# MIXING MECHANISMS: HOW LANGUAGE MODELS RETRIEVE BOUND ENTITIES IN-CONTEXT

Yoav Gur-Arieh♠♢, Mor Geva♠, Atticus Geiger♢♡

♠Blavatnik School of Computer Science and AI, Tel Aviv University
♢Pr(Ai)$^2$R Group
♡Goodfire

# Background

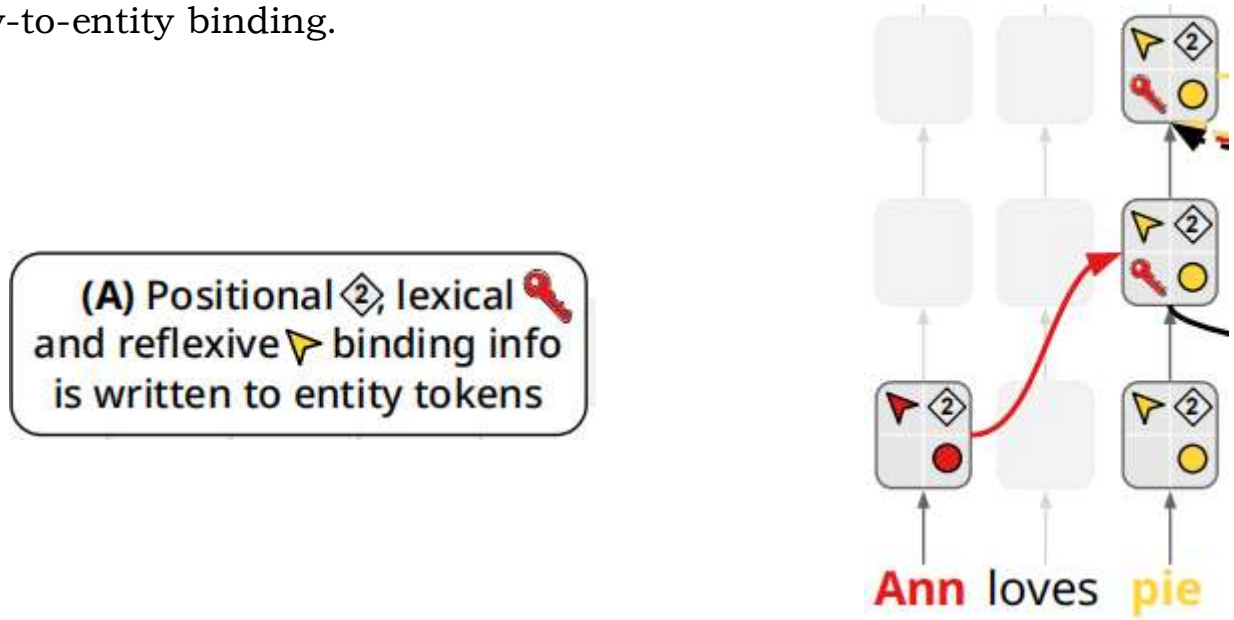


A key component of in-context reasoning is the ability of language models (LMs) to bind entities for later retrieval.

For example, an LM might represent Ann loves pie by binding Ann to pie, allowing it to later retrieve Ann when asked Who loves pie?

Previous studies have suggested that Ann is retrieved based on its position in context.In this study, the authors found, as the number of bound entities in context increases, the positional mechanism becomes noisy and unreliable in middle positions.

Therefore, the author argues that language models employ additional mechanisms beyond the position mechanism to facilitate entity-to-entity binding.



# Problem Setup



Continuing our example, define

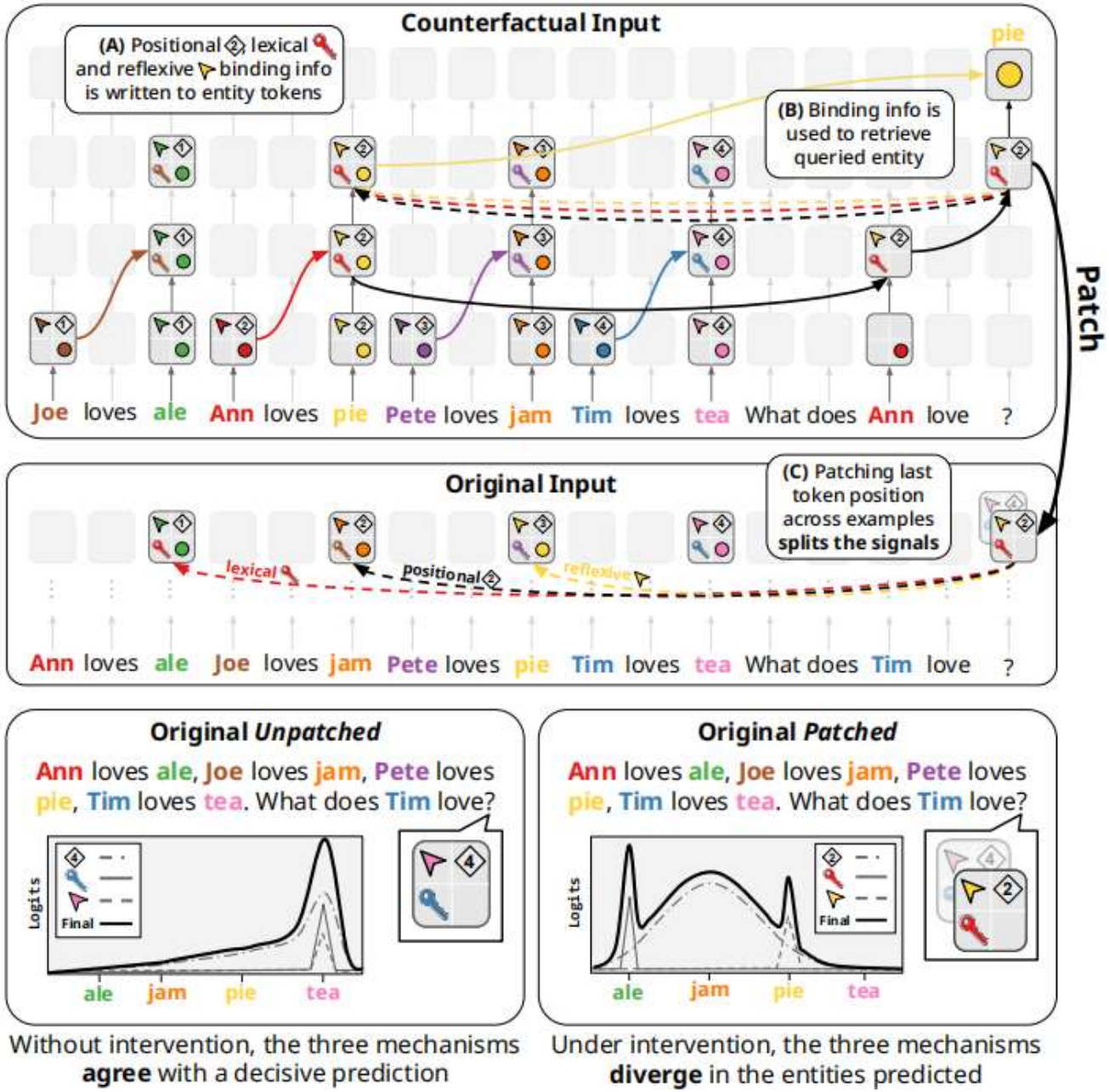
$$\mathcal{T}(\mathbf{G}, q, t) = G_1^1 \text{ loves } G_1^2, G_2^1 \text{ loves } G_2^2. \begin{cases} \textit{Who loves } q? & t_{\text{entity}} == 1 \\ \textit{What does } q \textit{ love?} & t_{\text{entity}} == 2 \end{cases}$$

and observe that

$$\mathcal{T}\left( \begin{bmatrix} \textit{Pete} & \textit{jam} \\ \textit{Ann} & \textit{pie} \end{bmatrix}, \textit{pie}, \textit{Ann} \right) = \textit{Pete loves jam, Ann loves pie. Who loves pie?}$$

- **Entity Roles:** Disjoint sets of entities $E_1, \ldots, E_m$ that will fill particular roles in a templatic text. For example, the set $E_1$ might be names of people {Ann, Pete, Tim, . . . }, and the set $E_2$ might be foods and drinks {ale, jam, pie, . . . }
- **Entity Groups:** An entity group is a tuple $G \in E_1 \times \cdots \times E_m$ containing entities that will be placed within the same clause in a template. For example, we could set $G_1$ = (Pete, jam) and $G_2$ = (Ann, pie). For convenience, we define G as a binding matrix wherein $G_i^j$ denotes the j-th entity in the i-th entity group.
- **A template (T ):** A function that takes as input a binding matrix G, the query entity q = $G_{q_{group}}^{q_{entity}}$, and the target entity t = $G_{q_{group}}^{t_{entity}}$, Here $q_{group}$ is a positional index of the entity group containing the target and query, and tentity = qentity index the positions of the target and query entities within that group, respectively

# Counterfactual Verification Method




Without intervention, the three mechanisms **agree** with a decisive prediction

Under intervention, the three mechanisms **diverge** in the entities predicted

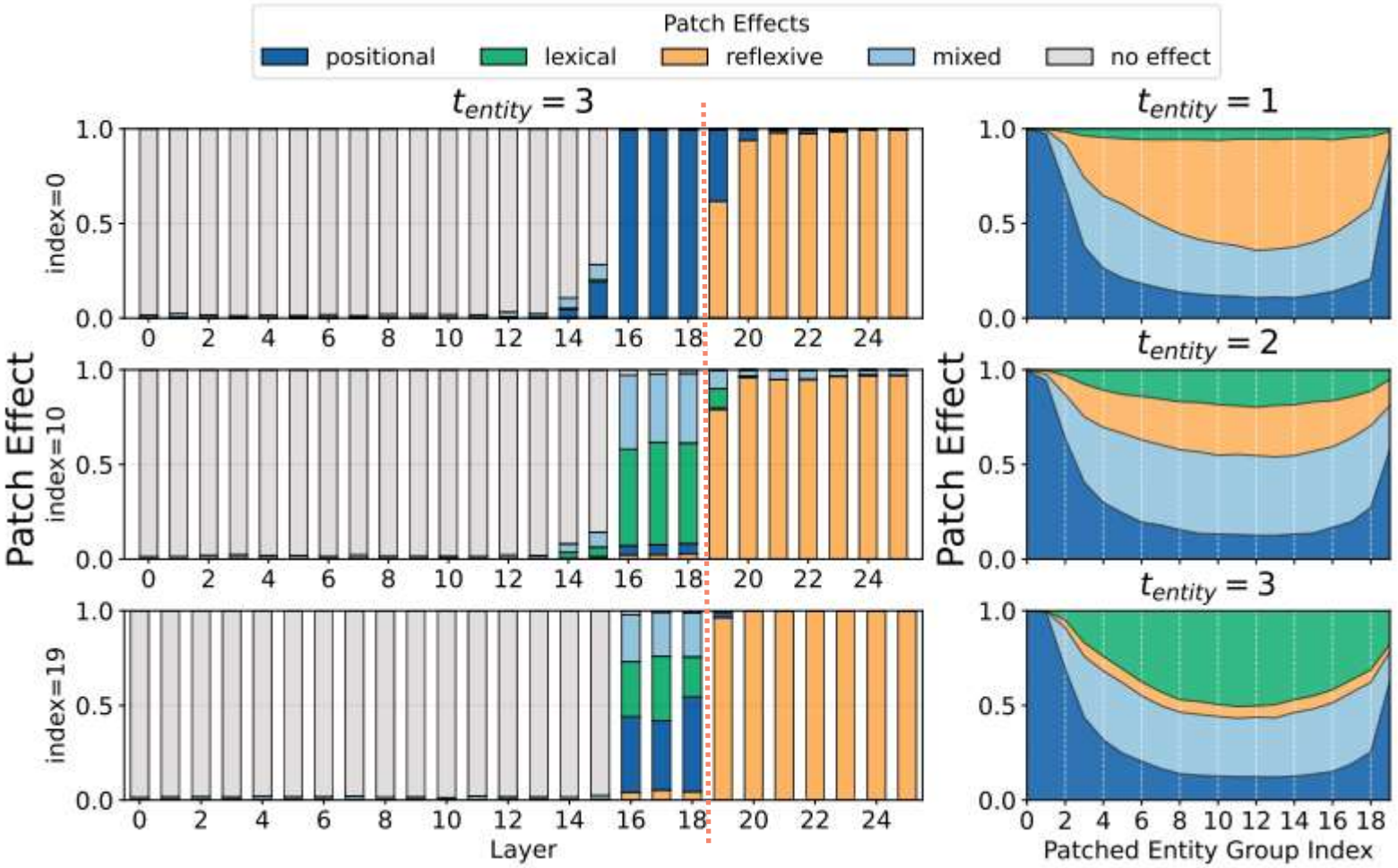
Original Input G and Counterfactual Input G’ :

$$G = \begin{bmatrix} \text{Ann} & \text{ale} \\ \text{Joe} & \text{jam} \\ \text{Pete} & \text{pie} \\ \text{Tim} & \text{tea} \end{bmatrix} \qquad G' = \begin{bmatrix} \text{Joe} & \text{ale} \\ \text{Ann} & \text{pie} \\ \text{Pete} & \text{jam} \\ \text{Tim} & \text{tea} \end{bmatrix}$$

**Three mechanisms have acquired three distinct contents, leading to three different outcomes:**

- **The Positional Mechanism(P):** The counterfactual input informs the retrieval system that the result belongs to the second group ($q'_{group}=2$). The residual stream retrieves (Joe loves jam → jam, $q_{group}=2$) after processing the original input group.
- **The Lexical Mechanism(L):** The counterfactual input group instructs the retriever to locate the group containing 'Ann'.After intervention, the model retrieves the entity linked to Ann in the original input (Ann loves ale → ale).
- **The Reflexive Mechanism(R):** In the counterfactual input, the target entity is 'pie'. After intervention, the model checks whether 'pie' appears in the original input (Pete loves pie → pie).

# Residual Current Replacement




**Left image:** The y-axis shows the indices of three representative entity groups (first, middle, and last)

The experiment revealed that in layers 16-18, the terminal labeling position carries binding information for retrieval.

**Right image:** The entity $\in \{1,2,3\}$ indicates that each clause may involve any three entities.。The U-shaped curve is characterized by the fact that the first and last indices are more dependent on the position mechanism, while the middle indices are more dependent on the lexical mechanism and the reflexive mechanism.

# Intermediate Frailty Effect of Position Mechanism



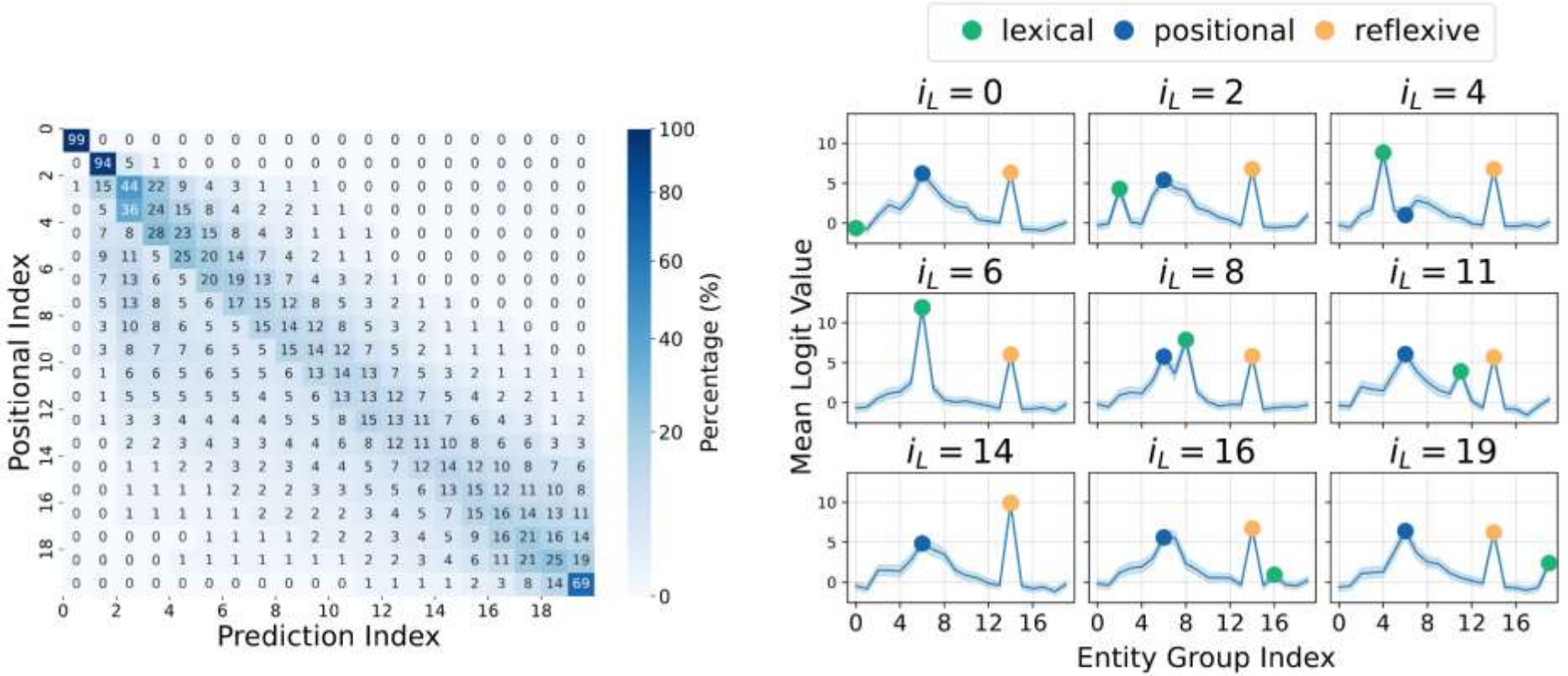



Figure 3: The positional mechanism is diffuse for middle entity groups. **Left**: Confusion matrix (%) of the patched positional index vs. gemma-2-2b-it's prediction after an interchange intervention (as in Figure 1). Counterfactual predictions cluster near the position promoted by the positional mechanism, decaying with distance. Only the *mixed* and positional patch effects from Figure 2 are shown; see Figure 28 for other models and tasks. **Right**: Mean logit distributions with $i_P = 6, i_R = 14$, and $i_L$ varied, illustrating interaction between the three mechanisms. The lexical and reflexive signals form one-hot peaks, while the positional is broader and more diffuse. These mechanisms also show additive and suppressive effects. See Figures 21, 22, and 23 for more distributions.

# Validating Of The Reflexive Mechanism



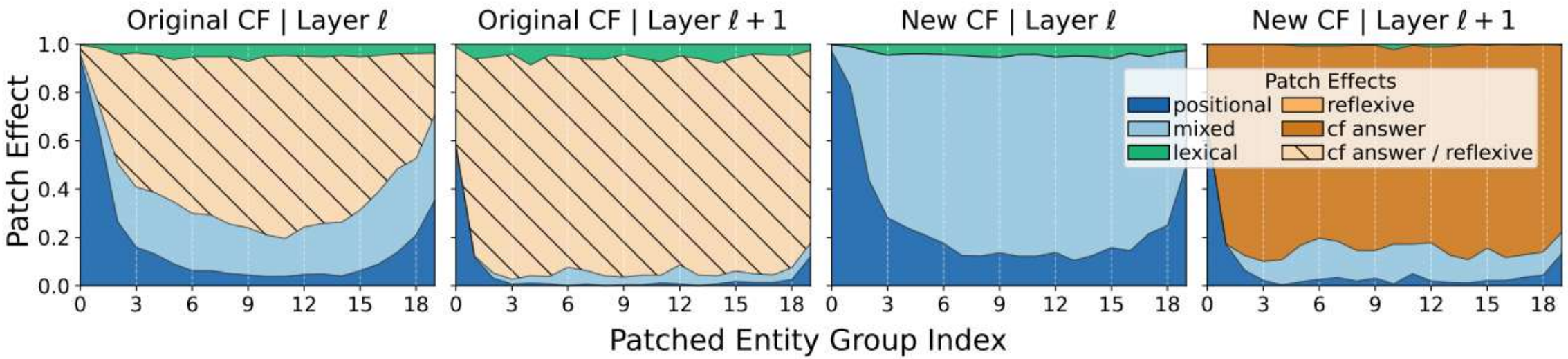



Figure 4: We distinguish the pointer in the reflexive mechanism from the answer entity with interchange interventions (gemma-2-2b-it, $t_{\text{entity}} = 1$, layers $\ell, \ell + 1$). **Left**: interventions on the original counterfactual dataset show that we can't distinguish between patching the pointer or the answer entity itself. **Right:** interventions on the modified counterfactuals (§3.4). At layer $\ell$ the model does not respond with the counterfactual answer entity which does not appear in the original context, indicating that the patched signal is a reflexive pointer that cannot be dereferenced. At layer $\ell + 1$, once the model has already retrieved the answer entity (§D.2), the patched signal becomes the answer entity itself. This shows that no confounding suppressive mechanism exists to prevent the model from answering with an entity not in its context.

# Simple Model Simulation



$$Y_i := \underbrace{w_{\text{pos}} \cdot \mathcal{N}\left(i \mid i_P, \sigma(i_P)^2\right)}_{\text{positional mechanism}} + \underbrace{w_{\text{lex}}[i_L] \cdot \mathbf{1}\{i = i_L\}}_{\text{lexical mechanism}} + \underbrace{w_{\text{ref}}[i_R] \cdot \mathbf{1}\{i = i_R\}}_{\text{reflexive mechanism}}$$

| Model | JSS ↑ | | |
|---|---|---|---|
| | $t_e = 1$ | $t_e = 2$ | $t_e = 3$ |
| *Comparing against the prevailing view* | | | |
| $\mathcal{M}$ ($L_{\text{one-hot}}; R_{\text{one-hot}}; P_{\text{Gauss}}$) | **0.95** | **0.96** | **0.94** |
| $\mathcal{P}_{\text{one-hot}}$ (prevailing view) | 0.42 | 0.46 | 0.45 |
| *Modifying the positional mechanism* | | | |
| $\mathcal{M}$ w/ $P_{\text{oracle}}$ | 0.96 | 0.98 | 0.96 |
| $\mathcal{M}$ w/ $P_{\text{one-hot}}$ | 0.86 | 0.85 | 0.85 |
| *Ablating the three mechanisms* | | | |
| $\mathcal{M} \setminus \{P_{\text{Gauss}}\}$ | 0.67 | 0.68 | 0.67 |
| $\mathcal{M} \setminus \{L_{\text{one-hot}}\}$ | 0.94 | 0.91 | 0.75 |
| $\mathcal{M} \setminus \{R_{\text{one-hot}}\}$ | 0.69 | 0.87 | 0.92 |
| $\mathcal{M} \setminus \{R_{\text{one-hot}}, L_{\text{one-hot}}\}$ | 0.69 | 0.84 | 0.74 |
| $\mathcal{M} \setminus \{P_{\text{Gauss}}, R_{\text{one-hot}}\}$ | 0.12 | 0.27 | 0.48 |
| $\mathcal{M} \setminus \{P_{\text{Gauss}}, L_{\text{one-hot}}\}$ | 0.55 | 0.41 | 0.20 |
| Uniform | 0.44 | 0.57 | 0.49 |

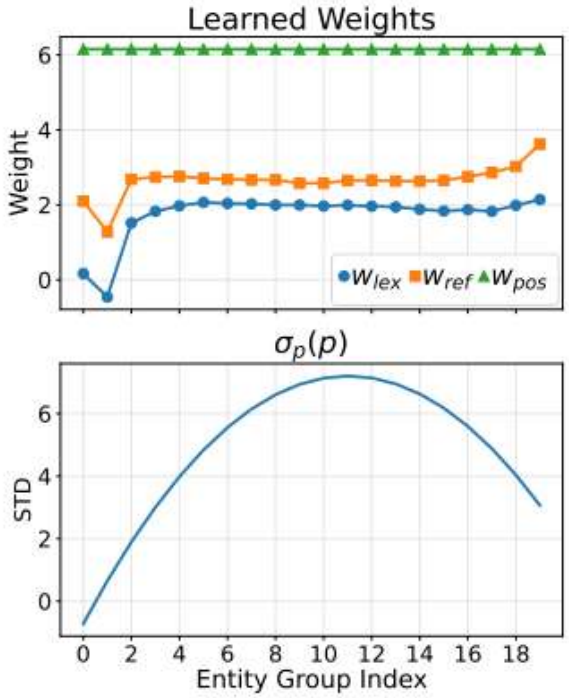



Figure 5: Results for training our full model $\mathcal{M}$ ($L_{\text{one-hot}}$, $R_{\text{one-hot}}$,$P_{\text{Gauss}}$), in addition to variants, baselines and ablations. **Left**: JSS scores for modeling the LM next token distribution over $i_P, i_L, i_R$. Evaluated on gemma-2-2b-it for the *music* binding task, with $t_e = t_{\text{entity}}$. Our model attains near-perfect JSS, slightly below the oracle. KL values (Table 3) show the same trend. All CIs are $< 0.02$; for $\mathcal{M}$ and $\mathcal{M}$ w/ oracle they are $< 0.002$. **Right**: Learned weights $w_{\text{lex}}, w_{\text{ref}}, w_{\text{pos}}$ and $\sigma$ curve, for $t_{\text{entity}} = 2$. Observe $\sigma$ widens for middle indices and narrows toward the end.

Follow Pˆıslar et al. (2025) in combining together multiple causal models (P, L, R) into a single causal model M that modulates between the mechanisms conditional on the input.

Theodora-Mara Pˆıslar, Sara Magliacane, and Atticus Geiger. Combining causal models for more ac_x0002_curate abstractions of neural networks In Fourth Conference on Causal Learning and Reasoning,
2025

# Introduction Free Form Text Into The Task



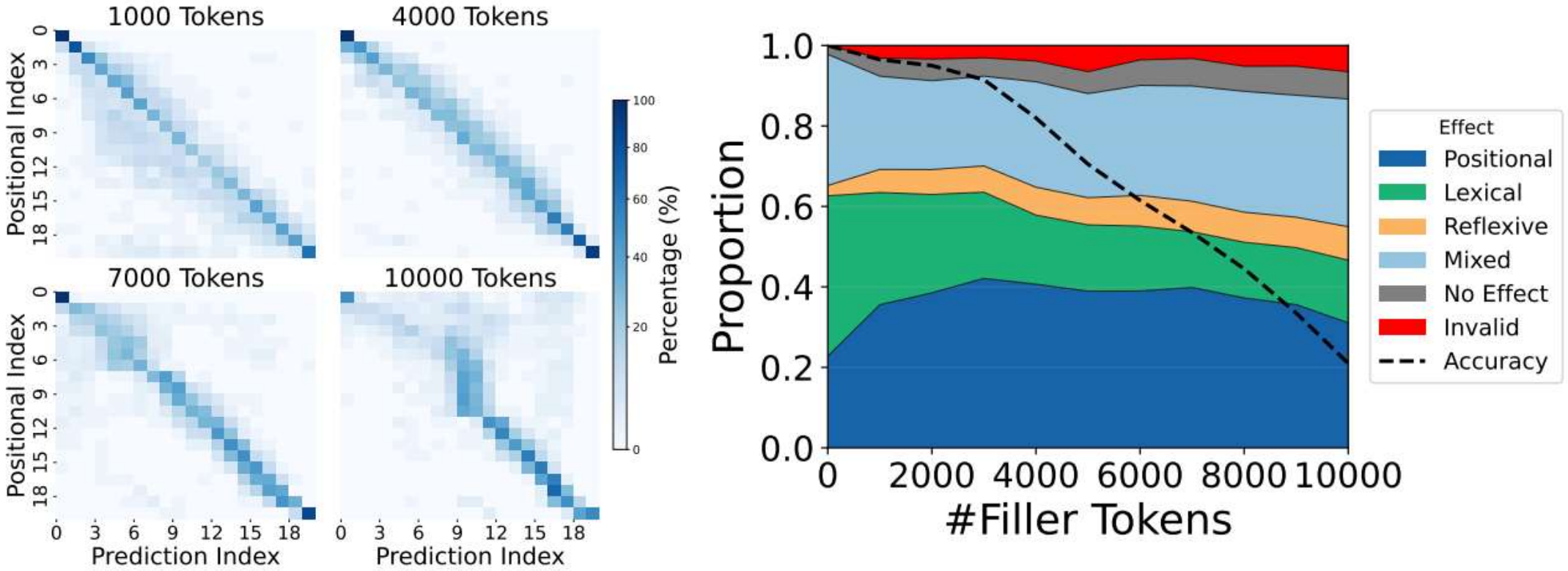



Figure 6: Padding results for gemma-2-2b-it on the *boxes* task. **Left**: Confusion matrix between the model's predicted index and the positional index patched in from the counterfactual. This gets increasingly fuzzy for early tokens as padding is increased. **Right**: Distribution of effects as padding is increased, showing the positional mechanism strengthens at the expense of the lexical mechanism.

Thanks