Promote, Suppress, Iterate: How Language Models Answer One-to-Many Factual Queries

Tianyi Lorena Yan, Robin Jia
{tianyi.yan, robinjia}@usc.edu

Allegro

USC University of Southern California

Paper Code

Two Subtasks of 1-to-N Factual Recall

`List three cities from` **`France`**`: 1.` **`Paris`** `2.` **`Marseille`** `3.` **`Lyon`**

- **Knowledge recall**: Given the query, retrieve relevant answers
- **Repetition avoidance**: Should not generate answers that have been generated

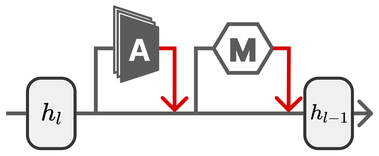
Exp Setting

Dataset	Example Template
Country-Cities	List three cities from <country>
Artist-Songs	Name three songs sung by <artist>
Actor-Movies	State three movie titles starring <actor>

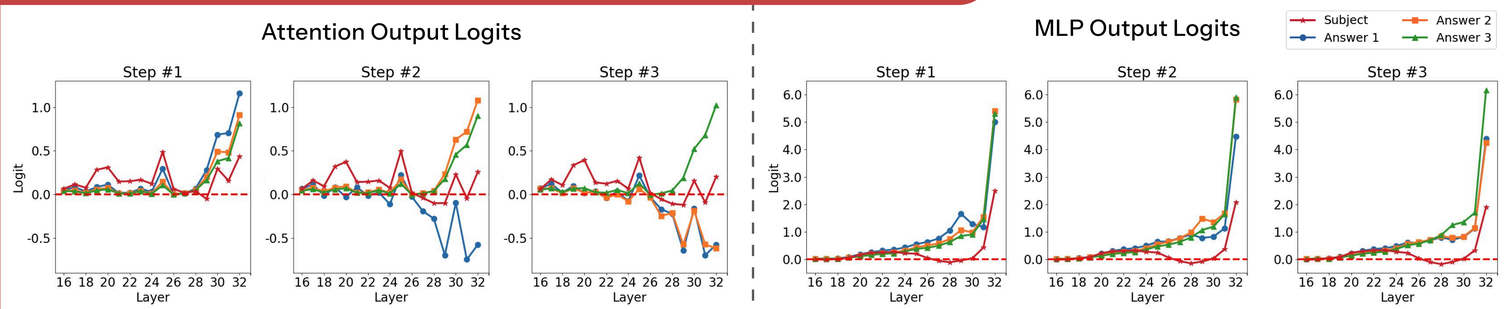
- Models: `Llama-3-8B-Instruct`, `Mistral-7B-Instruct-v0.2`
- Three templates for each model and dataset. Analyze correct cases.

① Overall Mechanism: Promote all answers then suppress the ones that have been generated

Method



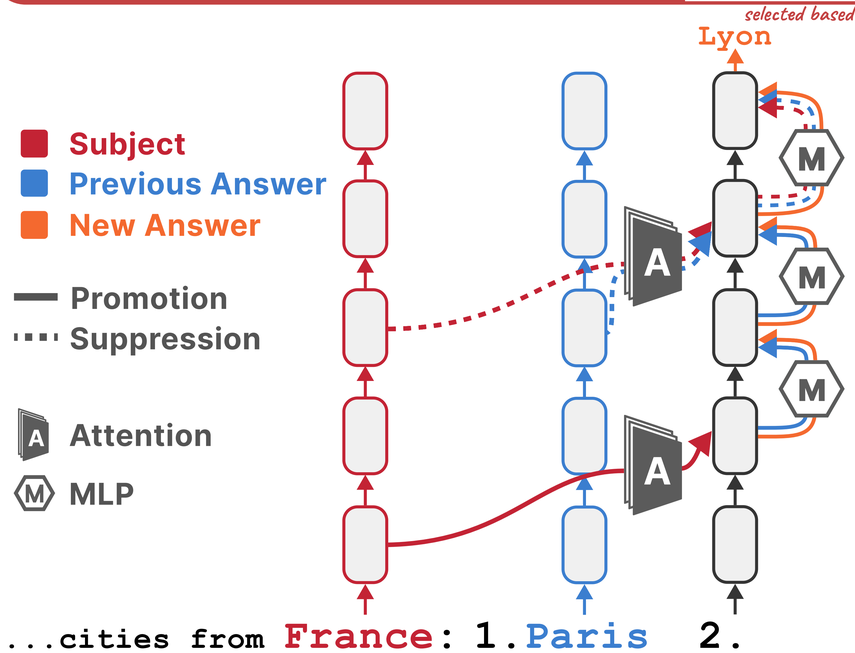
- Unembed attention/MLP output across layers
- Visualize logits of the first token of the subject & three answers
  📈 Positive: Promotion 📉 Negative: Suppression



(1) Attention propagates subject. (2) MLPs promote all answers. (3) Previous answers suppressed.

② How attention and MLPs use subject and previous answer tokens to implement the two subtasks

selected based on causal tracing results



Method

Token Lens

Attn Head Output $a^{(li)} = \mathrm{v} \cdot \mathrm{p}$
Unembed Attn Output $a^{l} = W_o^{(l)} \cdot \mathrm{Concat}(a^{(l_1)}, \ldots, a^{(l_n)})$

Attention Knockout

❌ Zero out attention score

Unembed($m^{(l)}$) — Unembed($m'^{(l)}$)