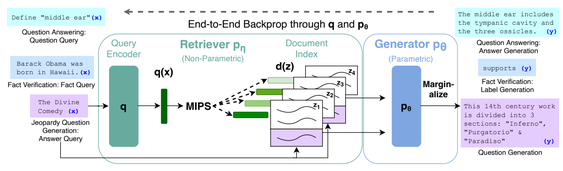
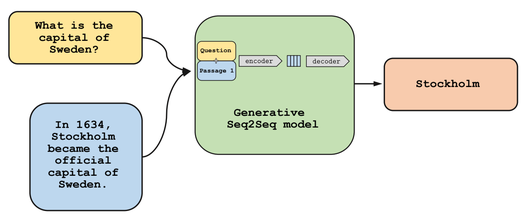
Deciphering the Interplay of Parametric and Non-parametric Memory in Retrieval-augmented Language Models

Mehrdad Farahani, Richard Johansson
Chalmers University of Technology and University of Gothenburg



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 in Sweden since 1634.	Stockholm	0.67
In 1634, Milan became the official capital of Italy.	Stockholm	0.64

Model behavior with different contexts for the question
"What is the capital of Sweden?"
(s: Sweden, o: Stockholm, r: capital of)



RAG Models

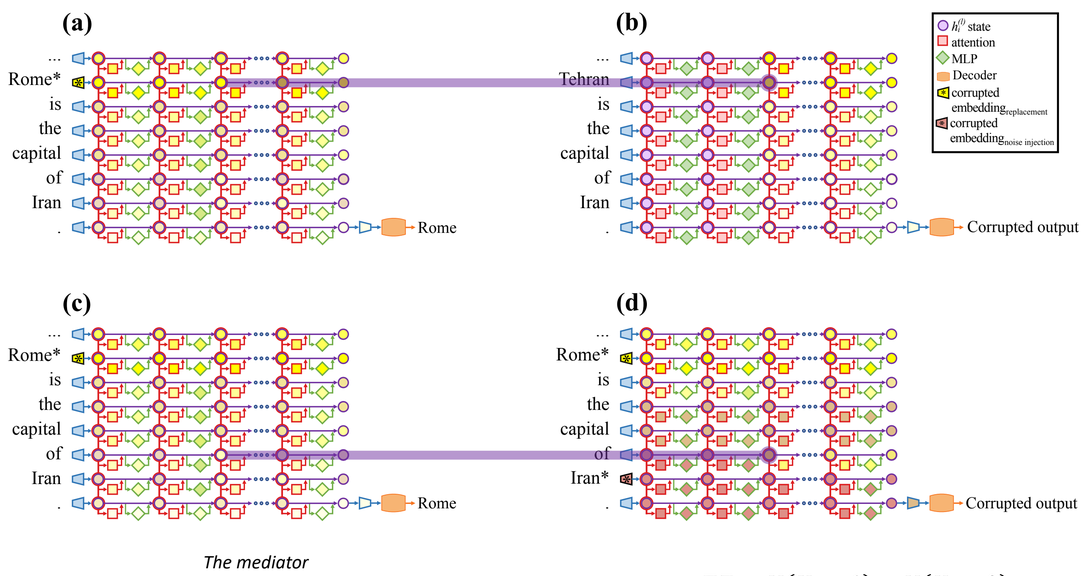
- Combine **parametric memory** (model's weights) and **non-parametric memory** (external knowledge)
- Support specialized knowledge tasks
- Improve factual accuracy through data retrieval

Atlas *the candidate model*

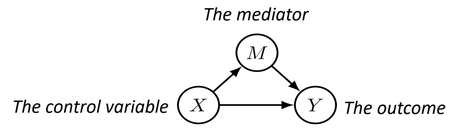
- Trains jointly on the Language Model (LM) and retriever
- Balances parametric and non-parametric knowledge efficiently

How well does Atlas balance between what it already **knows** and what it **fetches**?
To understand **when** and **why** Atlas favors one source of knowledge over the other?

RQ[1]: Which aspect of the model representation impacts the output in copying mode?
RQ[2]: What specific parts of the model trigger copying?



The mediator



*Captures how the presence of **X** changes **M**, which then affects **Y***

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

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

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

Experiments

Experiment 1: What is the balance between parametric and non-parametric behavior?

a. Modifying **object** representations with **counterfactual**
b. Keeping context unchanged ($\boldsymbol{X} \longleftarrow \mathbf{0}$)

Experiment 2: What makes the model decide to rely on the context?

a. Replacing **object tokens** with **counterfactuals** in the context
b. Adding **noise** to **subject/relation** tokens
c. Affecting subject/relation tokens with noise ($\boldsymbol{X} \longleftarrow \mathbf{0}$)

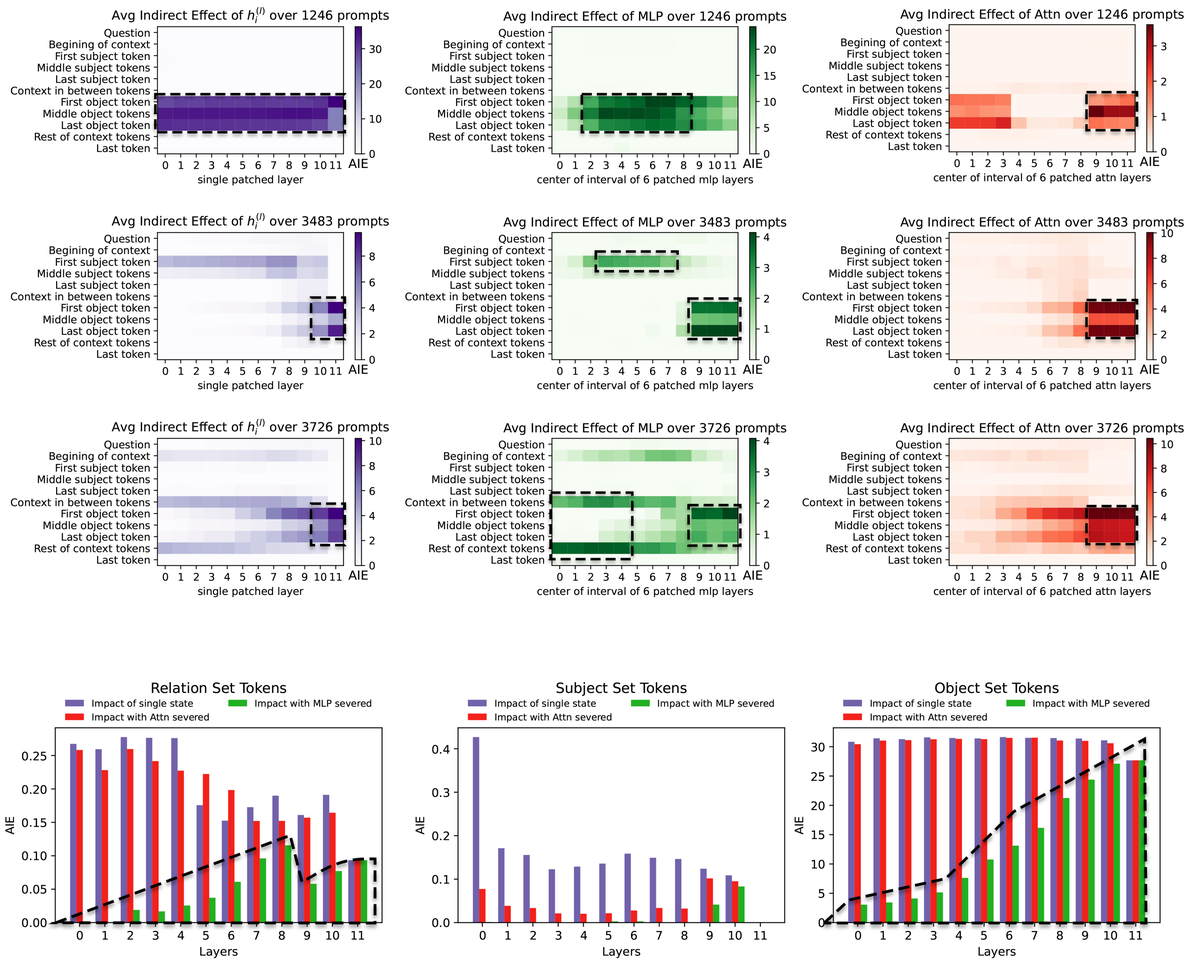
Experimental Design

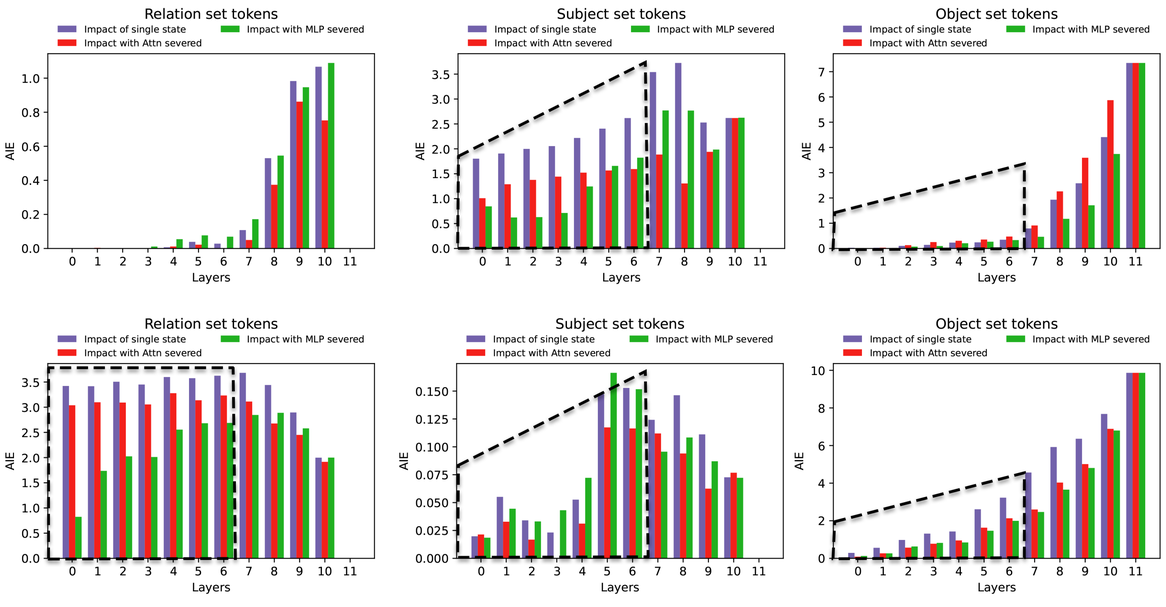
- Utilizing two datasets: PopQA and PrincetonEntityQuestion (PEQ)
- Focusing only on the LM of Atlas
- Applying synthetic context
- Retaining queries with correct answers, with and without synthetic context
- Ensuring the true answer is a substring of the actual answer
- Removing relations with few data points
- Using ***Y*** as the log ratio stabilizes values and highlights subtle shifts

Query: What is the capital of Sweden?
Context: Stockholm (Object) **is the capital of** (Relation) Sweden (Subject) .

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

Sample queries with synthetic context templates





AIE and PSE Results

Experiment 1:

- Highlights the impact of object tokens in copying mode
- The model performs a form of relevance evaluation
- MLP contributes to translating representations from the encoder to the decoder.
- Severing MLP reduces reliance on object tokens
- The lower impact of Attention suggests it is less involved in this process
- MLP shows a similar effect for relation tokens, indicating that both object and relation tokens undergo similar representation transformations for the decoder.

Experiment 2:

- In the early layers, the MLP indicates that the model focuses on subject and relation tokens
- As processing continues, the focus gradually shifts toward object tokens
- Attention focuses on the entire context and maintains coherence in this process
- MLP's role expands to help with the object extraction step
- MLP and Attention are key in transitioning from relevance evaluation to object extraction
- Both MLP and Attention work closely with subject and relations tokens in the early layers
- Changes in object tokens reveal that MLP must collaborate with Attention to extract object tokens accurately

ATE Results

- Reveal significant differences, with greater variability observed in the non-parametric subset. Comparatively, the overall behavior of the model aligns more closely with the non-parametric subset
- Indicate similar contributions from subjects and relations to relevance, with a slightly greater influence from subjects

Summary

- Exploring copying vs. recalling decisions through two experiments
- **Copying** from context by **assessing relevance** in the Atlas model
- **Contextualizing relevance and translating information** with early-to-middle MLP
- **Integrating context and extracting object tokens** with later Attention

Link to the code and paper

Limitations

- **Dataset Specificity:** Using PopQA and PEQ datasets, limiting generalizability due to differing memory behaviors
- **Context Manipulation:** Employing counterfactuals, which may not capture noisy or ambiguous contexts fully
- **Model Generalization:** Evaluating Atlas's adaptability to other RAG models, especially in varied contexts
- **Temporal Relevance:** Balancing parametric and non-parametric memory during temporal changes remains a challenge

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM