LLMs Process Lists With General Filter Heads

Arnab Sen Sharma, Giordano Rogers, Natalie Shapira, David Bau

filter.baulab.info

Specific Attn Heads focus on *filtered* items

Head [35, 19] in Llama-70B

Options: Cherry, Knife, Pants, Ambulance
Find the fruit.
Answer:

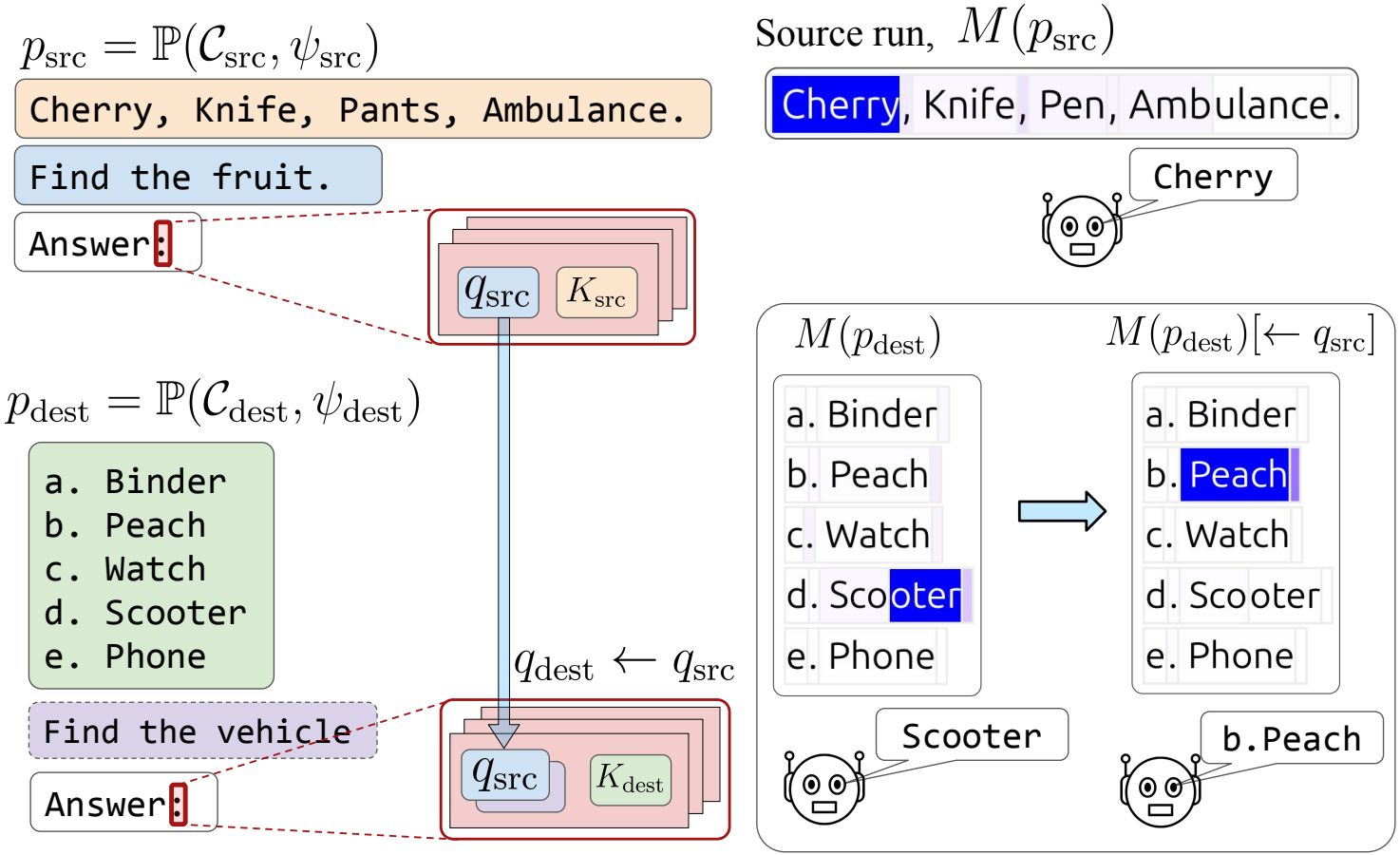
1. The Space Needle
2. Louvre Museum
3. Colosseum
4. Christ the Redeemer
5. State of Liberty
6. Big Ben
Which of these landmarks is located in Brazil?
Answer:

A random head

Options: Cherry, Knife, Pants, Ambulance
Find the fruit.
Answer:

1. The Space Needle
2. Louvre Museum
3. Colosseum
4. Christ the Redeemer
5. State of Liberty
6. Big Ben
Which of these landmarks is located in Brazil?
Answer:

We ask: How filter heads perform filtering?



Identifying Filter Heads

$$q_{-1}^{\ell j} \leftarrow \text{mask}^{\ell j} * q_{\text{src}}^{\ell j} + (1 - \text{mask}^{\ell j}) * q_{\text{dest}}^{\ell j}$$

Learn a binary mask over the heads optimizing for their functional role.

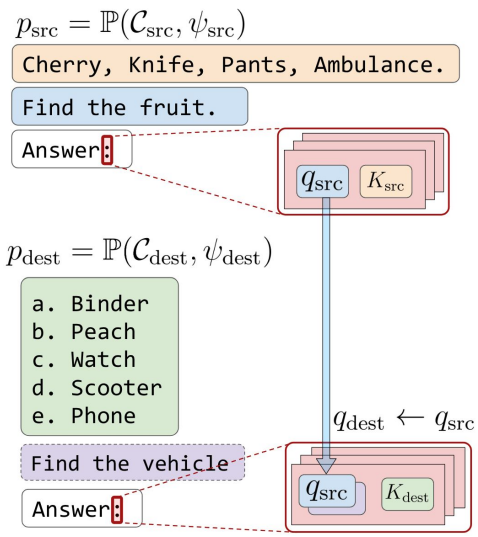
⇒ 2% of all the heads

⇒ Concentrated in the middle layers

Evaluation

For ***all*** the filter heads:

→ Transport query states from a source prompt to a destination prompt

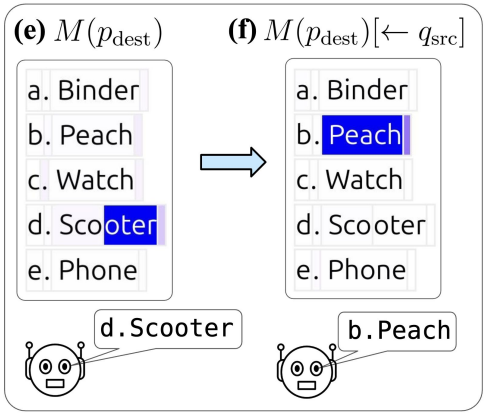


→ Check if the LM executes the source preedicate

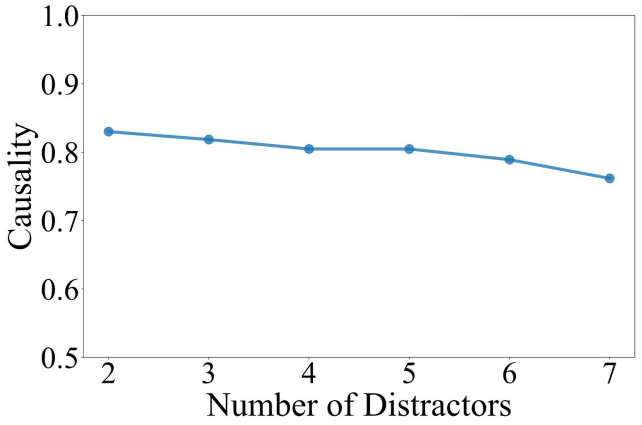
$$c^* = \underset{c \in \mathcal{C}_{\text{dest}}}{\text{argmax}} \left(M(p_{\text{dest}}) \left[q^{\ell j}_{-1} \leftarrow q^{\ell j}_{\text{src}} \mid \forall [\ell, j] \in \mathcal{H} \right] \right)_t$$

$$\text{Causality}(\mathcal{H}, p_{\text{src}}, p_{\text{dest}}) = \mathbb{1}\left[c^* \stackrel{?}{=} c_{\text{targ}} \right]$$

where $\mathcal{H}$ is the set of all selected filter heads



Generalization



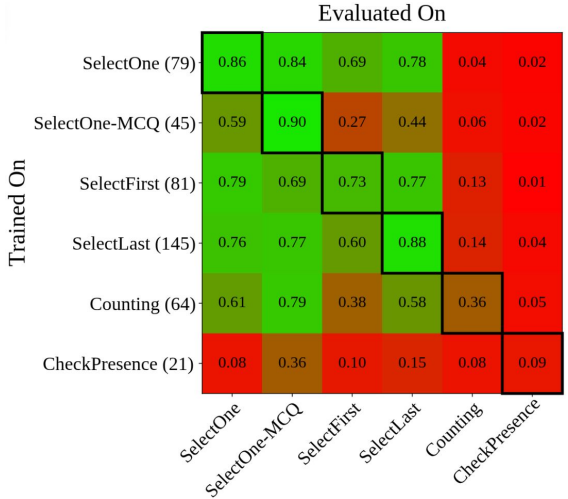
From	To				
	English	Spanish	French	Hindi	Thai
English	**0.863**	0.893	0.779	0.928	0.951
Spanish	0.857	**0.877**	0.775	0.875	0.891
French	0.938	0.932	**0.793**	0.931	0.9473
Hindi	0.920	0.920	0.885	**0.918**	0.957
Thai	0.897	0.928	0.887	0.940	**0.943**

(a) Cross-lingual transfer

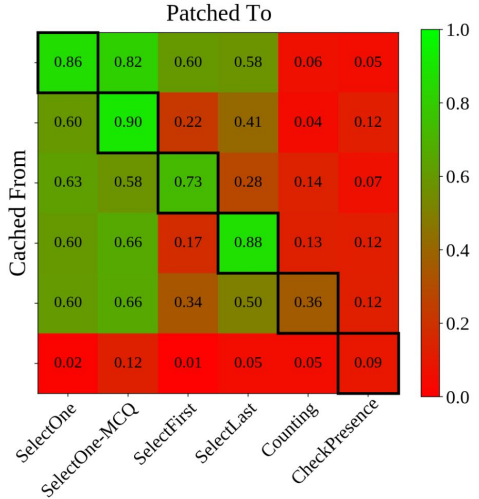
From	To	
	single line	bulleted
single line	**0.863**	0.842
bulletted	0.840	**0.848**

(b) Across option presentation style

Semantic Type	Causality	Δlogit
Object Category	0.863	+9.03
Person Profession	0.836	+7.33
Person Nationality	0.504	+5.04
Landmark in Country	0.576	+7.02
Word rhymes with	0.041	+0.65



(a) Heads evaluated across tasks



(b) Transferring q_{src} across tasks

🤔 Causality drops to ~zero in when question is presented before

```
Which one is a fruit in this list?
Options: Cherry, Knife, Pen, Ambulance.
Ans:
```

From	To	
	after	before
after	**0.863**	0.580
before	0.398	**0.020**

(c) Placement of the question

LLMs have two ways to search a list

LAZY Query presented **AFTER** items

```
(a) apple, (b) watch, (c) monkey.
Find the fruit.
Answer:
```

Map
Independent enrichment

apple ⟶

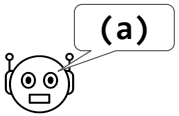
watch ⟶ 🕒

monkey ⟶

Filter
*Perform filtering, **lazily** when needed*

(a) apple, (b) watch, (c) monkey

Reduce
Reduction and formatting



Eager Query presented **BEFORE** items

```
Find the fruit.
(a) apple, (b) watch, (c) monkey.
Answer:
```

Map + Flag
Recall semantic info + eagerly check predicate

apple ⟶ ✓

watch ⟶ 🕒 ✗

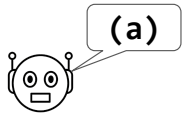
monkey ⟶ ✗

Retrieve + Format

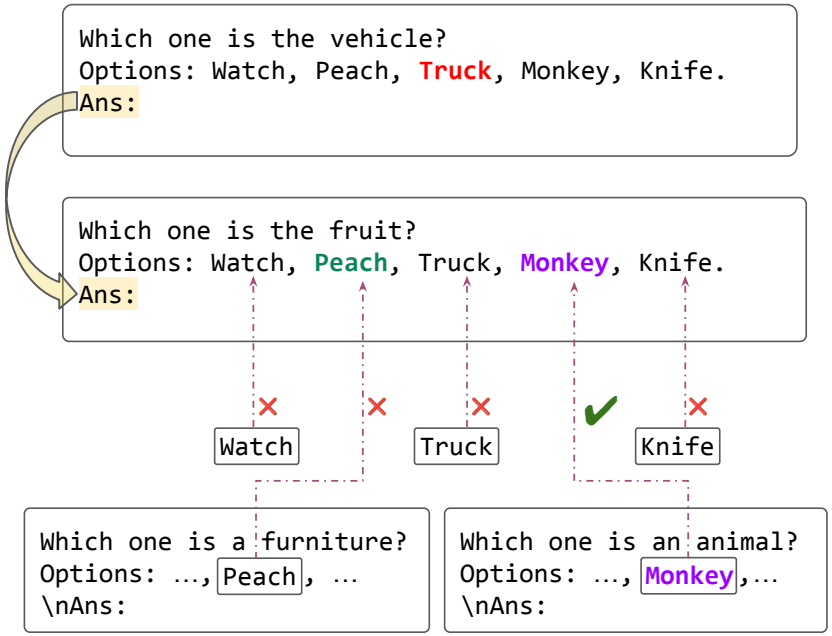
1. *Scan for ✓ flag*

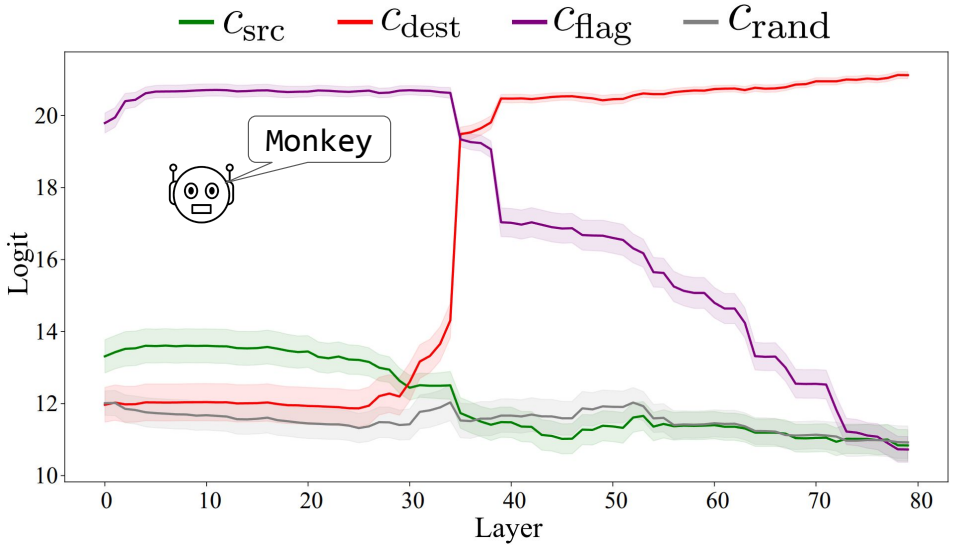
apple ✓

2. *Format and answer*



Validating Dual Implementation Hypothesis





Takeaways

- LMs perform list processing with specialized filter heads
- Predicate encoding is abstract and can trigger the same filtering op on a different context.
- LMs maintain multiple implementation for the same task: ***Lazy*** vs ***Eager*** implementation for filtering.

Visit us at ICLR to learn more!

Poster session 1, Thursday, 9:30 am

filter.baulab.info