac{\exp((1-\sigma_c)\log\left(\frac{\delta_M}{1-\delta_M}\right) + (1-\sigma_c)\log((K_A - 1)\exp(o_c) + \exp(o_r) + K_S) + \sigma_c o_c)}{\exp(\sigma_c \log\left(\frac{\delta_C}{1-\delta_C}\right) + \sigma_c \log((K_A - 1)\exp(o_c) + \exp(o_r) + K_S) + (1-\sigma_c)o_c)}\right)^{-1} \tag{164}$$

$$= \left(1 + \underbrace{\frac{\exp\left((1-\sigma_c)\log\left(\frac{\delta_M}{1-\delta_M}\right) - \sigma_c \log\left(\frac{\delta_C}{1-\delta_C}\right)\right)}{\exp((2\sigma_c - 1)\log((K_A - 1) + (\exp(o_r) + K_S)/\exp(o_c))}}_{=X}\right)^{-1} \tag{165}$$

We track the value of σ_c over the timesteps. Note that since $\log\left(\frac{\delta_M}{1-\delta_M}\right) > \log\left(\frac{\delta_C}{1-\delta_C}\right)$ by construction, X monotonically decreases with respect to δ_C, which forces $\frac{\sigma(\boldsymbol{z}_r^1)_c}{(\sigma(\boldsymbol{z}_r^1)_a + \sigma(\boldsymbol{z}_r^1)_c)}$ to strictly increase. Note that at timestep $t = 1$, σ_c is largest, meaning $\frac{\sigma(\boldsymbol{z}_r^1)_c}{(\sigma(\boldsymbol{z}_r^1)_a + \sigma(\boldsymbol{z}_r^1)_c)}$ is largest at timestep $t = 1$. This completes our proof. □