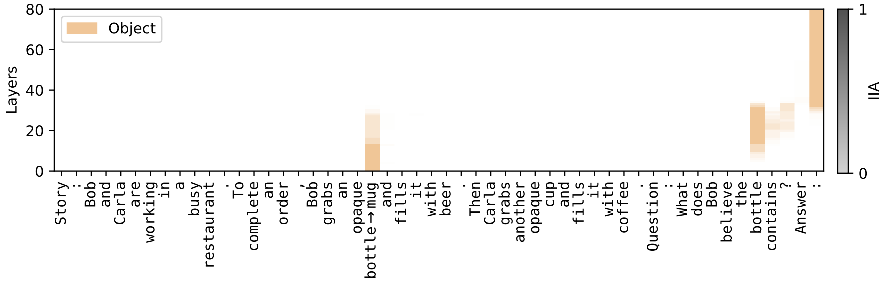

Figure 11: Information flow of object input tokens using causal mediation analysis.

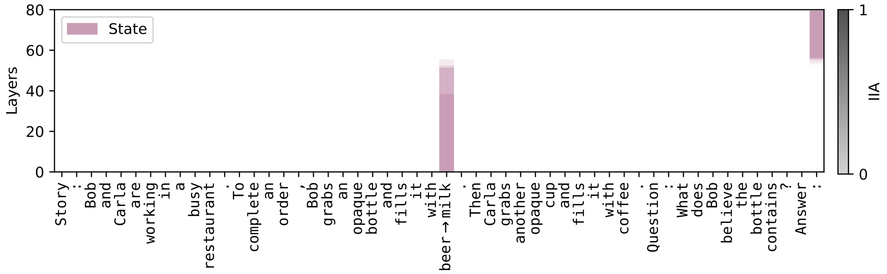

Figure 12: Information flow of state input tokens using causal mediation analysis.

## D Desiderate Based Patching Via Causal Abstraction

**Causal Models and Interventions** A deterministic causal model $\mathcal{M}$ has *variables* that take on *values*. Each variable has a *mechanism* that determines the value of the variable based on the values of *parent variables*. Variables without parents, denoted $\mathbf{X}$, can be thought of as inputs that determine the setting of all other variables, denoted $\mathcal{M}(\mathbf{x})$. A *hard intervention* $A \leftarrow a$ overrides the mechanisms of variable $A$, fixing it to a constant value $a$.

**Interchange Interventions** We perform *interchange interventions* (Vig et al., 2020; Geiger et al., 2020) where a variable (or set of features) $A$ is fixed to be the value it would take on if the LM were processing *counterfactual input* $\mathbf{c}$. We write $A \leftarrow \mathsf{Get}(\mathcal{M}(\mathbf{c}), A)$ where $\mathsf{Get}(\mathcal{M}(\mathbf{c}), A)$ is the value of variable $A$ when $\mathcal{M}$ processes input $\mathbf{c}$. In experiments, we will feed a *original input* $\mathbf{o}$ to a model under an interchange intervention $\mathcal{M}_{A \leftarrow \mathsf{Get}(\mathcal{M}(\mathbf{c}),A))}(\mathbf{o})$.

**Featurizing Hidden Vectors** The dimensions of hidden vectors are not an ideal unit of analysis (Smolensky, 1986), and so it is typical to *featurize* a hidden vector using some invertible function, e.g., an orthogonal matrix, to project a hidden vector into a new variable space with more interpretable dimensions called "features"(Mueller et al., 2024). A feature intervention $\mathbf{F}_{\mathbf{h}} \leftarrow \mathbf{f}$ edits the mechanism of a hidden vector $\mathbf{h}$ to fix the value of features $\mathbf{F}_{\mathbf{h}}$ to $\mathbf{f}$.

**Alignment** The LM is a *low-level causal model* $\mathcal{L}$ where variables are dimensions of hidden vectors and the hypothesis about LM structure is a *high-level causal model* $\mathcal{H}$. An *alignment* $\Pi$ assigns each high-level variable $A$ to features of a hidden vector $\mathbf{F}^A_{\mathbf{h}}$, e.g., orthogonal directions in the activation space of $\mathbf{h}$. To evaluate an alignment, we perform intervention experiments to evaluate whether high-level interventions on the variables in $\mathcal{H}$ have the same effect as interventions on the aligned features in $\mathcal{L}$.

**Causal Abstraction** We use interchange interventions to reveal whether the hypothesized causal model $\mathcal{H}$ is an abstraction of an LM $\mathcal{L}$. To simplify, assume both models share an input and output space. The high-level model $\mathcal{H}$ is an abstraction of the low-level model $\mathcal{L}$ under a given alignment when each high-level interchange intervention and the aligned low-level intervention result in the same output. For a high-level intervention on $A$ aligned with low-level features $\mathbf{F}^A_{\mathbf{h}}$ with a counterfactual input $\mathbf{c}$ and original input $\mathbf{b}$, we write

$$\mathsf{GetOutput}(\mathcal{L}_{\mathbf{F}^A_{\mathbf{h}} \leftarrow \mathsf{Get}(\mathcal{L}(\mathbf{c}),\mathbf{F}^A_{\mathbf{h}}))}(\mathbf{o})) = \mathsf{GetOutput}(\mathcal{H}_{A \leftarrow \mathsf{Get}(\mathcal{H}(\mathbf{c}),A))}(\mathbf{o})) \tag{1}$$

If the low-level interchange intervention on the LM produces the same output as the aligned high-level intervention on the algorithm, this is a piece of evidence in favor of the hypothesis. This extends naturally to multi-variable interventions (Geiger et al., 2024).

**Graded Faithfulness Metric** We construct *counterfactual datasets* for each causal variable where an example consists of a base prompt and a counterfactual prompt . The *counterfactual label* is the expected output of the algorithm after the high-level interchange intervention, i.e., the right-side of Equation 1. The interchange intervention accuracy is the proportion of examples for which Equation 1 holds, i.e., the degree to which $\mathcal{H}$ faithfully abstracts $\mathcal{L}$.

**Aligning Features to Causal Variables** In our experiments, we use Singular Vector Decomposition (SVD) to featurize residual stream vectors, i.e., features are the orthogonal singular vectors. For a given transformer layer and token location, we collect the residual stream vectors across a large number of examples and compute the singular vectors. Given singular vector features $\mathbf{F_h}$ of a hidden vector $\mathbf{h}$ in the residual stream of the LM $\mathcal{L}$, we select features to align with a causal variable $A$ in causal model $\mathcal{H}$ using Desiderata-based Component Masking (DCM) (De Cao et al., 2020; Davies et al., 2023; Prakash et al., 2024). Given original input $\mathbf{o}$ and counterfactual input $\mathbf{c}$, we train a mask $\mathbf{m} \in [0,1]^{|\mathbf{F_h}|}$ on the following objective

$$\mathsf{CE}\Big(\mathsf{GetLogits}\big(\mathcal{L}_{\mathbf{F_h} \leftarrow \mathbf{m} \circ \mathsf{Get}(\mathcal{L}(\mathbf{c}), \mathbf{F_h}))}(\mathbf{b})\big), \mathsf{GetLogits}\big(\mathcal{H}_{A \leftarrow \mathsf{Get}(\mathcal{H}(\mathbf{c}), A))}(\mathbf{b})\big)\Big) \tag{2}$$

## E Pseudocode for the Belief Tracking High-Level Causal Model

**Algorithm 2** High-level causal model for the no visibility

```
procedure BeliefTracking($c_1, o_1, s_1, c_2, o_2, s_2, q_c, q_o$)
    Ordering ID assignment
    $c_1^{OI}, o_1^{OI}, s_1^{OI} \leftarrow$ AssignOIs($[c_1, o_1, s_1], 1$)
    $c_2^{OI}, o_2^{OI}, s_2^{OI} \leftarrow$ AssignOIs($[c_2, o_2, s_2], 2$)

    Binding lookback mechanism
    binding_address$_1 \leftarrow$ (copy($c_1^{OI}$), copy($o_1^{OI}$))
    binding_address$_2 \leftarrow$ (copy($c_2^{OI}$), copy($o_2^{OI}$))

    $q_c^{OI} \leftarrow$ copy($\{c_1 : c_1^{OI}, c_2 : c_2^{OI}\}[q_c]$)
    $q_o^{OI} \leftarrow$ copy($\{o_1 : o_1^{OI}, o_2 : o_2^{OI}\}[q_o]$)
    binding_pointer $\leftarrow (q_c^{OI}, q_o^{OI})$

    if binding_address$_1 =$ binding_pointer then
        binding_payload $\leftarrow$ copy($s_1^{OI}$)
    else if binding_address$_2 =$ binding_pointer then
        binding_payload $\leftarrow$ copy($s_2^{OI}$)
    end if

    Answer lookback mechanism
    answer_pointer $\leftarrow$ binding_payload
    answer1_address $\leftarrow s_1^{OI}$
    answer2_address $\leftarrow s_2^{OI}$
    if answer1_address $=$ answer_pointer then
        answer_payload $\leftarrow s_1$
    else if answer2_address $=$ answer_pointer then
        answer_payload $\leftarrow s_2$
    end if
    return answer_payload
end procedure
```

## F Desiderata-based Component Masking

While interchange interventions on residual vectors reveal where a causal variable might be encoded in the LM's internal activations, they do not localize the variable to specific subspaces. To address this, we apply the *Desiderata-based Component Masking* technique (De Cao et al., 2020; Davies et al., 2023; Prakash et al., 2024), which learns a sparse binary mask $\mathbf{m}$ over the singular vectors of the LM's internal activations. We first cache the internal activations from 500 samples at the token positions specified in the main text for each experiment. Next, we apply *Singular Value Decomposition* to compute the singular vectors as a matrix $V \in \mathbb{R}^{d \times 500}$ where $d$ is the dimensionality of the residual stream. We then masked this matrix using a learnable binary vector $\mathbf{m} \in [0, 1]^d$ to choose a subset of singular vectors

$$V_{masked} = V\mathbf{m} \tag{3}$$

The chosen subset of vectors is used to construct a *projection matrix* $W_{proj} \in \mathbb{R}^{d \times d}$.

$$W_{proj} = V_{masked} V_{masked}^T \tag{4}$$

Then, we perform subspace-level interchange interventions (rather than replacing the entire residual vector) using the following equations:

$$h_{new} = W_{proj} h_c + (I - W_{proj}) h_o \tag{5}$$