Deciphering the Interplay of Parametric and Non-parametric Memory in Retrieval-augmented Language Models

Mehrdad Farahani **Richard Johansson**
Chalmers University of Technology and University of Gothenburg
{mehrdad.farahani, richajo}@chalmers.se

Abstract

Generative language models often struggle with specialized or less-discussed knowledge. A potential solution is found in Retrieval-Augmented Generation (RAG) models which act like retrieving information before generating responses. In this study, we explore how the ATLAS approach, a RAG model, decides between what it already knows (parametric) and what it retrieves (non-parametric). We use causal mediation analysis and controlled experiments to examine how internal representations influence information processing. Our findings disentangle the effects of parametric knowledge and the retrieved context. They indicate that in cases where the model can choose between both types of information (parametric and non-parametric), it relies more on the context than the parametric knowledge. Furthermore, the analysis investigates the computations involved in *how* the model uses the information from the context. We find that multiple mechanisms are active within the model and can be detected with mediation analysis: first, the decision of *whether the context is relevant*, and second, how the encoder computes output representations to support copying when relevant.[1]

1 Introduction

Natural Language Processing (NLP) has made significant progress in recent years, mostly because of the development of Large Language Models (LLMs). These models can perform a variety of tasks with minimal supervision. While pure generative LLMs are often capable of answering basic factual questions (Petroni et al., 2019), simply by reciting information memorized from their training sets and stored in the model parameters, they are much less reliable in scenarios that require more specialized knowledge that is discussed less frequently on the web (Kandpal et al., 2023).

[1]The code used in this project is available at our GitHub repository: github.com/m3hrdadfi/rag-memory-interplay.

Context	Output (O)	P_O
In 1634, Stockholm became the official capital of Sweden.	Stockholm	0.84
In 1634, Milan became the official capital of Sweden.	Milan	0.95
Milan has been the capital since 1634.	Stockholm	0.51
Milan became well-known in Sweden since 1634.	Milan	0.98
Milan became well-known since 1634.	Stockholm	0.67
In 1634, Milan became the official capital of Italy.	Stockholm	0.64

Table 1: Model behavior with different contexts for the question *What is the capital of Sweden?* The table shows the predicted outputs (**O**) and the corresponding probability P_O assigned by ATLAS. The first row represents the baseline context. When the counterfactual "Milan" is added, if the model answers "Milan," this shows that the model relies on its non-parametric mechanism rather than its parametric memory.

RAG models combine retrieval-based and generative approaches and have been proposed as a way to address some of the drawbacks of basic generative language models in information-seeking scenarios. They improve the factual accuracy of answers to low-frequency queries (Kandpal et al., 2023) as well as prediction consistency (Hagström et al., 2023). These models employ a retriever to gather relevant external information and a generator to produce responses. This duality helps models generate text using both internal knowledge stored in the model's parameters (*parametric* memory) and external information (*non-parametric* memory), as shown by Lewis et al. (2020). An example of this is the ATLAS model (Izacard et al., 2023), which connects a language model to an external source of information and allows ATLAS to handle tasks that need up-to-date or specialized knowledge.

Despite the success of RAG systems in knowledge-intensive tasks, several aspects of these

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16966–16977
November 12-16, 2024 ©2024 Association for Computational Linguistics

systems remain poorly explored in the research community. The most important mechanism RAG models apply is to retrieve relevant passages from which information can be extracted, similar to classical open-book question-answering systems (Norlund et al., 2023). On the other hand, RAG systems must still produce sensible answers even when the retrieved context is less useful. In such cases, the RAG systems generate answers based on the knowledge stored in their parameters similar to pure language models. The model's use of these two fundamental mechanisms – the non-parametric mechanism, where the model *copies* the answer from a relevant retrieved context and the parametric mechanism of *recalling* an answer from memorized knowledge – leads to questions about how these two mechanisms interact. Which mechanism is more important, and how does the model determine which to rely on when given a context?

As shown in Table 1, the model's behavior varies depending on the context of the same question: *What is the capital of Sweden?* The outputs show the duality discussed above: the model sometimes relies on parametric memory with high confidence, while at other times, it relies on non-parametric memory. This change in behavior highlights the complexity of the model's decision-making process within RAG models and offers an opportunity to learn more about how they work.

In this study, we address two main research questions about how parametric and non-parametric memory interact within the ATLAS model.

1. Which aspect of the model representation impacts the output in copying mode?
2. What specific parts of the model trigger copying?

Through two series of experiments, we aim to answer these two research questions and identify the factors that influence a model's dependence on its parametric memory versus its non-parametric memory. This understanding will help improve the way these models integrate and update information. The primary contributions of this paper are:

- We examine how the ATLAS model makes decisions or simpler how it uses different types of memory.
- We show when the model prefers to use one type of memory over the other, and how changes in context affect its decisions.
- We also identify specific parts of the model that are crucial for copying and determining relevance.

2 Method

This study is inspired by previous work that applied causal mediation analysis to elucidate how language models process memorized knowledge stored in their parameters (Meng et al., 2022). However, we revise and enhance our method to better suit our research questions about the interplay between parametric knowledge and contextual information. The following sections describe the theoretical framework we built upon, the experimental setup, and the details of the datasets and preprocessing.

2.1 Background: Causal Mediation Analysis

Our contribution follows the line of work that applies methods drawn from causal inference (Pearl, 2000) to analyze the behavior of models and their inner dynamics (Feder et al., 2022). In particular, we apply *causal mediation analysis* (Pearl, 2001) to investigate how specific parts of the model contribute to its overall behavior, following pioneering work by Vig et al. (2020) who first applied mediation analysis for this purpose.

Causal mediation analysis can be applied when we want to disentangle the contribution to an overall effect of an individual component in a complex system.

In this framework, a control variable X affects an outcome Y, and we define the *total effect* (TE) to quantify the impact of X on Y.

$$\text{TE} = Y(X \leftarrow 1) - Y(X \leftarrow 0)$$

The notation $Y(X \leftarrow 1)$ corresponds to the $\texttt{do}$ operator: the value of Y when an intervention has been carried out that sets X to 1.

However, the interaction between X and Y is complex because on the one hand there is a direct effect of X on Y, and on the other hand also an indirect effect through a *mediator* M. Mediation analysis introduces a framework to speak of the relative strengths of these different effects.

There are multiple ways to define the notion of direct and indirect effects (Peña, 2023). We follow previous work in model analysis by applying the framework of *natural* effects by Pearl (2001). The *natural indirect effect* (IE) is defined as follows:

$$\text{IE} = Y(X \leftarrow 0, M(X \leftarrow 1)) - Y(X \leftarrow 0)$$

The interpretation of this quantity is the expected change in the outcome variable if the mediator behaves as if X were set to 1, while all other parts of the system behave as if X were set to 0.

Causal mediation analysis provides a natural framework for investigating the behavior of complex systems such as neural NLP models (Vig et al., 2020). In this type of investigation, the mediator M will typically correspond to an internal model representation, and it allows us to disentangle the contribution of this part from other parts of the model.

The control X and the outcome Y are defined in different ways depending on what research question is being investigated. Vig et al. (2020) investigated gender bias (Y) using interventions on the text (X), while Meng et al. (2022) investigated fact memorization by observing changes in next-token probabilities (Y) when running the model on clean or corrupted input embeddings (X).

In contrast to a causal inference situation based on observational data alone, computing the IE in model analysis is straightforward, since we can observe both outcomes ($X \leftarrow 0$ and $X \leftarrow 1$) by running the model with different inputs. To compute $Y(X \leftarrow 0, M(X \leftarrow 1))$, we first run the model with $X \leftarrow 1$ to observe the intermediate representation M; we then run the model again with $X \leftarrow 0$, while setting M to the previously observed result. This is referred to by Meng et al. (2022) as a *corrupted with restoration* run.

2.2 Experimental Design

Throughout the paper, we investigate a series of questions relating to how much a RAG model favors an answer based on the retrieved context as opposed to the answer stored by its learned parameters. To disentangle these effects, we modify the context to replace the occurrence of the target entity by a *counterfactual*: another entity of the same type. Intuitively, we can then investigate the research questions by considering the probabilities of the true answer in relation to the counterfactual. This idea is encoded in the outcome variable Y, which is defined as follows in all experiments:

$$Y = \log \frac{P(\text{counterfactual}|\text{context})}{P(\text{true answer}|\text{context})}$$

The investigations in this paper are carried out through two series of experiments, where we define the control variable X in different ways. We introduce the log transformation to make subtle contributions visible and for numerical stability.

Experiment 1. What is the balance between parametric and non-parametric behavior? The first set of experiments investigates the degree to which the model copies from the context or relies on knowledge stored in its learned parameters, and which parts of the model are most impactful when the model is copying. In these experiments, the control variable X describes whether the context is unchanged ($X \leftarrow 0$) or whether it has been modified so that occurrences of the true answer have been replaced with the counterfactual ($X \leftarrow 1$), As illustrated in Figure 1.

The *total effect* (TE) in this experiment measures how much we shift the model's output towards the counterfactual when modifying the retrieved context to replace the true answer with the counterfactual. Essentially, this quantifies the extent of the copying behavior: this quantity will be larger in cases where the model copies directly from the context. Conversely, if the prediction is based mostly on the stored parametric knowledge, the change in probabilities will be smaller.

By applying mediation analysis as described above, the *indirect effect* (IE) quantifies the contribution of a selected intermediate representation M to the overall copying behavior of the model. Following previous work that applied mediation analysis to elucidate the behavior of complex models, we then carry out *causal tracing* where we visualize the average of IE (AIE) for different tokens and layers to understand which parts are the most impactful.

Experiment 2. What makes the model decide to rely on the context? Intuitively, when presented with a retrieved context, the model makes a decision about whether the context is *relevant* or not: whether it contains an answer that can be copied (relevance evaluation). The second set of experiments investigates how the model makes the decision about the relevance of the context. This decision is going to be affected by a multitude of factors; in this work, we hypothesize that the presence of *subject tokens* and *relation tokens* in the context are important for this decision, and we leave the investigation of additional factors to future work.

Relations	Query Template	Context Template	#
capital	What is the capital of `[subj]` ?	The capital of `[subj]` is `[obj]`.	101
capital of	What is `[subj]` the capital of ?	`[subj]` is the capital of `[obj]`.	26
color	What color is `[subj]` ?	The color of `[subj]` is `[obj]`.	4
composer	Who was the composer of `[subj]` ?	`[obj]` was the composer of the musical work `[subj]`.	4
country	In what country is `[subj]` ?	The `[subj]` is located in `[obj]`.	101
father	Who is the father of `[subj]` ?	`[obj]` is the father of `[subj]`.	3
genre	What genre is `[subj]`?	The work titled `[subj]` belongs to the `[obj]` genre.	17
occupation	What is `[subj]`'s occupation ?	The occupation of `[subj]` is `[obj]`.	4
place of birth	In what city was `[subj]` born ?	`[subj]` was born in the city of `[obj]`.	13
religion	What is the religion of `[subj]` ?	`[subj]` practices the `[obj]` religion.	15
sport	What sport does `[subj]` play ?	The `[subj]` team plays the sport of `[obj]`.	20
P17	Which country is `[subj]` located in ?	`[subj]` is located in the country of `[obj]`.	101
P19	Where was `[subj]` born ?	According to records, `[subj]` was born in `[obj]`.	101
P20	Where did `[subj]` die ?	`[subj]` passed away in `[obj]`.	101
P36	Where was `[subj]` born ?	According to records, the capital of `[subj]` is `[obj]`.	83
P69	Where was `[subj]` educated ?	`[subj]` received their education at `[obj]`.	16
P106	What kind of work does `[subj]` do ?	`[subj]` is employed as a `[obj]` according to structured data.	14
P127	Who owns `[subj]` ?	`[subj]` is owned by `[obj]`.	24
P131	Where is `[subj]` located ?	`[subj]` is located in `[obj]`.	14
P159	Where is the headquarter of `[subj]` ?	The headquarters of `[subj]` is located in `[obj]`.	101
P175	Who performed `[subj]` ?	`[obj]` performed the song `[subj]`.	16
P176	Which company is `[subj]` produced by ?	The `[subj]` is produced by the company `[obj]`.	66
P276	Where is `[subj]` located ?	The `[subj]` took place in `[obj]`.	25
P407	Which language was `[subj]` written in ?	`[subj]` was written in the `[obj]` language.	101
P413	What position does `[subj]` play ?	`[subj]` plays in the position of `[obj]`.	14
P495	Which country was `[subj]` created in ?	`[subj]` was created in `[obj]`.	101
P740	Where was `[subj]` founded ?	`[subj]` was founded in `[obj]`.	60

Table 2: Full list of the queries that were built using synthetic context templates derived from both datasets. `[subj]` and `[obj]` serve as placeholders for subject and object entities. Bold and italic styles are used to differentiate between the two datasets (PopQA and PEQ, respectively).

In these experiments, we work with retrieved context documents where the true answers have been replaced with a counterfactual. Similarly to the setup by Meng et al. (2022), the control variable X in this case corresponds to whether the embeddings of the context subject tokens or relationship tokens have been affected by noise ($X \leftarrow 0$) or are set to their original values ($X \leftarrow 1$), As shown in Figure 1. In this second set of experiments, the average of TE (ATE) measures the shift towards the counterfactual when providing the model with uncorrupted subject or relation embeddings. The purpose of this measurement is to quantify the general impact of context subject tokens or context relation tokens on the model's decision to rely on the context or its learned parameters. The AIE in this experiment shows how a selected representation M contributes to this decision, and again we carry out causal tracing over the model to find the most impactful model components.

2.3 Datasets

We used two datasets for our study: PopQA (Mallen et al., 2023) and PrincetonEntityQuestion (PEQ) (Sciavolino et al., 2021). They both comprise entity-centric Question-Answer pairs (QAs) and include factual triples (subject, relation, object) associated with natural language queries (Kwiatkowski et al., 2019). The PopQA prioritizes popular entities and includes a variety of relations, while the PEQ focuses on sentences found in Wikipedia that are rich in presence (more than 2,000 instances) and form straightforward questions. There are more than 10,000 factual questions in each dataset.

2.4 The RAG System under Investigation: The ATLAS Model

Our investigation examines ATLAS (Izacard et al., 2023) as an example of a RAG model that integrates parametric and non-parametric components to utilize external data effectively. This integration, and the fact that ATLAS is pre-trained jointly with a retriever, makes it well suited to studying how information is synthesized in response to a prompt. For this study, we use the version of ATLAS that has been fine-tuned on Google's Natural Questions (Kwiatkowski et al., 2019). We focus on only the

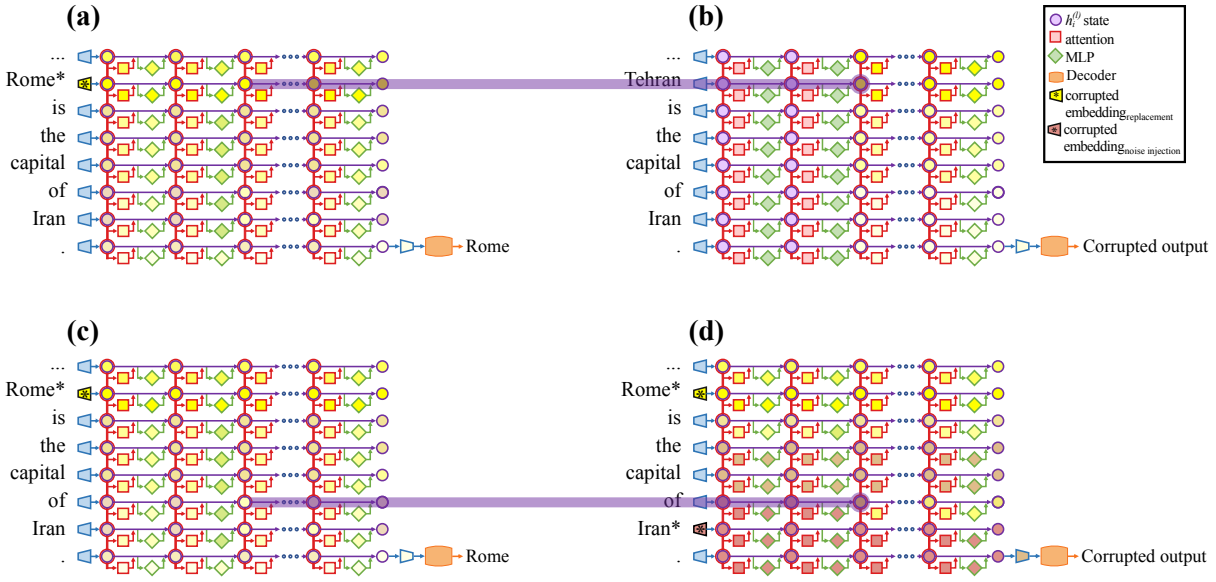

Figure 1: The first row represents the first experiment, while the second row represents the second experiment for the subjects in relation to the "restoration run". (a) shows the representations with injected counterfactual embeddings for the query "What is the capital of Iran?" (b) depicts how restoration occurs at token i and layer l. Moving on to the second experiment, which is very similar to Meng et al. (2022): (c) shows the representations when we replace the object tokens ("Tehran") with a counterfactual ("Rome"). (d) demonstrates how restoration occurs after adding noise to the subject tokens ("Iran"). We have a similar implementation for the second experiment for relations.

language model (the sequential-to-sequential component) of ATLAS, which uses the retrieved document related to the query to generate the answers.

2.5 Data Preparation

Actual retrieved documents have different quality of relations (everything in the context except the subject and object). Some retrieved documents may contain high-quality relational data that directly addresses or expands on the query, while others may introduce noise or irrelevant information. To be able to isolate the effects of relations, have fair comparison across data points, and create a consistent experimental setup, we consider the retrieved document for each query as a controlled template (a synthetic context), as shown in Table 2. In Appendix A, we show that the results on actual documents extracted using the built-in retriever in ATLAS are similar.

Following Meng et al. (2022), to ensure that the model's parametric knowledge represents the answer, we retain only those samples for which the model generates the correct answer with and without their context. The correct answer may be a substring of the actual answer, such as *Zaragoza* being a correct answer for *Zaragoza, Spain* or *Zaragoza city*. We removed relations where we had just a few data points after filtering. Table 2 shows the complete set of relations used in the experiments.

2.6 Path Specific Effects (PSE): Implementation Details

In our previous setup, we conducted experiments to investigate the impact of individual tokens at each layer concerning copying behavior and context relevance. However, we did not explore the separate effects of each module (*MLP* and *Attention*) to understand their individual contributions. To this end, we utilize the other experiment introduced by Meng et al. (2022) as PSE. We start by collecting the embeddings of each *MLP* and *Attention* module with corrupted input before restoration as the zero states (the baseline condition with corrupted input). Then, to isolate the effect of each module during the restoration, we replace the representation of the module at token i and layer l with the one we have in zero states. In simpler terms, if we want to investigate the effect of *MLP*, we first store the embedding representation of *MLP* for all tokens and layers. Then, when moving to restoration, for instance, we want to restore $Attention_l^i$ concerning token i and layer l, we do that and then restore all the *MLP* layers to the zero states that we already stored. Unlike PSE, in standard IE, we do not restore the zero states after restoration.

2.7 Causal Tracing: Implementation Details

For causal tracing, we averaged causal traces across a set of prompts for each template and over all the templates. In these experiments, we computed the AIE at three points of the transformer modules: the *hidden states* (the output of a transformer block), the *MLP*, and the *Attention*.

We follow Meng et al. (2022) and aggregate over tokens, and we extend this approach to consider the context in addition to the question. To generate the counterfactual contexts, we replaced the object tokens with another set of tokens appearing as the object in some other example in the same relation.

Following Meng et al. (2022), we implemented the *corrupted-with-restoration run* by restoring the clean run's result in six consecutive layers in *MLP* and *Attention* modules. We will later conduct PSE experiments to investigate the special role of *MLP* and *Attention* modules when we compute the impact of *hidden states*. Compared to what Meng et al. (2022) introduced, we divided the token spaces into 11 divisions for all the experiments to compute the average effect. These divisions include question, beginning of context, first subject token, middle subject tokens, last subject token, context in between tokens, first object token, middle object tokens, last object token, rest of context tokens, and last token.

3 Results and Discussion

We discuss the interpretation of the ATE, AIE, and PSE (via causal tracing) for the two experiments.

Experiment 1. Balance between parametric and non-parametric behavior The first set of experiments evaluated the model's responses when the object in the context was replaced with a counterfactual one. To understand the overall system's behavior, we categorized the results of experiments according to parametric and non-parametric. If the model consistently produces the correct answer despite counterfactual contexts, it should be classified as parametric; otherwise, it shows non-parametric behavior.

A t-test (Student, 1908) and effect size analysis (Cohen, 1988) (p-value=1.60e-4, Cohen's d=-0.9851), as shown in Figure 3, reveal statistically significant differences between these two categories, with the non-parametric subset showing much greater variability. This suggests that when the model engages in non-parametric behavior (copying from the context), it is susceptible to changes in the context. By analogy, it can be seen that the overall behavior of the model (as the general subset) is similar to the non-parametric subset, indicating a strong tendency of the model to copy from the contextual information.

Impactful tokens in copying situations The causal tracing results (Figure 2a–2c) clearly show that object tokens are the most impactful when the model is in copying mode. The AIEs are close to zero for other token positions – such as subject and relation tokens – in the context. In these cases, the model performs a form of relevance evaluation. Once the model determines the context to be relevant for answering the query, it then copies relevant object tokens into the output.

Impactful components in copying situations The causal tracing (Figure 2a–2c) across different model components (*MLP* and *Attention*) provides additional insights into how the model handles copying behavior. We observe that the object token representations flow directly through the model without being strongly affected by the surrounding context. Moreover, the *MLP* in mid-layers plays a crucial role in translating representations from the encoder to the decoder. It needs to ensure that the copied object tokens can be passed into the decoder so that they can be generated as output. Since the encoder and decoder are in different latent spaces, the *MLP* likely functions as the mechanism for this translation. This also explains that the *Attention* shows lower AIEs in copying situations compared to the *MLP*. *Attention* may play a more supportive role, ensuring that the copied tokens stay coherent with the rest of the context.

PSE analysis (Figure 4a–4c) provides a better resolution of impactful components during copying. We have observed that when the *MLP* is severed, the model's ability to rely on object tokens is substantially reduced, particularly in the mid-layers. This supports *MLP* is responsible for translating the object tokens into a form that can be passed to the decoder. The lower impact of *Attention* suggests that it is less involved in this process. Interestingly, the *MLP* also shows a similar effect for relation tokens, which may indicate that both object and relation tokens go through similar representation transformations before being passed to the decoder.

Experiment 2. Impact of rest of the context on the relevance mechanism In this experiment,

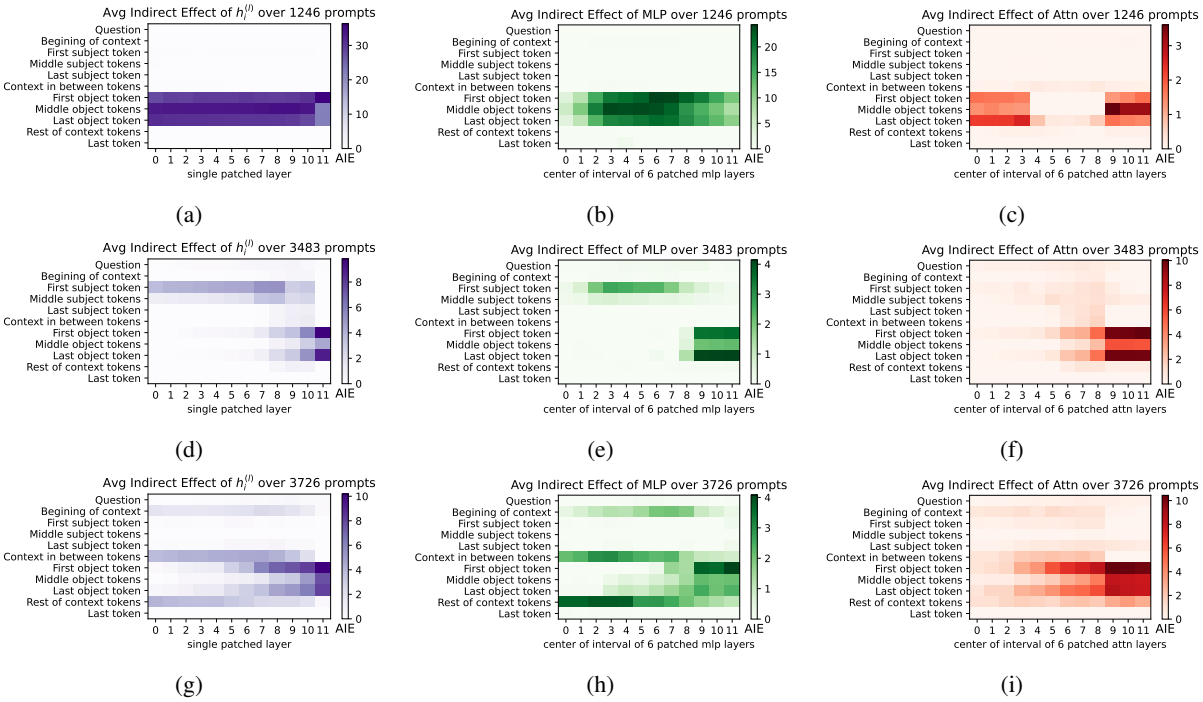

Figure 2: The figures demonstrate the AIE results of the copying behavior in ATLAS across different modules and layers. (a – c) represent the AIEs of *hidden states* ($h^{(l)}$), *MLP*, and *Attention* modules over the whole data points, which show that the object tokens are the dominant component in copying behavior. (d – i) similarly, show the AIEs for the second experiment on subject and relations tokens respectively, highlighting the vital role of these two components in determining context relevancy.

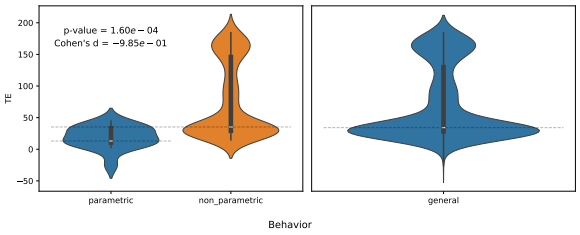

Figure 3: The left side illustrates the TE distribution across parametric and non-parametric behaviors, while the right side shows the overall distribution. The dominant distribution, represented in orange, indicates that the model's responses shift towards counterfactuals when the contexts are altered. Similarly, it reflects the model's general tendency to rely on the context to extract the answer (essentially, copying from the context).

we investigated the questions related to the earlier observations about how the model views context as relevant. We performed separate analyses on the two categories of tokens (subject and relation tokens) to determine their relative importance in relevance evaluation. The results, shown in Figure 5, indicate that while there is a statistically significant difference between the effects of subject and relation tokens (p-value=3.57e-3), the effect size is quite small (Cohen's d=-6.87e-2), meaning that both types of tokens contribute similarly to the relevance process. Interestingly, the ATE distribution for subjects shows a slightly larger spread than relation tokens, suggesting that subjects may have a marginally greater influence on relevance.

Model layers in relation to context relevance We observe an interesting pattern by looking at AIE values (Figure 2d–2i) on how the model evaluates relevance. Low AIE values for object tokens in the early layers show that the model mainly focuses on subject and relation tokens in these layers. It is as if the model is first trying to determine the relevance of the context. As the processing moves to the middle and later layers, the higher AIE values in the last layer show that the focus gradually shifts toward object tokens. The *MLP* and *Attention* are key in transitioning from relevance evaluation to object extracting. The *MLP* processes subject and relation tokens in the early layers, contributing to the context relevancy. Meanwhile, *Attention* integrates these tokens by focusing on the entire context, ensuring the model maintains coherence across the context in this process. In the later layers, the *MLP*'s role expands to help transform object token representations for the object extraction step,

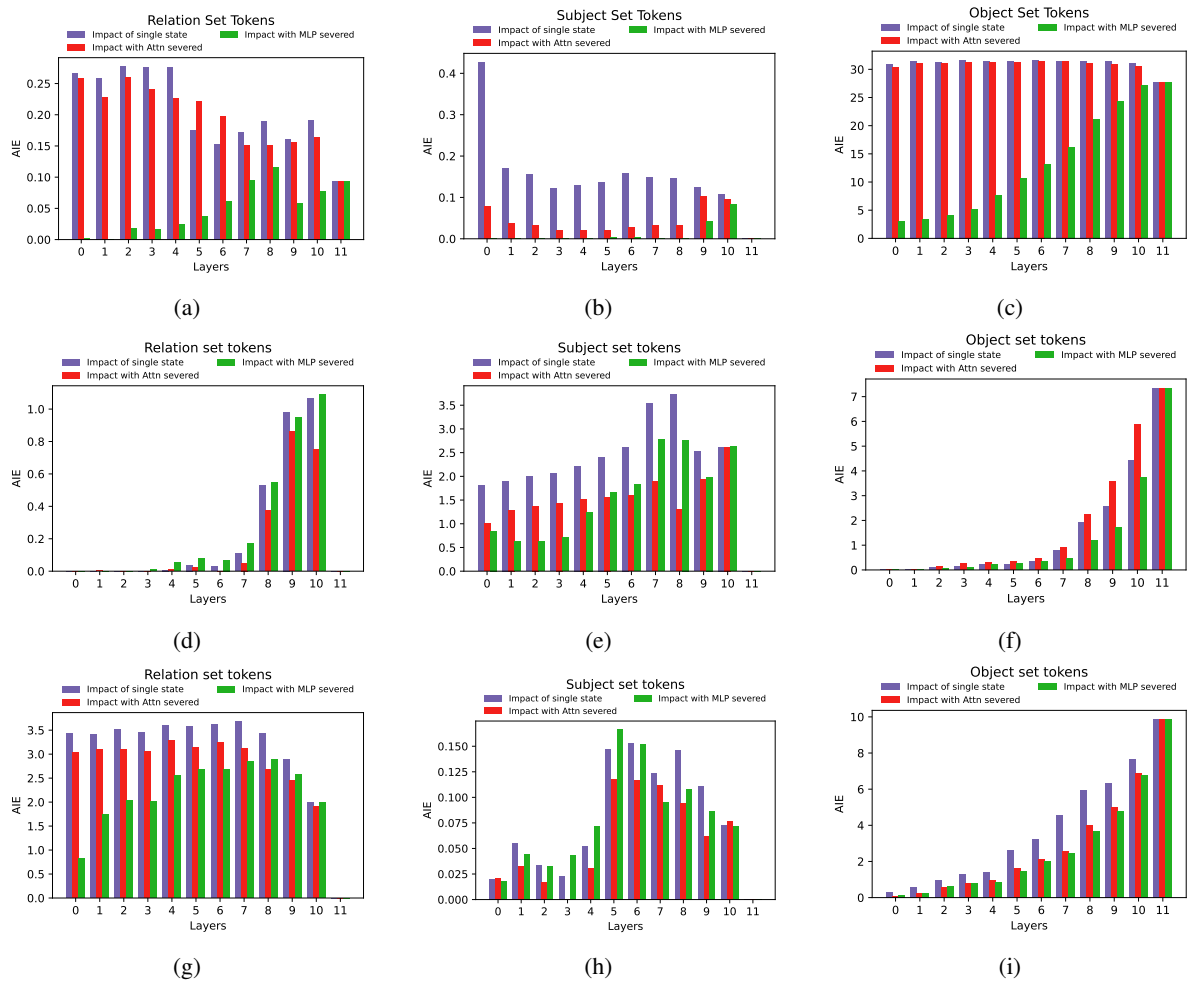

Figure 4: These figures illustrate the impact of *MLP* and *Attention* on both earlier experiments. We consider the average impact over all the subject, object, and relation tokens as set tokens. (a–c) show the contribution of *MLP* blocks from the early to the middle layers are key contributors to the model's ability to translate object token representations from the encoder space to the decoder while the *Attention* plays a minor role in the later layers. (d – i) depict the contribution of both model components from the early to the later layers, aligning with the processing of context relevance and the extraction of object tokens.

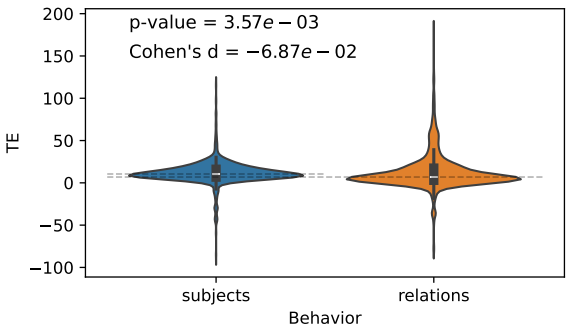

Figure 5: This plot shows the TE distribution across subjects and relation tokens.

indicating the dual role of the *MLP*.

The PSE results show that in the early layers, both the *MLP* and *Attention* work closely with subject and relation tokens (Figure 4d–4i). Put simply, the model evaluates these tokens together to figure out if the context is relevant. As we move into the later layer, the *Attention* becomes increasingly engaged with the object tokens. The change in object tokens shows that the *MLP*, which was initially focused on subject and relation tokens, now works with *Attention* to accurately extract the object tokens as the final answer.

4 Related Work

Recent research has increasingly focused on studying how Language Models (LMs) behave in specific situations. Although this field is still developing, several important studies have emerged. For example, Wu et al. (2024) look into how large language models (LLMs) handle retrieved information

that contains incorrect information. Their work involves in building small to large incorrect information to observe how models react. While parallel studies have been conducted on memory interplay in RAG models (Wadhwa et al., 2024), there is still a need to investigate the behavior of these models further. Roberts et al. (2020) show that LLMs can answer questions by using the knowledge they learned during pre-training without needing external information. Similarly, Chen et al. (2022) study how LMs remember facts when they find conflicting information from different sources. De Cao et al. (2021) suggest methods to update factual knowledge in models without needing much retraining. Their method uses a hyper-network to update the knowledge stored in the model's parameters.

In a related study, Longpre et al. (2021) explore how conflicts between contextual and parametric knowledge affect question-answering systems. Building on this idea, Wang et al. (2023) develop a way to test how well models can find and resolve conflicts in contextual information. Their results show that while models can spot conflicting information, they often struggle to identify the exact parts that are in conflict and have difficulty producing clear responses that address all the different pieces of information.

5 Conclusion

The study offers valuable insights into the inner workings of the ATLAS model – a fine-tuned RAG model – and how this model processes information from external sources (non-parametric memory) and learned parameters (parametric memory) in different queries. To clarify this phenomenon, we conducted two sets of experiments to understand how the model decides to choose copying from external sources over recalling from learned parameters.

In the first experiment, we replaced the object tokens with counterfactuals to separate the effects of context from parametric knowledge. The results revealed that the model relies heavily on the context and tends to copy from it. The second experiment revealed how the model decides to rely on non-parametric knowledge and which mechanism causes the model to choose to copy over recall. Our results suggested that the model performs a relevance evaluation to ensure that the context is useful. If that is the case, it shifts focus towards identifying entities that might be replicated as the answer (the object tokens). The subject and relation tokens are essential here in determining context relevance.

Furthermore, both experiments' results indicated that the *MLP* in the early to middle layers is involved in contextualizing the relevance of the tokens, while *Attention* in the later layers helps the model focus on integrating this information to extract object tokens as the final answer. These findings explain how these models behave and open doors to controlling this behavior in future research.

Acknowledgements

We express our gratitude to the anonymous reviewers for their insightful feedback and to Milad Malekipirbazari and Mohammad M. Ahmadpanah for their valuable input. This research was conducted in the project *Representation Learning for Conversational AI* under the Wallenberg AI, Autonomous Systems, and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation. We also appreciate the Swedish National Infrastructure for Computing (SNIC) for providing computational resources through grant agreement no. 2023/22-1025. We also would like to recognize our use of Grammarly[2] for paraphrasing and providing grammar suggestions.

Limitations

In this section, we discuss the limitations of our study for future research.

Dataset Specificity: We conducted our experiments using specific datasets (PopQA and PEQ) – parametric and non-parametric memory behavior may differ between different datasets – which may limit our findings' generalizability.

Context Manipulation: In our study, we have used counterfactuals, which might not fully capture the actual situation where the context is noisy or ambiguous.

Model Generalization: Although the ATLAS model performed well in our experiments, Its versatility to other RAG models remains unknown. ATLAS has been exposed to contexts of varying quality during training and must develop ways to adapt to poor-quality contexts, relying on parametric knowledge when necessary. It is unclear whether similar behavior can be observed in *in-context* RAG implementations based on off-the-shelf LLMs that have

[2] grammarly.com

not been trained in a RAG setup (Ram et al., 2023), or whether such models behave differently when deciding whether a context passage is relevant or not.

Temporal Relevance: There is a potential limitation when dealing with outdated or rapidly changing information based on non-parametric memory. It is necessary to understand how the model adjusts to the temporal changes in knowledge and how effectively it can choose parametric versus non-parametric memory when these changes occur.

Ethics Statement

Our work is in the area of analysis of existing models and we do not release any new model as part of this project, so we see no obvious ways to abuse the results presented here.

References

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*, 2nd edition. Routledge, New York.

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

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Lovisa Hagström, Denitsa Saynova, Tobias Norlund, Moa Johansson, and Richard Johansson. 2023. The effect of scaling, retrieval augmentation and form on the factual consistency of language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5457–5476, Singapore. Association for Computational Linguistics.

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

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *ArXiv*, abs/2005.11401.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Tobias Norlund, Ehsan Doostmohammadi, Richard Johansson, and Marco Kuhlmann. 2023. On the generalization ability of retrieval-enhanced transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1485–1493, Dubrovnik, Croatia. Association for Computational Linguistics.

Judea Pearl. 2000. *Causality: Models, reasoning, and inference*. Cambridge University Press.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, San Francisco, USA. Morgan Kaufmann Publishers Inc.

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

José M. Peña. 2023. Alternative measures of direct and indirect effects. *ArXiv*, abs/2306.01292.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Hitesh Wadhwa, Rahul Seetharaman, Somyaa Aggarwal, Reshmi Ghosh, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. 2024. From rags to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *ArXiv*, abs/2310.00935.

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence.

A Experiments on Real Contexts

We use the built-in retriever in ATLAS to fetch 20 documents per query from a given dataset for the actual context section. We then filter these documents based on specific criteria:

1. We keep only documents with one subject and one object in their context, ensuring that the object does not appear in the question and one subject also appears in the query. The reason for this filter is to ensure that it is possible to output a correct answer given a context.
2. We apply a second filter to retain only those samples for which the model can generate the correct answer even without their retrieved documents. We then expand the datapoints by considering one document from the filtered set for each question that generated the correct answer.

We previously mentioned using a context template for each relation for simplicity and complete control over the context. We conducted the first and second experiments using the actual retrieved document by the ATLAS model's retriever component, as depicted in Figure 6. In the first experiment, object tokens continue to play a dominant role in the context as part of the translation process, where the model transforms object token representations from the encoder to the decoder for output generation (Figure 6a–6c). In the second experiment, subject tokens demonstrate a significant impact on context relevance evaluation (Figure 6d–6f), helping the model decide whether the context is suitable for extracting object tokens.

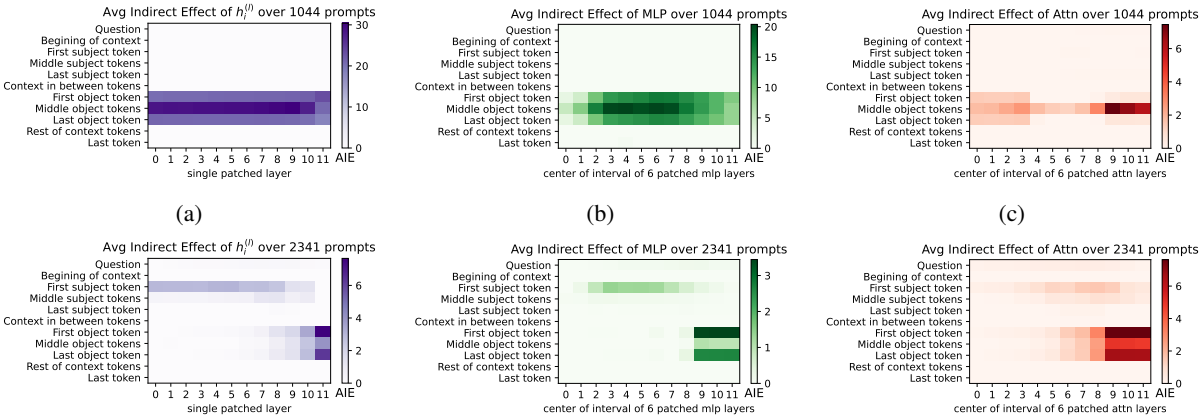

Figure 6: The figures demonstrate the AIE results of the copying behavior in ATLAS across different modules and layers for actual documents retrieved by the ATLAS model's retriever. (a – c) represent the AIEs of *hidden states* ($h^{(l)}$), MLP, and *Attention* modules over the whole prompts. (d – f) similarly, show the AIEs for the second experiment on subject tokens.