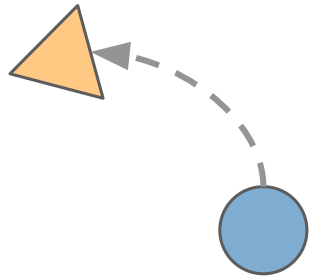
How LMs Use Pointers to Bind and Retrieve Entities

Yoav Gur-Arieh Mor Geva Atticus Geiger



LMs Are Powerful In-Context Reasoners

- Can perform in-context learning without new training
- Generalize to unseen tasks

Brown et al. 2020

Entity Binding is Necessary for Reasoning

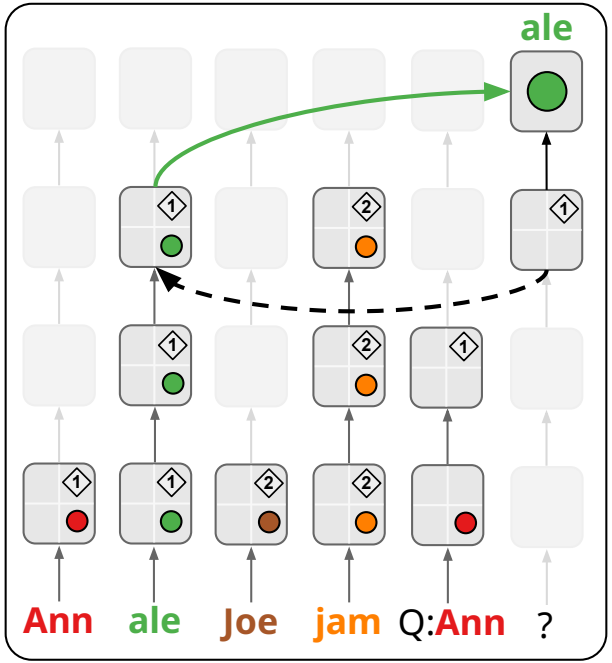
Ann loves **ale**, **Joe** loves **jam**.

What does **Ann** love?

Entity Binding is Necessary for Reasoning

What does **Ann** love?

LMs Rely on Position for Binding

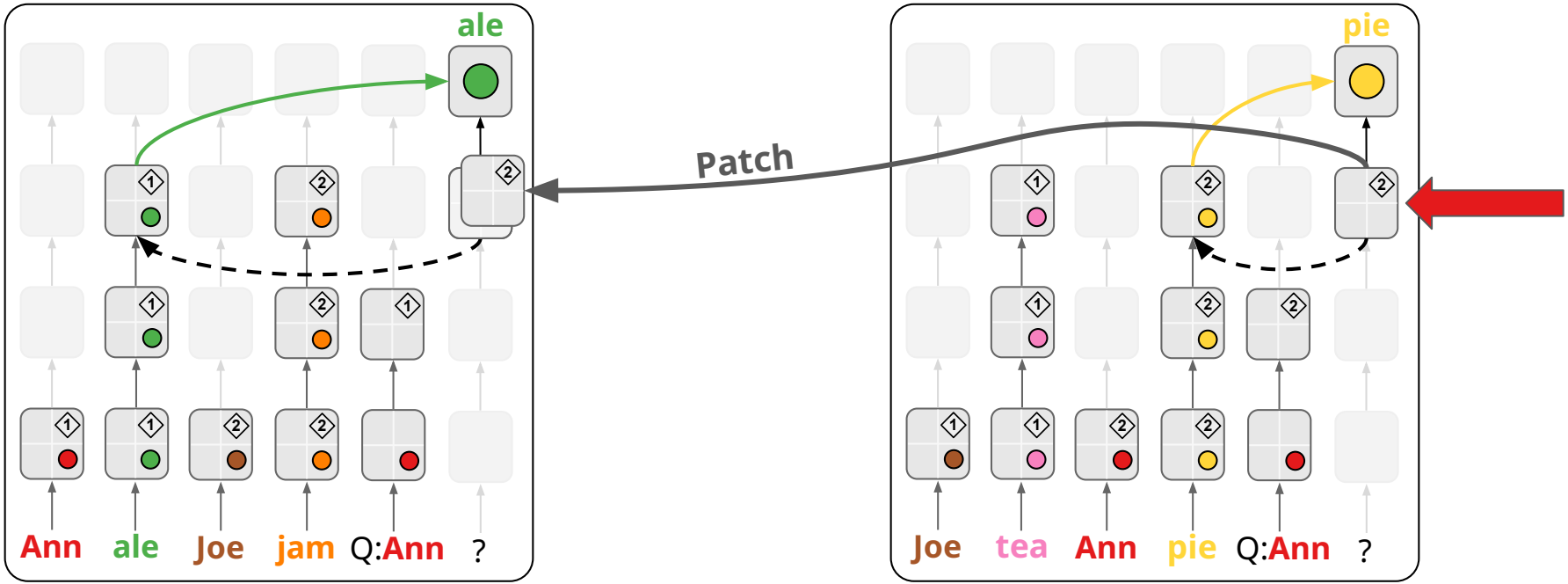


Ann loves **ale**, **Joe** loves **jam**. What does **Ann** love?

original

Prakash et al. 2025

LMs Rely on Position for Binding



Ann loves **ale**, **Joe** loves **jam**. What does **Ann** love?

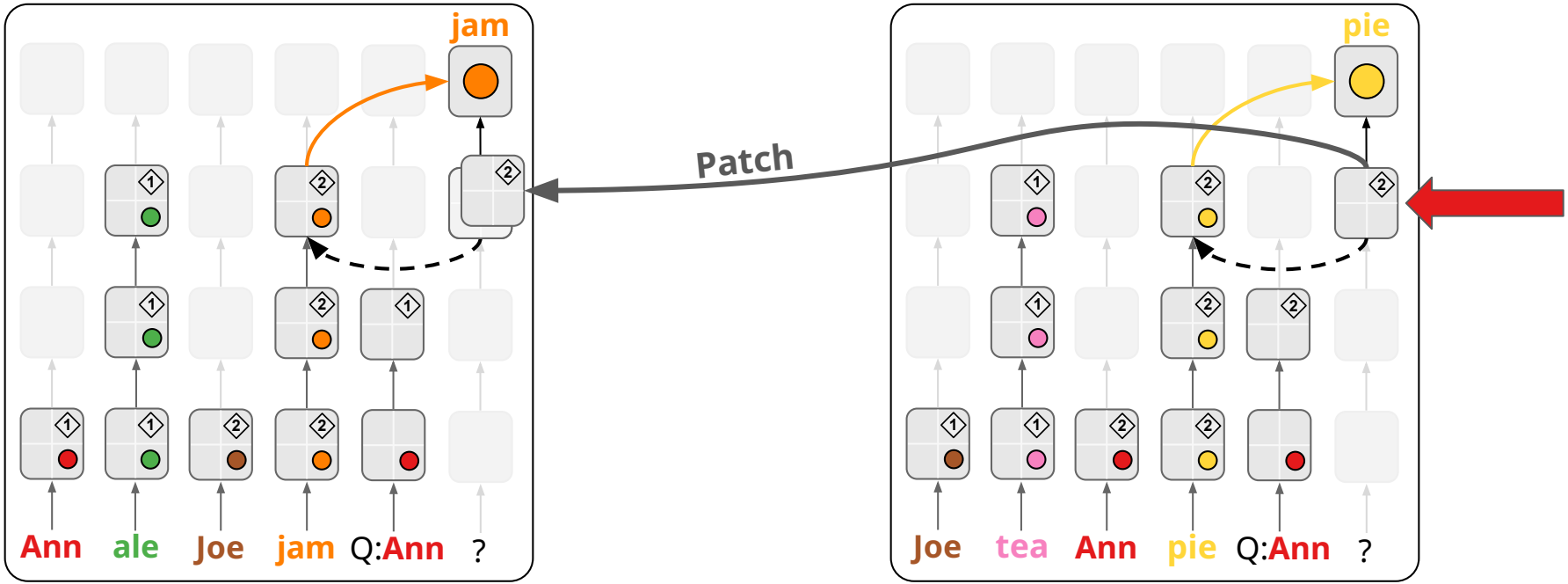
original

Joe loves **tea**, **Ann** loves **pie**. What does **Ann** love?

counterfactual

Prakash et al. 2025

LMs Rely on Position for Binding



Ann loves **ale**, **Joe** loves **jam**. What does **Ann** love?

original

Joe loves **tea**, **Ann** loves **pie**. What does **Ann** love?

counterfactual

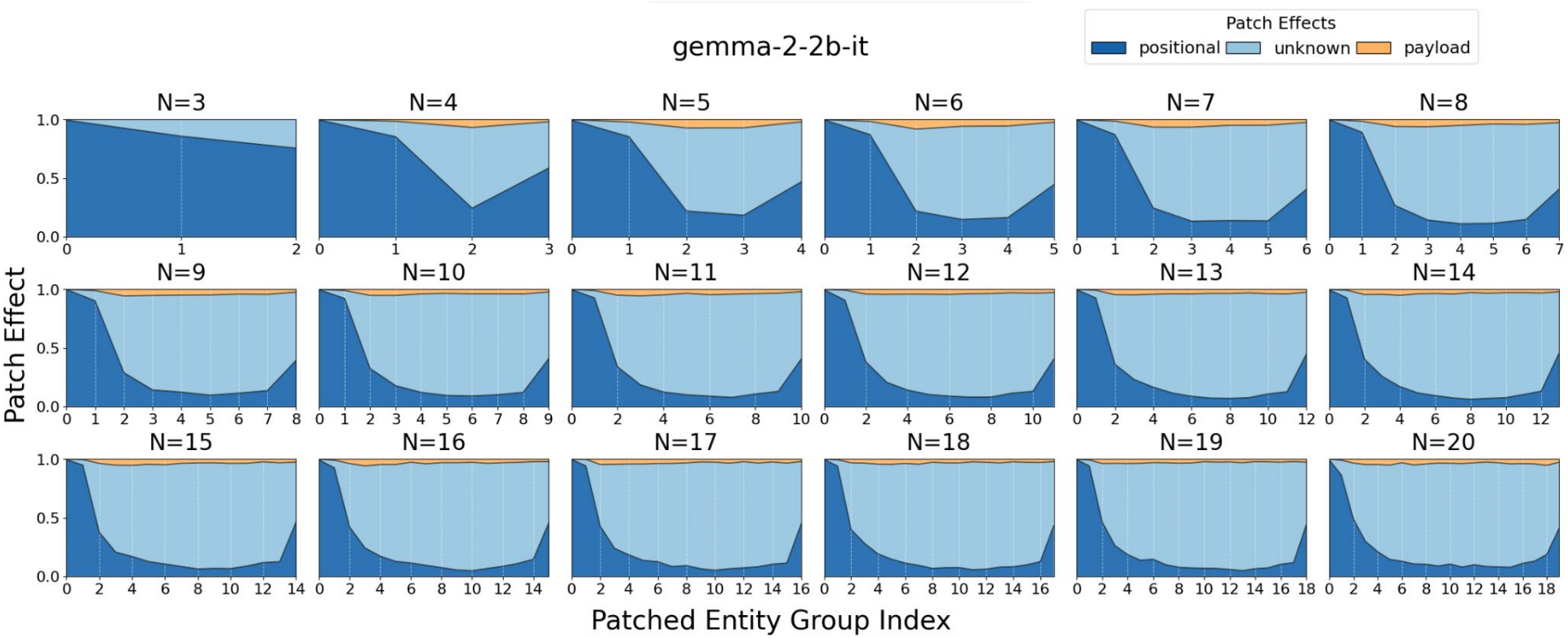
Prakash et al. 2025

Is this always true?

The setup key limitations:

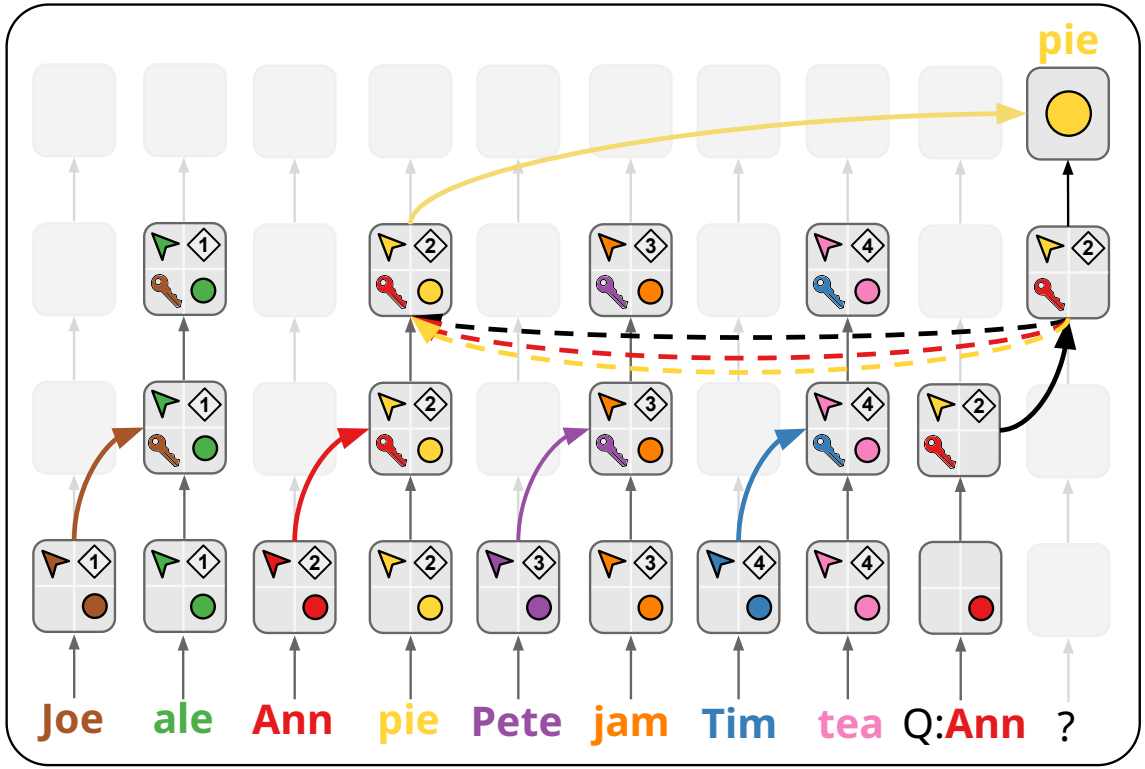
- Two models
- Single binding task
  - Each entity group always has three entities
- Always query the last entity within a group
- **Always two entity groups (N)**

Evaluating with more entity groups



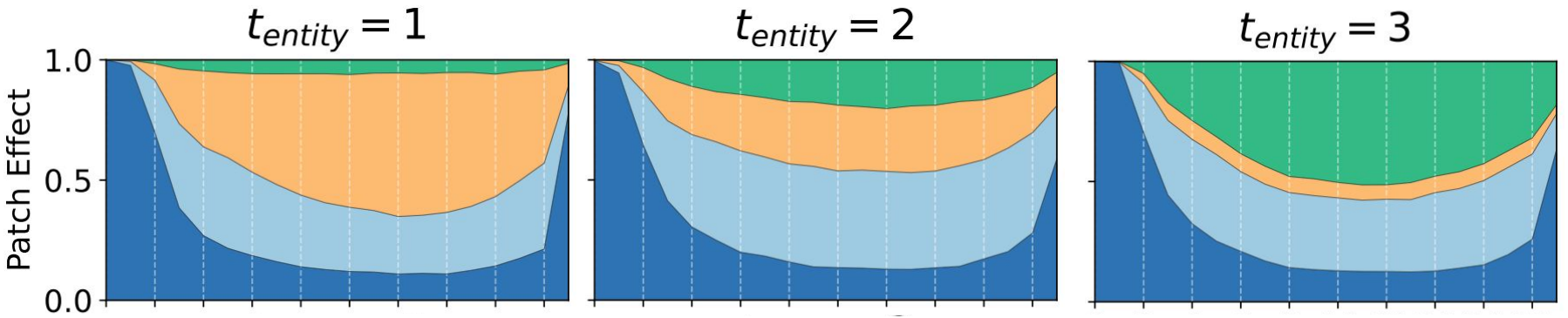
LLMs Mix Mechanisms

To bind entities, we find that LMs rely on **positional**, **lexical** and **reflexive** mechanisms.



How The Mechanisms Interact

Binding in **first** and **last** entities is mainly **positional**, but **middle** mixes **lexical** and **reflexive**.



Causal Model

- Created a simple causal model predicting model behavior

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

Causal Model

- Created a simple causal model predicting model behavior
- Achieved 95% agreement

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

Causal Model

- Created a simple causal model predicting model behavior
- Achieved 95% agreement
- Prevailing view achieved 44%

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

Takeaways

- LMs struggle to rely on positional information for binding (lost in the middle)
- Thus LMs rely on three mechanisms for binding: ***positional***, ***lexical*** and ***reflexive***.
- Findings are robust across 9 models (2B-72B params) and 10 binding tasks.

Paper

Interactive blog post