

# How Language Models Compute Negation: A Mechanistic Study of Factual Negation

Jongwook Yoon, Jongwon Lim, Yohan Jo  
Graduate School of Data Science, Seoul National University



**Negation is not a late standalone answer-flip. In factual recall, it first interacts with relation information and later weakens subject-based support for the original object.**

## 1. Why Study Factual Negation?

- LLMs often fail on simple negated factual prompts.
- Example: “Paris is not the capital of ...” may still lead to “France”.
- To better understand the source of these failures, we analyze the internal mechanisms by which LLMs process negation.

**Main question:** Does the model recall first and then flip, or does negation modify recall earlier in the pipeline?

## 2. Experimental Setup

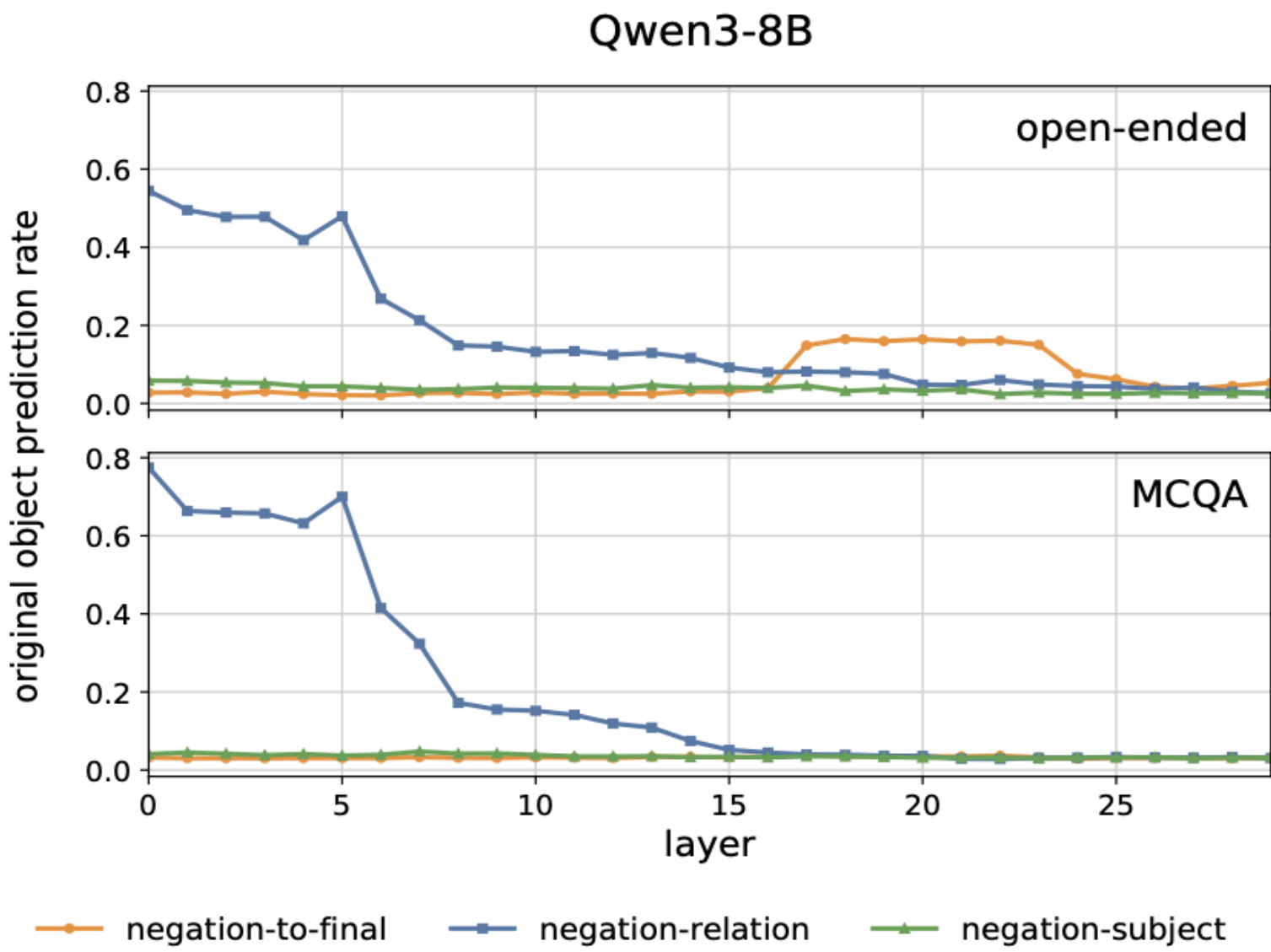
- Datasets:** CounterFact and ZSRE with minimally edited negated prompts.
- Models:** Qwen3-8B / 8B-Base; Gemma-3-12B-PT / IT
- We analyze only cases with a correct original prediction and a changed prediction under negation.
- Methods:** attention knockout, activation patching, direct logit attribution, and attention pattern analysis.

### Prediction change under negation (open-ended setting)

| Model          | CounterFact Changed (%) | ZSRE Changed (%) |
|----------------|-------------------------|------------------|
| Qwen3-8B-Base  | 36.1                    | -                |
| Qwen3-8B       | 23.4                    | 23.8             |
| Gemma-3-12B-PT | 40.2                    | -                |
| Gemma-3-12B-IT | 29.5                    | 27.2             |

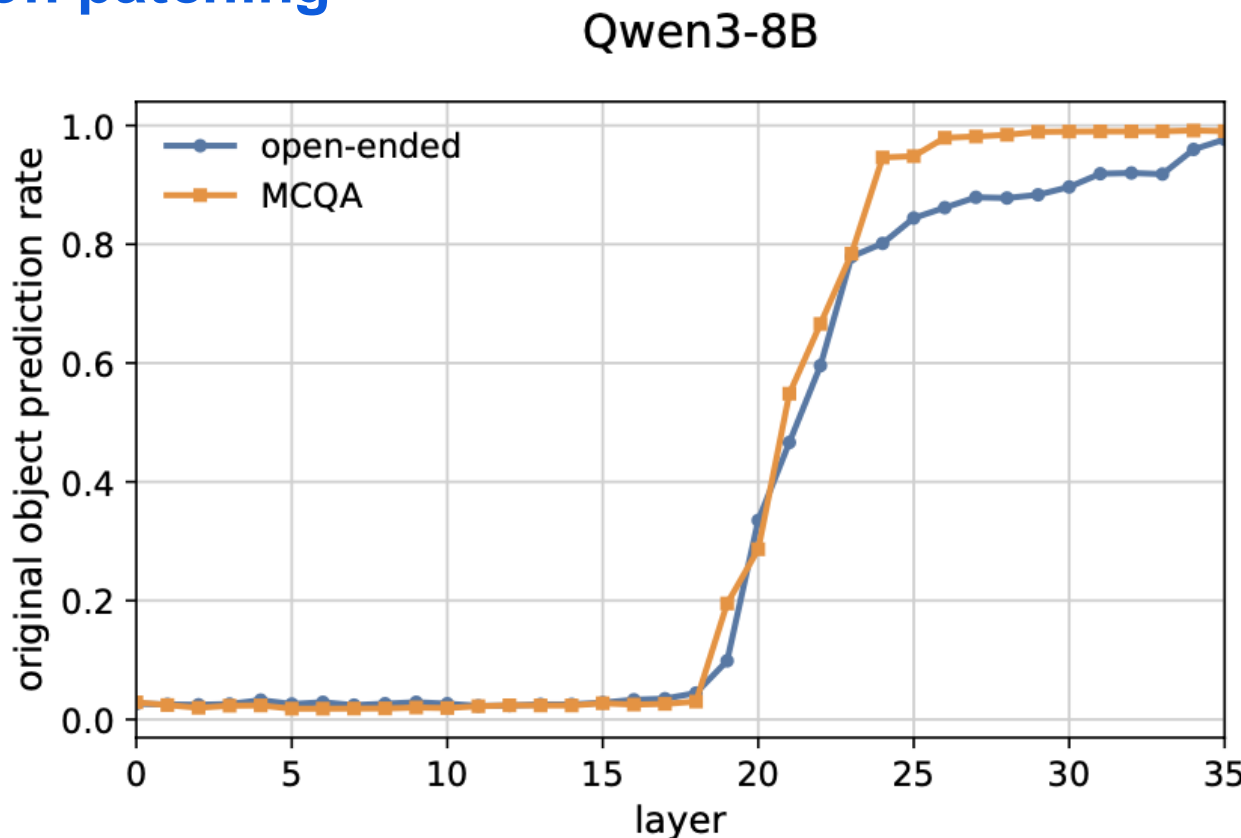
## 4. Negation Interacts with Relation in Early Layers

### Attention knockout



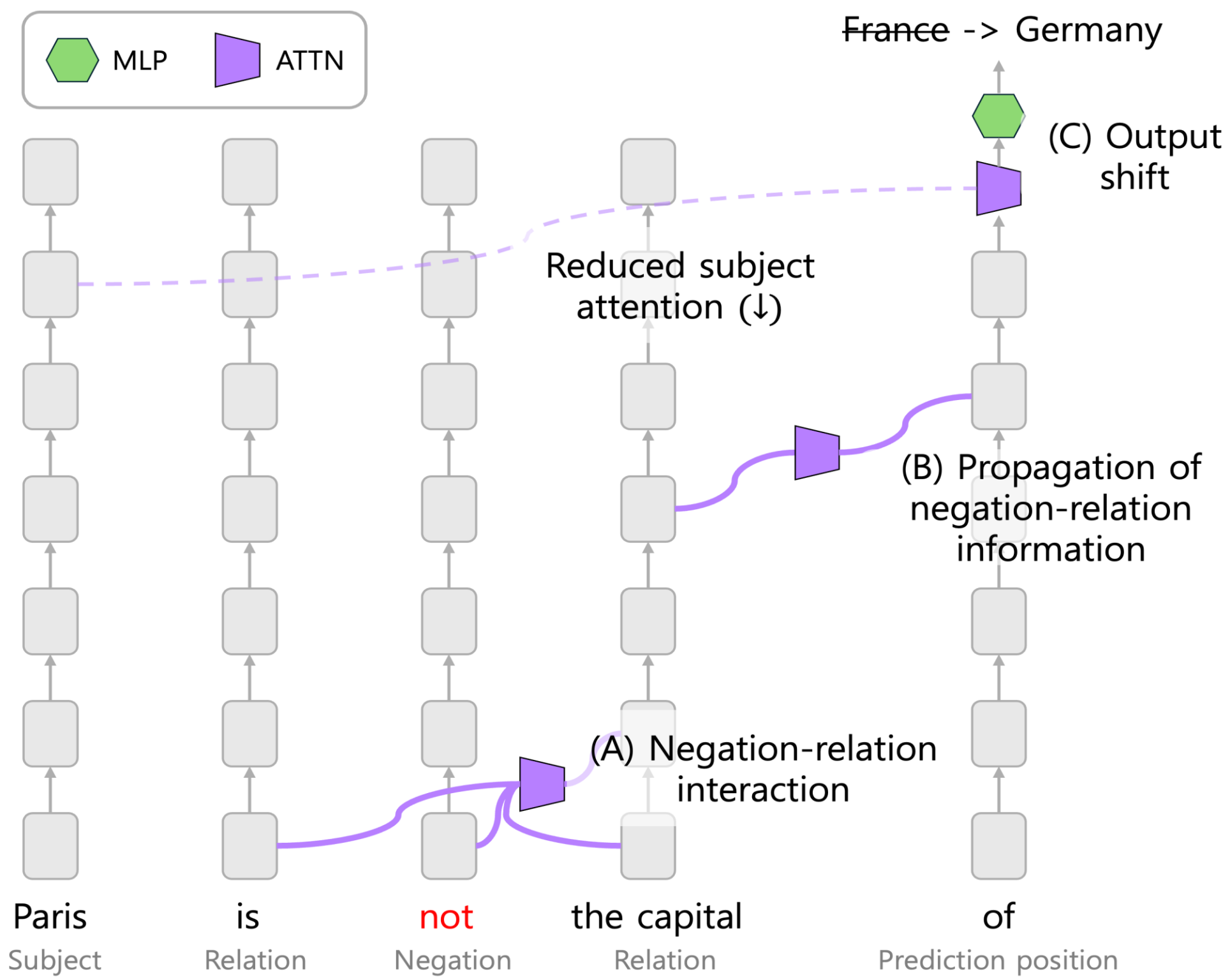
- Masking negation-relation interactions most strongly restores the original object in early layers. Direct negation-to-final access has a much weaker effect.

### Activation patching



- Replacing the final-position hidden state has little effect early, but sharply restores the original object in middle-to-late layers.
- Negation-relation information is assembled early, then propagated to the final position later.

## 3. Proposed Mechanism



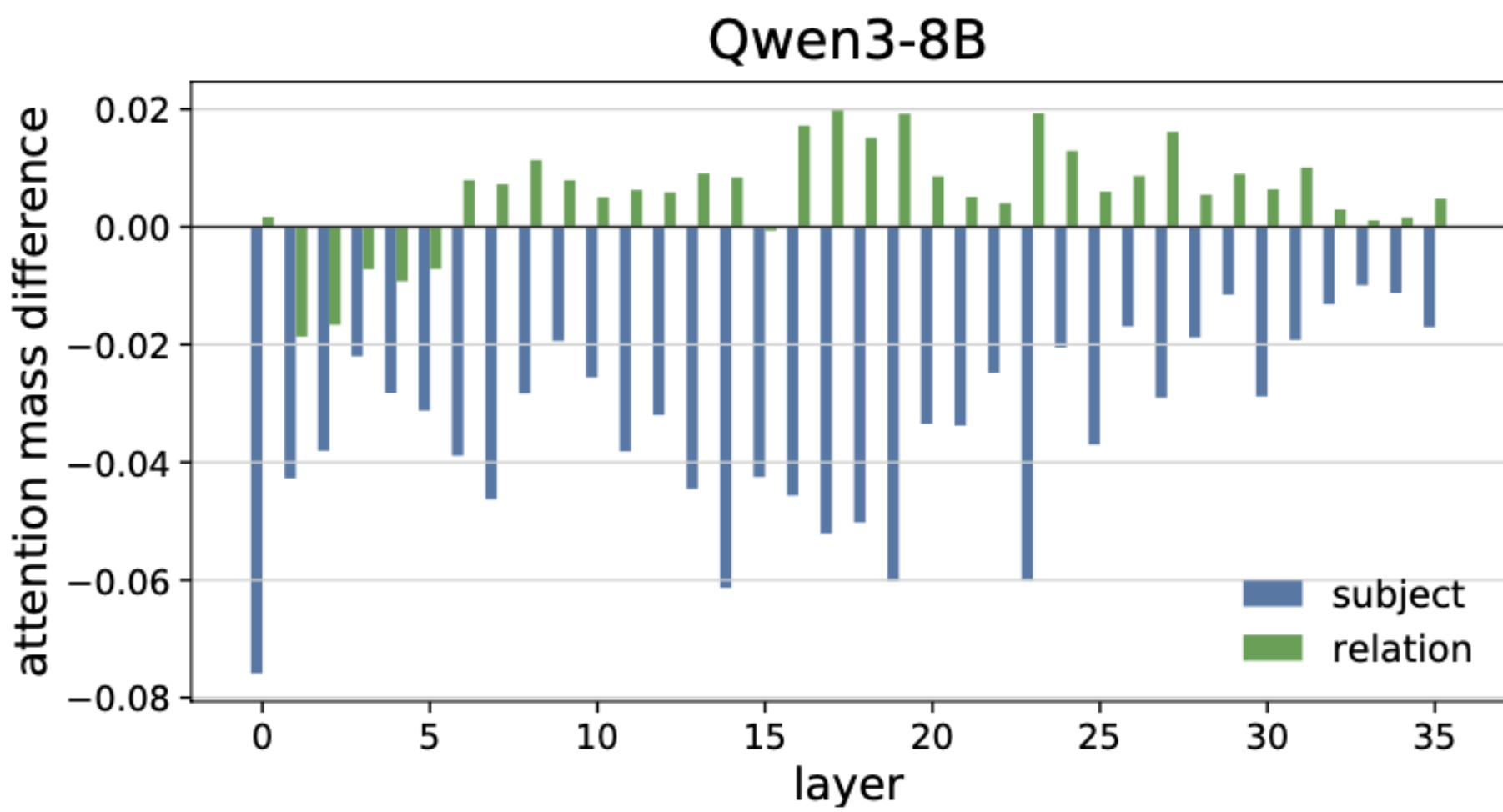
- (A) Negation interacts with relation tokens in early layers.
- (B) Negation–relation information is propagated to the final prediction position in middle-to-late layers.
- (C) Later attention and MLP components shift the output away from the original object, while reduced subject attention may weaken subject-derived support.

## 5. Output Suppression Emerges in Later Layers

### Direct logit attribution

- Direct logit attribution shows that the largest positive contributions come from middle-to-late and late attention/MLP components.

### Attention weight analysis



- Layer-wise differences in final-token attention mass between negated and original prompts for Qwen3-8B.
- Negation consistently reduces final-token attention to subject tokens.
- Potential mechanism: Reduced final-token attention to subject tokens may weaken subject-derived support for the original object.

## 6. Limitations / Future work

- Limited to two model families and changed-prediction cases, leaving unchanged failures.
- Future work should causally test the pathway and reduced subject attention.
- Broader datasets and negation forms are needed to confirm generality.