ABSTRACTION OF INTERMEDIATE REPRESENTATIONS FOR MECHANISTIC INTERPRETABILITY VIA SHAPLEY INDIRECT EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

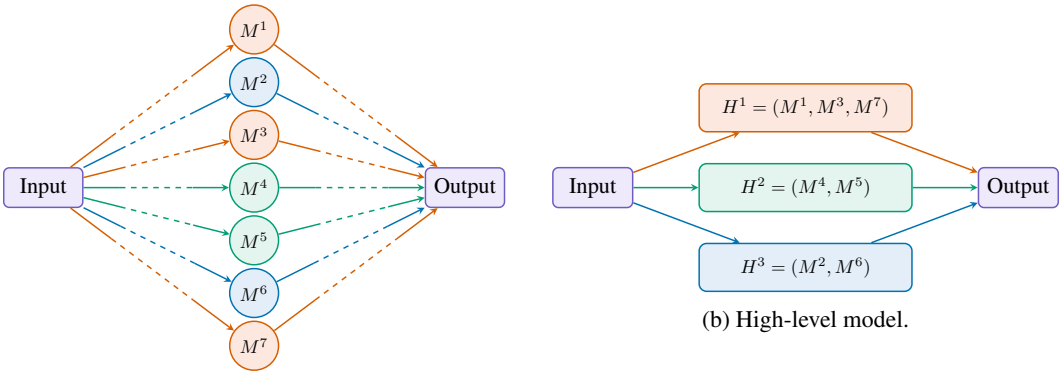
Providing interpretable explanations of neural networks is a central challenge in modern machine learning. We study how semantically meaningful changes in input attributes propagate through intermediate representations to affect model outputs, using causal mediation analysis and interchange interventions. Explaining large systems may require abstracting many low-level components into a smaller number of meaningful mechanisms, motivating the abstraction of intermediate representations. We first introduce a Shapley value-based indirect effect that treats causally non-ordered mediators as players in a cooperative game, averages each mediator's marginal contribution across coalitions, and satisfies the desired additive decomposition property. Using this definition, we formalize abstraction quality in terms of completeness, requiring each input attribute's effect to be concentrated in a dominant high-level mediator, and distinctness, requiring different input attributes to be associated with different dominant mediators. We then propose a greedy algorithm for constructing the abstraction. Finally, experiments on GPT-2 small and Qwen3.5-9B-base show that clear abstractions emerge at early and intermediate layers, with weaker completeness at later layers.

1 INTRODUCTION

Providing interpretable explanations of complex systems, such as neural networks, is a central challenge in modern machine learning. This paper studies *influence analysis*: characterizing how a change in an input propagates through a system to affect its output. Identifying the roles of internal components requires quantifying how much of the resulting output change is transmitted through each component. In causal mediation analysis, Pearl (2001; 2009) formalized the natural indirect effect by contrasting counterfactual outcomes under mediator values induced by different treatment levels, thereby isolating the effect transmitted through a specified mediator. Analogously, interchange interventions in neural network interpretability construct counterfactual outcomes by replacing an internal representation with the value it would attain under a different input (Vig et al., 2020; Chintam et al., 2023; Weng et al., 2024). More broadly, causal mediation analysis has become an important framework for characterizing and identifying mediating components in mechanistic interpretability (Mueller et al., 2026).

Although such interchange interventions can quantify the influence transmitted through individual internal representations, explaining a large system may require organizing its many low-level components into a smaller number of meaningful mechanisms. Causal abstraction formalizes this objective by mapping a complex low-level causal model, $\mathcal{M}^L$, to a simpler and more interpretable high-level model, $\mathcal{M}^H$ (Rubenstein et al., 2017; Beckers & Halpern, 2019; Beckers et al., 2019; Geiger et al., 2025). This framework has been applied to mechanistic interpretability, particularly to investigate whether neural networks and large language models implement interpretable high-level causal structures (Geiger et al., 2021; Wu et al., 2023). By summarizing low-level computations in terms of a smaller set of causally meaningful variables, an abstraction can provide a more comprehensible explanation of an otherwise opaque system. Automatically discovering such abstractions, however, remains challenging. Existing approaches based on interchange interventions have primarily evaluated whether a neural network implements a human-specified high-level causal model (Geiger et al., 2021; Wu et al., 2023), rather than deriving the high-level model directly from the network's patterns of mediated influence. Circuit discovery identifies sparse causal subnetworks or feature-level structures supporting specific model behaviors (Marks et al., 2025).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Low-level model.

(b) High-level model.

Figure 1: Causal abstraction of intermediate representations. Colors indicate the correspondence between low-level mediators and their high-level abstract mediator.

Formally, we consider a system with a potentially large collection of mediating variables,

$$\boldsymbol{M} = (M^1, \ldots, M^I), \tag{1}$$

between the input $\boldsymbol{X}$ and the output Y. We assume that the mediators are causally non-ordered: no mediator M^i is a parent or descendant of another mediator M^j. This setting arises naturally in LLMs when the mediators are selected from the same layer or computational stage. We further assume that each change in a component of the input vector

$$\boldsymbol{X} = (X^1, \ldots, X^K) \tag{2}$$

has a meaningful semantic interpretation. A change in an individual component can therefore be viewed as an intervention on a specific input attribute, allowing us to examine how that intervention propagates through the mediators to affect the output. The semantics of the input attributes provide a basis for interpreting the semantic roles of the abstract mediators. Our objective is to characterize and decompose the effect of changing each input component X^k.

Before turning to model abstraction, we first address a fundamental issue in defining mediator-specific effects. For an interpretable mediator-level attribution, the mediator-specific indirect effects should satisfy a *decomposition property*: their sum should equal the joint indirect effect. With multiple mediators, traditional path-specific effects (Pearl, 2001) do not generally sum to the joint indirect effect. Inspired by the Shapley value (Shapley, 1951), we formulate the effect on the output as the value of a cooperative game in which the mediators are the players and define the *Shapley value-based indirect effect*. Rather than fixing the remaining mediators at a single reference configuration, this measure averages each mediator's marginal contribution over all possible coalitions. By the efficiency property of the Shapley value, the resulting mediator-specific effects sum exactly to the total mediated effect, which coincides with the total effect in our fully mediated setting. The proposed measure, therefore, provides a principled allocation of the total effect among causally non-ordered mediators, including contributions arising from their interactions.

Next, to improve interpretability through abstraction, we aggregate the low-level mediators into a smaller collection of disjoint high-level mediators:

$$\boldsymbol{H} = (H^1, \ldots, H^J) = \mathrm{group}_{\mathcal{P}}(\boldsymbol{M}), \qquad J \ll I, \tag{3}$$

where $\mathcal{P}$ is a partition of the low-level mediator indices. This grouping is an instance of a *constructive causal abstraction* (Beckers & Halpern, 2019). As illustrated in Figure 1, the high-level model summarizes the low-level system using fewer causally meaningful units. For high interpretability, we introduce two criteria that the abstraction should satisfy:

- **Completeness.** For each input component X^k, the effect of changing X^k is concentrated in a single dominant high-level mediator; that mediator's indirect effect is therefore close to the total effect.

- **Distinctness.** Different input components are associated with different dominant high-level mediators, so the abstraction separates their corresponding mechanisms.

Completeness makes each input effect easier to interpret by attributing it to a single dominant high-level mediator, with negligible effects through the others, while distinctness separates the roles associated with different input attributes. Together, these criteria encourage a simple and interpretable mapping from input attributes to high-level mediators. We next use the Shapley value-based indirect effects to quantify how well a given mediator abstraction satisfies these criteria and propose a greedy algorithm for automatically constructing such abstractions.

Finally, using the proposed methodology, we investigate the mechanistic interpretability of LLMs through two applications—grammatical feature processing and gender-related influence—and show that clear abstractions emerge at early and intermediate layers in both GPT-2 small and Qwen3.5-9B-base, while completeness degrades at later layers.

The contributions of this paper are:

- We introduce a Shapley value-based indirect effect for multiple causally non-ordered mediators that provides a decomposition of the total effect.
- We propose a framework for abstracting low-level mediators into a smaller set of high-level mediators, together with the completeness and distinctness criteria for evaluating abstraction quality and a greedy algorithm for constructing such abstractions.
- We apply the framework to mechanistic interpretability in GPT-2 small and Qwen3.5-9B-base, studying grammatical feature processing and gender-related influence, and show that clear abstractions emerge at multiple intervention layers.

2 Motivating Example: Mechanistic Interpretability of Large Language Models

We illustrate a motivating example in LLM interpretability. Consider prompts of the form

"Every day, {I, You, He, She, They} ...."

The output is the model's preference between the continuations "read" and "reads". An English learner may be taught to select the appropriate verb form based on grammatical number (singular vs. plural) and person (third person vs. non-third person). Then, a natural question is whether an LLM develops a similar mechanism and, if so, where in the network that mechanism is represented.

A good abstraction can answer this question by providing a simple and explicit characterization of the roles of the mediators. An ideal abstraction is illustrated in Figure 2, which successfully separates the mediators responsible for grammatical number from those responsible for grammatical person. The indirect effects under this abstraction satisfy the following properties:

- The indirect effect of changing from singular to plural is mediated entirely through mediator set A; the indirect effects through all other mediator sets are zero.
- The indirect effect of changing from third person to non-third person is mediated entirely through mediator set B, which is disjoint from mediator set A; the indirect effects through all other mediator sets are zero.

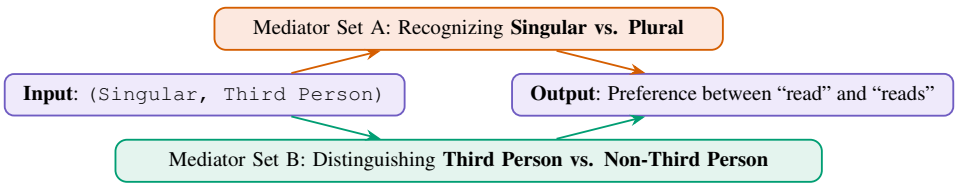


Figure 2: Illustrative example of causal abstraction.

3 PRELIMINARIES

Setup We formulate the problem with a view toward LLM interpretability. Let $\mathcal{I} := \{1, \ldots, I\}$ denote the index set of low-level mediators, where each mediator represents an internal unit of the system, such as a head within a layer. Let $\boldsymbol{T}$ denote an input text, and let Y denote a scalar output quantity of interest. We assume the following low-level deterministic causal model:

$$\mathcal{M}^L : M^i = f_{M^i}(\boldsymbol{T}), \;\; i \in \mathcal{I}, \;\; Y = f_Y(M^1, \ldots, M^I), \tag{4}$$

where the mediators are causally non-ordered and the input affects the output only through the mediators, with no direct effect. Deterministic causal models are a central object of study in causal abstraction (Geiger et al., 2025). For an LLM, this assumption is appropriate when all intermediate variables are defined within the same layer or computational stage. It generally does not hold across layers because activations in later layers causally depend on those in earlier layers.

Characterization Vector We map the original input $\boldsymbol{T}$ to a characterization vector $\boldsymbol{X} = \varphi(\boldsymbol{T})$, where changes in its components have intuitive semantic interpretations. A complete characterization of the input is often impossible. We therefore consider the augmented vector $(\boldsymbol{X}, U)$, where $\boldsymbol{U}$ represents factors not captured by $\boldsymbol{X}$. We assume that $\boldsymbol{T}$ and $(\boldsymbol{X}, \boldsymbol{U})$ are in one-to-one correspondence, with inverse mapping ψ, so that $\boldsymbol{T} = \psi(\boldsymbol{X}, \boldsymbol{U})$. We denote the real-world joint distribution of $(\boldsymbol{X}, \boldsymbol{U})$ by $\mathbb{P}_{(\boldsymbol{X},\boldsymbol{U})}$. For example, for prompts of the form $\boldsymbol{T}$ = "Every day, {I, You, He, She, They}", the characterization vector $\boldsymbol{X}$ may encode semantic attributes such as grammatical number and person, while U captures the remaining lexical and contextual information. The structure of the characterization vector should reflect the semantic attributes that are relevant to the researcher's intended interpretation of the abstract mediators.

Effect of Input We study influence analysis by examining how changes in the input affect the output, for example, changing "He" to "I". Let $\boldsymbol{X}^{-k} := (X^1, \ldots, X^{k-1}, X^{k+1}, \ldots, X^K)$, and define $\boldsymbol{X}_{x^k} := (X^1, \ldots, X^{k-1}, x^k, X^{k+1}, \ldots, X^K)$ and $\boldsymbol{X}_{x'^k} := (X^1, \ldots, X^{k-1}, x'^k, X^{k+1}, \ldots, X^K)$. Let $P_{-k,\boldsymbol{U}}$ denote the reference distribution of $(\boldsymbol{X}^{-k}, \boldsymbol{U})$. For simplicity, for any mediator configuration $\boldsymbol{m} = (m^1, \ldots, m^I)$, we write $Y(\boldsymbol{m}) := f_Y(\boldsymbol{m})$. For each mediator i, let $M^i(\boldsymbol{X}_{x^k}, \boldsymbol{U}) := f_{M^i}(\psi(\boldsymbol{X}_{x^k}, \boldsymbol{U}))$.

Definition 1. *The (average) total effect with respect to component k, from x'^k to x^k, is*

$$\begin{aligned} TE(k; x^k, x'^k) := \mathbb{E}_{(\boldsymbol{X}^{-k},\boldsymbol{U})\sim P_{-k,\boldsymbol{U}}}[&Y(M^1(\boldsymbol{X}_{x^k}, \boldsymbol{U}), \ldots, M^I(\boldsymbol{X}_{x^k}, \boldsymbol{U})) \\ &- Y(M^1(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), \ldots, M^I(\boldsymbol{X}_{x'^k}, \boldsymbol{U}))]. \end{aligned} \tag{5}$$

This is average over all input components other than X^k. Following the traditional path-specific effect framework of Pearl (2001), we define the following mediator-specific indirect effect.

Definition 2. *The (average natural) indirect effect with respect to component k, from x'^k to x^k, through the low-level mediator M^i is defined as*

$$\begin{aligned} &IE^{\mathcal{M}^L}_{\text{single}}(k, i; x^k, x'^k) := \mathbb{E}_{(\boldsymbol{X}^{-k},\boldsymbol{U})\sim P_{-k,\boldsymbol{U}}}[\\ &Y(M^1(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), \ldots, M^{i-1}(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), M^i(\boldsymbol{X}_{x^k}, \boldsymbol{U}), M^{i+1}(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), \ldots, M^I(\boldsymbol{X}_{x'^k}, \boldsymbol{U})) \\ &- Y(M^1(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), \ldots, M^{i-1}(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), M^i(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), M^{i+1}(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), \ldots, M^I(\boldsymbol{X}_{x'^k}, \boldsymbol{U}))]. \end{aligned} \tag{6}$$

This expression changes only mediator M^i to its value under x^k, while keeping all other mediators at their natural values under x'^k. The remaining input components, together with $\boldsymbol{U}$, follow the reference distribution $P_{-k,\boldsymbol{U}}$ and are not intervened on. The indirect effect is therefore averaged over all input components other than X^k and over $\boldsymbol{U}$. This measure provides a natural notion of the indirect effect transmitted through an individual mediator. However, this measure does not generally provide an additive decomposition of the total effect, i.e., in general $\sum_{i\in\mathcal{I}} IE^{\mathcal{M}^L}_{\text{single}}(k, i; x^k, x'^k) \neq TE(k; x^k, x'^k)$. Consequently, this definition is not suitable for studying the completeness property, which requires quantifying the proportion of the total effect attributed to each mediator.

4 SHAPLEY VALUE-BASED INDIRECT EFFECTS

4.1 SHAPLEY VALUE-BASED INDIRECT EFFECTS UNDER LOW-LEVEL MODEL

We first define a Shapley value-based indirect effect for allocating the total mediated effect among multiple causally non-ordered mediators. Originating in cooperative game theory (Shapley, 1951), the Shapley value allocates an outcome among individual contributors by averaging each contributor's marginal contribution across all possible coalitions. Inspired by cooperative game theory, we view the effect on the output transmitted through multiple mediators as the joint contribution of a coalition defined by the mediators' intervention states.

Our Shapley value-based indirect effects differ from causal Shapley values (Heskes et al., 2020) and do-Shapley values (Jung et al., 2022) in their attribution targets and coalition construction. Causal Shapley values attribute an individual model prediction to input features while accounting for their causal relationships. Do-Shapley values quantify each variable's contribution to a joint intervention effect by averaging its marginal effect over coalitions of intervened variables, leaving the remaining variables unintervened. In contrast, our measure allocates the effect of changing a single input attribute across multiple causally non-ordered mediators by averaging each mediator's marginal contribution over coalitions of treatment- and reference-induced mediator states.

Definition of the Shapley value-based Indirect Effect For each $S \subseteq \mathcal{I}$ and $i \in \mathcal{I}$, we define

$$M_i^{(S)}(\boldsymbol{X}_{x^k}, \boldsymbol{X}_{x'^k}, \boldsymbol{U}) \coloneqq \begin{cases} M^i(\boldsymbol{X}_{x^k}, \boldsymbol{U}), & i \in S \\ M^i(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), & i \notin S \end{cases}, \tag{7}$$

and $\boldsymbol{M}^{(S)}(\boldsymbol{X}_{x^k}, \boldsymbol{X}_{x'^k}, \boldsymbol{U}) := (M_i^{(S)}(\boldsymbol{X}_{x^k}, \boldsymbol{X}_{x'^k}, \boldsymbol{U}))_{i\in\mathcal{I}}$. Then, the coalition value associated with $S \subseteq \mathcal{I}$ is defined by

$$v_k^L(S; x^k, x'^k) \coloneqq \mathbb{E}_{(\boldsymbol{X}^{-k}, \boldsymbol{U})\sim P_{-k,\boldsymbol{U}}}[Y(\boldsymbol{M}^{(S)}(\boldsymbol{X}_{x^k}, \boldsymbol{X}_{x'^k}, \boldsymbol{U}))]. \tag{8}$$

Thus, the mediators indexed by S take their values under x^k, whereas the remaining mediators take their values under x'^k. For $i \notin S$, the difference

$$v_k^L(S \cup \{i\}; x^k, x'^k) - v_k^L(S; x^k, x'^k) \tag{9}$$

represents the marginal effect of changing the mediator M^i from its value under x'^k to its value under x^k, with the mediators indexed by S already taking their values under x^k.

Definition 3 (Shapley value-based indirect effect)**.** *The Shapley value-based indirect effect of changing component k from x'^k to x^k attributed to the low-level mediator M^i is*

$$IE_{\mathrm{SHAP}}^{\mathcal{M}^L}(k, i; x^k, x'^k) \coloneqq \sum\nolimits_{S\subseteq\mathcal{I}\setminus\{i\}} \frac{|S|!\,(I-|S|-1)!}{I!}[v_k^L(S\cup\{i\}; x^k, x'^k) - v_k^L(S; x^k, x'^k)]. \tag{10}$$

By the efficiency property of the Shapley value (Shapley, 1951), these Shapley value-based indirect effects provide an exact additive decomposition of the total effect.

Proposition 1 (Decomposition Property)**.** *The following decomposition property holds:*

$$\sum\nolimits_{i\in\mathcal{I}} IE_{\mathrm{SHAP}}^{\mathcal{M}^L}(k, i; x^k, x'^k) = v_k^L(\mathcal{I}; x^k, x'^k) - v_k^L(\varnothing; x^k, x'^k) = TE(k; x^k, x'^k). \tag{11}$$

Beyond this decomposition property, the remaining Shapley axioms (Shapley, 1951) also admit natural interpretations in the mediation setting. First, suppose two mediators M^i and M^j have the same marginal contributions for every coalition $S \subseteq \mathcal{I} \setminus \{i, j\}$: $v_k^L(S \cup \{i\}; x^k, x'^k) - v_k^L(S; x^k, x'^k) = v_k^L(S \cup \{j\}; x^k, x'^k) - v_k^L(S; x^k, x'^k)$. Then $\mathrm{IE}_{\mathrm{SHAP}}^{M^L}(k, i; x^k, x'^k) = \mathrm{IE}_{\mathrm{SHAP}}^{M^L}(k, j; x^k, x'^k)$. Thus, mediators with the same marginal contributions across all relevant coalitions receive equal indirect-effect attributions, regardless of their labels or ordering. Second, if switching mediator M^i from its value under x'^k to its value under x^k never changes the coalition value, $v_k^L(S \cup \{i\}; x^k, x'^k) = v_k^L(S; x^k, x'^k)$ for every $S \subseteq \mathcal{I} \setminus \{i\}$, then $\mathrm{IE}_{\mathrm{SHAP}}^{M^L}(k, i; x^k, x'^k) = 0$. Thus, a mediator receives zero attribution if changing its state never changes the output, regardless of the states of the other mediators.

4.2 SHAPLEY VALUE-BASED INDIRECT EFFECTS UNDER HIGH-LEVEL MODEL

Abstraction Next, we study the abstraction of mediators. We begin with the low-level mediators $M = (M^i)_{i\in\mathcal{I}}$. Let $\mathcal{J} \coloneqq \{1, \dots, J\}$. To construct a high-level model, let $\mathcal{P} = \{\mathcal{I}_1, \dots, \mathcal{I}_J\}$ be a partition of $\mathcal{I}$; that is,

$$\mathcal{I}_j \neq \varnothing, \quad \mathcal{I}_j \cap \mathcal{I}_{j'} = \varnothing \ (j \neq j'), \quad \bigcup\nolimits_{j\in\mathcal{J}} \mathcal{I}_j = \mathcal{I}. \tag{12}$$

The j-th abstract mediator is defined as the tuple of low-level mediators indexed by $\mathcal{I}_j$:

$$H^j(\boldsymbol{X}, \boldsymbol{U}) \coloneqq (M^i(\boldsymbol{X}, \boldsymbol{U}))_{i\in\mathcal{I}_j}. \tag{13}$$

Thus, H^j groups the corresponding low-level mediators without discarding information. Because $\mathcal{P}$ is a partition of $\mathcal{I}$, the complete abstract-mediator vector $H = (H^1, \dots, H^J)$ contains the same information as M, up to regrouping. This grouping defines *a constructive causal abstraction* (Beckers & Halpern, 2019): each high-level mediator corresponds to a disjoint subset of low-level mediators, and an interchange intervention on a high-level mediator corresponds to jointly replacing its constituent low-level mediators by their values under the corresponding input. Let $\mathrm{ungroup}_{\mathcal{P}}$ denote the operation that reconstructs M from H. The high-level outcome function is induced by the low-level outcome function as $Y^H(H^1, \dots, H^J) \coloneqq Y(\mathrm{ungroup}_{\mathcal{P}}(H^1, \dots, H^J))$.

Shapley value-based Indirect Effects Under High-Level Model Similarly to the low-level model, we define Shapley value-based indirect effects under the high-level model. For each $A \subseteq \mathcal{J}$, define the hybrid abstract-mediator vector $H^{(A)}(\boldsymbol{X}_{x^k}, \boldsymbol{X}_{x'^k}, \boldsymbol{U})$ componentwise as

$$H_r^{(A)}(\boldsymbol{X}_{x^k}, \boldsymbol{X}_{x'^k}, \boldsymbol{U}) \coloneqq \begin{cases} H^r(\boldsymbol{X}_{x^k}, \boldsymbol{U}), & r \in A \\ H^r(\boldsymbol{X}_{x'^k}, \boldsymbol{U}), & r \notin A \end{cases}, \tag{14}$$

and $\boldsymbol{H}^{(A)}(\boldsymbol{X}_{x^k}, \boldsymbol{X}_{x'^k}, \boldsymbol{U}) \coloneqq (H_r^{(A)}(\boldsymbol{X}_{x^k}, \boldsymbol{X}_{x'^k}, \boldsymbol{U}))_{r\in\mathcal{J}}$. The high-level coalition value associated with A is defined as

$$v_k^H(A; x^k, x'^k) \coloneqq \mathbb{E}_{(\boldsymbol{X}^{-k}, \boldsymbol{U})\sim P_{-k,U}}[Y^H(\boldsymbol{H}^{(A)}(\boldsymbol{X}_{x^k}, \boldsymbol{X}_{x'^k}, \boldsymbol{U}))]. \tag{15}$$

Definition 4 (High-level Shapley value-based indirect effect)**.** *The Shapley value-based indirect effect attributed to the abstract mediator H^j is defined as*

$$IE_{\mathrm{SHAP}}^{\mathcal{M}^H}(k, j; x^k, x'^k) \coloneqq \sum\nolimits_{A\subseteq\mathcal{J}\setminus\{j\}} \frac{|A|!\,(J - |A| - 1)!}{J!} [v_k^H(A \cup \{j\}; x^k, x'^k) - v_k^H(A; x^k, x'^k)]. \tag{16}$$

By the grouping construction, $v_k^H(A; x^k, x'^k) = v_k^L(\bigcup_{j\in A} \mathcal{I}_j; x^k, x'^k)$ for every $A \subseteq \mathcal{J}$. Consequently, $v_k^H(\mathcal{J}; x^k, x'^k) - v_k^H(\varnothing; x^k, x'^k) = TE(k; x^k, x'^k)$. Therefore, by the efficiency property of the Shapley value, the decomposition of the total effect is preserved under the high-level model $TE(k; x^k, x'^k) = \sum_{j\in\mathcal{J}} IE_{\mathrm{SHAP}}^{\mathcal{M}^H}(k, j; x^k, x'^k)$.

Limitation of Shapley value-based Indirect Effects Shapley value-based indirect effects preserve the additive decomposition of the total effect at both the low-level mediator model and the high-level abstract mediator model, but they do not generally preserve mediator-specific attributions across these two levels. In particular, in general, $IE_{\mathrm{SHAP}}^{\mathcal{M}^H}(k, j; x^k, x'^k) \neq \sum_{i\in\mathcal{I}_j} IE_{\mathrm{SHAP}}^{\mathcal{M}^L}(k, i; x^k, x'^k)$, where $\mathcal{I}_j$ denotes the set of low-level mediators grouped into the high-level mediator H_j.

5 EVALUATING AND CONSTRUCTING HIGH-LEVEL ABSTRACTIONS

An Ideal Abstraction of Intermediate Representations Using the total and Shapley value-based indirect effects under $\mathcal{M}^H$, we formulate an ideal causal abstraction for explaining the underlying complex model in terms of indirect effects that satisfy completeness and distinctness.

For each input component k, define a dominant abstract mediator by

$$j_k^\star \in \arg\max_{j\in\mathcal{J}} |IE_{\mathrm{SHAP}}^{\mathcal{M}^H}(k, j; x^k, x'^k)|. \tag{17}$$

For each k, the dominant abstract mediator should explain the total effect of changing X^k. Ideally,

$$\textbf{Completeness}: \quad IE^{\mathcal{M}_H}_{\mathrm{SHAP}}(k, j; x^k, x'^k) = 0, \quad \text{for all } j \neq j^\star_k,\ k \in \{1, \ldots, K\}. \tag{18}$$

Additionally, distinct input components should be explained by distinct dominant abstract mediators. Ideally,

$$\textbf{Distinctness}: \quad j^\star_k \neq j^\star_{k'}, \quad \text{for all } k \neq k'. \tag{19}$$

The complete distinctness condition is feasible only when $J \geq K$.

Measuring the Quality of an Abstracted High-Level Model The ideal conditions cannot generally be achieved in practice. We therefore measure the quality of an abstracted model by quantifying its deviation from these ideal conditions. For each input component k, we measure completeness by

$$L(k; \mathcal{P}) \coloneqq \frac{\sum_{j \neq j^\star_k} |IE^{\mathcal{M}^H}_{\mathrm{SHAP}}(k, j; x^k, x'^k)|}{|TE(k; x^k, x'^k)| + \varepsilon}, \tag{20}$$

where $\varepsilon > 0$ is a small constant that prevents division by zero. A smaller value of $L(k; \mathcal{P})$ indicates that the total effect of X^k is more completely concentrated on its dominant abstract mediator.

We summarize the overall non-dominant leakage by averaging over all input components:

$$L(\mathcal{P}) \coloneqq K^{-1} \sum\nolimits_{k=1}^{K} L(k; \mathcal{P}). \tag{21}$$

Since each $L(k; \mathcal{P}) \geq 0$, and $L(k; \mathcal{P}) = 0$ if and only if Eq. (18) holds for input component k, we have $L(\mathcal{P}) = 0$ if and only if Eq. (18) holds for every $k \in \{1, \ldots, K\}$. We next separately measure distinctness by

$$D(\mathcal{P}) \coloneqq 1 - \frac{2}{K(K-1)} \sum\nolimits_{1 \leq k < k' \leq K} \mathbb{I}(j^\star_k = j^\star_{k'}), \tag{22}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. $D(\mathcal{P}) = 1$ if and only if all input components have distinct dominant abstract mediators.

Algorithm for Representation Abstraction We seek an abstraction that balances completeness and distinctness by solving the following optimization problem:

$$\min_{\mathcal{P}} \quad \Phi_\lambda(\mathcal{P}) \coloneqq L(\mathcal{P}) + \lambda[\delta - D(\mathcal{P})]_+, \tag{23}$$

where $[a]_+ \coloneqq \max\{a, 0\}$, $\delta \in [0, 1]$ specifies the desired degree of distinctness, and $\lambda > 0$ controls the penalty for violation. Minimizing $L(\mathcal{P})$ promotes completeness, whereas the penalty term promotes distinct dominant abstract mediators. Larger values of λ place greater emphasis on satisfying $D(\mathcal{P}) \geq \delta$. Solving this optimization problem is computationally expensive, as the number of possible partitions grows combinatorially with the number of mediators.

We then develop a heuristic greedy algorithm for constructing the abstraction. Starting from the singleton partition of the low-level mediators, the algorithm repeatedly merges two groups until J abstract mediators remain. At each iteration, it evaluates every candidate pairwise merge using permutation-based estimates of the high-level Shapley value-based effects and selects the candidate minimizing the penalized objective $\widehat{\Phi}_\lambda$. Because this procedure is greedy and uses a penalty rather than a hard constraint, it does not guarantee that the returned partition satisfies $D(\widehat{\mathcal{P}}) \geq \delta$. Algorithm 1 presents the pseudocode for the greedy abstraction procedure.

Algorithm 1: Greedy abstraction

Input: Mediators $\mathcal{I}$, target groups J, distinctness target δ, penalty λ, permutation samples N
Output: Partition $\widehat{\mathcal{P}}$
$\mathcal{P} \leftarrow \{\{i\} : i \in \mathcal{I}\}$
while $|\mathcal{P}| > J$ **do**
  $\widehat{\Phi}^\star_\lambda \leftarrow \infty$
  for each pair $\mathcal{I}_a, \mathcal{I}_b \in \mathcal{P}$ **do**
    $\widetilde{\mathcal{P}} \leftarrow \mathrm{Merge}(\mathcal{P}, \mathcal{I}_a, \mathcal{I}_b)$
    Estimate $\widehat{IE}^{\mathcal{M}^H}_{\mathrm{SHAP}}(k, j)$ for all k, j under $\widetilde{\mathcal{P}}$ using N permutations
    Compute $\widehat{\Phi}_\lambda \leftarrow \widehat{L}(\widetilde{\mathcal{P}}) + \lambda[\delta - \widehat{D}(\widetilde{\mathcal{P}})]_+$
    if $\widehat{\Phi}_\lambda < \widehat{\Phi}^\star_\lambda$ **then**
      $(\widehat{\Phi}^\star_\lambda, \mathcal{P}^\star) \leftarrow (\widehat{\Phi}_\lambda, \widetilde{\mathcal{P}})$
    end if
  end for
  $\mathcal{P} \leftarrow \mathcal{P}^\star$
end while
return $\widehat{\mathcal{P}} \leftarrow \mathcal{P}$

Table 1: (Third-Person Singular -s) Selected three-group abstractions and estimated effects across intervention layers. Shapley value-based IEs are (IE via H_1, IE via H_2, IE via H_3).

Layer	Selected abstraction	Feature	TE	IEs	Dominant
0	$H_1 = \{0, 2, 5, 6, 7, 13, 14, 19, 21, 24, 26, 27, 28, 30, 31\}$, $H_2 = \{1, 3, 4, 8, 9, 11, 12, 15, 16, 17, 18, 20, 22, 23, 25, 29\}$, $H_3 = \{10\}$	Number	0.358	$(0.017,\ \mathbf{0.309},\ 0.032)$	H_2
		Person	0.202	$(\mathbf{0.192},\ 0.007,\ 0.003)$	H_1
5	$H_1 = \{0, 1, 2, 4, 13, 15, 18, 21, 24, 25, 26, 28, 29, 30\}$, $H_2 = \{3, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 17, 19, 22, 23, 27, 31\}$, $H_3 = \{20\}$	Number	0.358	$(\mathbf{0.353},\ -0.000,\ 0.005)$	H_1
		Person	0.202	$(0.019,\ \mathbf{0.182},\ 0.001)$	H_2
10	$H_1 = \{0, 2, 4, 5, 7, 11, 13, 15, 17, 20, 21, 22, 24, 26, 27, 28, 30\}$, $H_2 = \{1, 6, 19, 25\}$, $H_3 = \{3, 8, 9, 10, 12, 14, 16, 18, 23, 29, 31\}$	Number	0.358	$(\mathbf{0.248},\ 0.019,\ 0.091)$	H_1
		Person	0.202	$(0.092,\ 0.006,\ \mathbf{0.104})$	H_3

6 Mechanistic Interpretability of LLMs

6.1 Third-Person Singular -*s*

Setup Using the developed methodology, we investigate the problem introduced in Section 2. We evaluate the abstraction procedure on GPT-2 small (`gpt2`), which has 12 layers and hidden dimension $d = 768$. We consider interventions at the inputs to layers 0, 5, and 10. Due to computational constraints, we partition the 768-dimensional representation in advance into $|\mathcal{I}| = 32$ contiguous blocks of width 24, which serve as the low-level mediators. The outcome is the conditional probability of the singular verb form,

$$P(\texttt{``reads''} \mid \boldsymbol{T}) / \{P(\texttt{``reads''} \mid \boldsymbol{T}) + P(\texttt{``read''} \mid \boldsymbol{T})\}. \tag{24}$$

The experimental details are provided in Appendix A.1. The code is publicly available at `https://anonymous.4open.science/r/ICLR2027_Abstraction_Shapley_Indirect_Effects-4C54/`. All experiments were conducted using an NVIDIA A40 (48GB) GPU and an Intel Xeon Gold 6342 CPU.

Results We report the resulting abstractions across layers for $J = 3$ high-level mediators, as shown in Table 1. As a baseline, we additionally report the Shapley indirect effects without abstraction in Tables 4–6 in Appendix A.1. The results without abstraction are less interpretable because the mediated effects are distributed across many low-level components. For $J = 3$, Table 1 shows that layer 5 yields the clearest abstraction. At layer 5, the number effect is almost entirely concentrated in H_1 (0.353 of a total effect of 0.358), while the person effect is predominantly concentrated in H_2 (0.182 of 0.202). Thus, the two input features are associated with distinct dominant high-level mediators, with little residual effect assigned to the remaining groups. At layer 0, the number and person effects are also strongly separated: the number effect is primarily attributed to H_2 (0.309 of 0.358), whereas the person effect is primarily attributed to H_1 (0.192 of 0.202). At layer 10, the dominant mediators remain distinct, with the number effect primarily attributed to H_1 (0.248 of 0.358) and the person effect primarily attributed to H_3 (0.104 of 0.202). However, both effects are less concentrated than at layer 5, particularly for person, with a substantial H_1 contribution (0.092).

Additional Experiments Table 3 in Appendix A.1 reports additional experiments varying the number of low-level mediators, $|\mathcal{I}|$, across intervention layers 0, 5, and 10. These results indicate that the learned abstraction is reasonably robust to the choice of mediator granularity, although performance is weaker at coarser granularities, particularly for $|\mathcal{I}| = 4$ and 8, while the clearest separation between grammatical number and person is obtained at layers 0 and 5, with weaker completeness at layer 10. To assess robustness to random sampling, we repeat the abstraction procedure using different random seeds and report the results in Table 7 in Appendix A.1. At $|\mathcal{I}| = 32$, we observe a similar qualitative pattern: layers 0 and 5 yield low non-dominant leakage with distinct dominant mediators, whereas layer 10 exhibits weaker completeness. We also include experiments across different values of J in Figures 3–5 in Appendix A.1. The results show how completeness and dominant-group size evolve as the number of high-level mediators J decreases during the greedy abstraction procedure. Across the experiments, as J gradually decreases, both the IE captured by the dominant group and the dominant-group size generally increase. Moreover, the IE captured by

Table 2: (Gender Bias) Selected three-group abstractions and estimated effects. Shapley value-based IEs are (IE via H_1, IE via H_2, IE via H_3). (Occup. = occupation; Adj. = adjective)

Layer	Selected abstraction	Feature	TE	IEs	Dominant
0	$H_1 = \{0, 4, 9, 10, 12, 13, 14, 17, 21, 25, 26, 31\}$, $H_2 = \{1, 2, 3, 5, 7, 8, 11, 15, 16, 18, 19, 20, 22, 23, 24, 27, 28, 29, 30\}$, $H_3 = \{6\}$	Occup.	0.659	$(0.078,\ \mathbf{0.582},\ -0.001)$	H_2
		Adj.	0.207	$(\mathbf{0.113},\ 0.094,\ 0.001)$	H_1
5	$H_1 = \{0, 2, 3, 4, 5, 6, 9, 14, 15, 16, 20, 22, 23, 26, 27, 28, 31\}$, $H_2 = \{1\}$, $H_3 = \{7, 8, 10, 11, 12, 13, 17, 18, 19, 21, 24, 25, 29, 30\}$	Occup.	0.659	$(\mathbf{0.503},\ 0.005,\ 0.151)$	H_1
		Adj.	0.207	$(0.052,\ 0.001,\ \mathbf{0.155})$	H_3
10	$H_1 = \{0, 1, 7, 12, 18, 19, 20, 21, 24, 26, 31\}$, $H_2 = \{2, 3, 4, 5, 6, 8, 9, 10, 11, 14, 15, 16, 22, 23, 25, 27, 28, 29, 30\}$, $H_3 = \{13, 17\}$	Occup.	0.659	$(0.166,\ \mathbf{0.493},\ 0.000)$	H_2
		Adj.	0.207	$(\mathbf{0.105},\ 0.100,\ 0.002)$	H_1

the dominant group tends to increase with dominant-group size, suggesting that aggregating more low-level mediators into the dominant high-level mediator yields a more complete abstraction. We additionally evaluate Qwen3.5-9B-base and observe clearer separation between grammatical number and person at layers 0, 10, and 20 than at layer 30, as detailed in Appendix B.1.

6.2 Gender Bias

Setup In addition, we apply our abstraction framework to the study of gender bias. We consider prompts that combine an occupation with a gender-associated adjective, ``The [adjective] [occupation] said that ...'', where both the adjective and occupation may carry male- or female-associated cues. We measure the model's relative preference for the continuations ``he'' and ``she'' using $P(\text{`` he''} \mid \boldsymbol{T})/\{P(\text{`` he''} \mid \boldsymbol{T}) + P(\text{`` she''} \mid \boldsymbol{T})\}$. One may expect gender-related influence to arise from interactions between occupation and adjective effects. We aim to provide an interpretable abstraction that identifies which components primarily mediate occupation-related effects and which primarily mediate adjective-related effects. The experimental details are provided in Appendix A.2.

Results We report the resulting abstractions across layers for $J = 3$ high-level mediators, as shown in Table 2. For $J = 3$, Table 2 shows the clearest abstraction at layer 5, where the occupation effect is concentrated in H_1 $(0.503/0.659)$ and the adjective effect in H_3 $(0.155/0.207)$. At layers 0 and 10, the dominant mediators are also distinct, but the adjective effect is less concentrated, particularly at layer 10, where it is nearly split between H_1 and H_2 $(0.105$ and $0.100)$. We provide additional results, analogous to those for the first experiment, in Appendices A.2 and B.2.

7 Conclusion

We introduced an abstraction framework that groups mediators according to how they causally contribute to changes in model outputs. Our findings provide a foundation for systematically discovering interpretable high-level mechanisms from low-level intermediate representations. Our abstraction objective differs from those of Zennaro et al. (2023); Zhu et al. (2024); Massidda et al. (2024); D'Acunto et al. (2025); Saulus et al. (2026). Prior work has focused (i) on selecting abstractions that preserve variation in low-level variables at the higher-level model or (ii) on recovering latent high-level causal structure from low-level observations. In contrast, our objective is to group mediators according to how they causally contribute to changes in the model output.

This paper focuses on indirect effects through individual mediators. In causal inference, path-specific effects are sometimes used to characterize effects transmitted jointly through multiple mediators along specified causal pathways (Pearl, 2001; Avin et al., 2005). Extending our abstraction framework to path-specific effects is an important direction for future work, although doing so may incur substantial computational costs. We also restrict attention to deterministic models evaluated on a fixed input $\boldsymbol{T}$, which is sufficient for the mechanistic interpretation setting considered here, including applications to LLMs. Extending the framework to scientific settings involving probabilistic models would require counterfactual reasoning. Developing such an extension is another important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing, language refinement, organization, formatting, and code development. AI-assisted code was reviewed and tested by the authors before being used in the experiments. Generative AI tools were also used to assist with mathematical exposition and the presentation of experimental results. We did not use generative AI tools to autonomously generate experimental findings or determine the scientific conclusions of the work. All AI-assisted text, code, and mathematical content were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pp. 357–363, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.

Sander Beckers and Joseph Y. Halpern. Abstracting causal models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstraction. *Uncertain Artif Intell*, 2019, Jul 2019.

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar Van Der Wal. Identifying and adapting transformer-components responsible for gender bias in an english language model. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 379–394, 2023.

Gabriele D'Acunto, Fabio Massimo Zennaro, Yorgos Felekis, and Paolo Di Lorenzo. Causal abstraction learning based on the semantic embedding principle. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL `http://jmlr.org/papers/v26/23-0058.html`.

Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Bloebaum, and Elias Bareinboim. On measuring causal contributions via do-interventions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10476–10501. PMLR, 17–23 Jul 2022.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=I4e82CIDxv`.

Riccardo Massidda, Sara Magliacane, and Davide Bacciu. Learning causal abstractions of linear structural causal models. *arXiv preprint arXiv:2406.00394*, 2024.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis. *Computational Linguistics*, 52(1), March 2026.

Judea Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pp. 411–420, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558608001.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, August 2017. URL http://auai.org/uai2017/proceedings/papers/11.pdf.

Théo Saulus, Simon Lacoste-Julien, and Dhanya Sridhar. Unsupervised causal abstractions discovery. *arXiv preprint arXiv:2606.19594*, 2026.

Lloyd S Shapley. Notes on the n-person game—ii: The value of an n-person game. 1951.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15669–15680, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. *Advances in neural information processing systems*, 36:78205–78226, 2023.

Fabio Massimo Zennaro, Paolo Turrini, and Theodoros Damoulas. Quantifying consistency and information loss for causal abstraction learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023. ISBN 978-1-956792-03-4.

Yuchen Zhu, Sergio Hernan Garrido Mejia, Bernhard Schölkopf, and Michel Besserve. Unsupervised causal abstraction. In *NeurIPS 2024 Causal Representation Learning Workshop*, 2024. URL https://openreview.net/forum?id=BP0e8RvFwd.

APPENDIX TO "ABSTRACTION OF INTERMEDIATE REPRESENTATIONS FOR MECHANISTIC INTERPRETABILITY VIA SHAPLEY INDIRECT EFFECTS"

A ADDITIONAL EXPERIMENTAL INFORMATION FOR GPT-2 SMALL

In this appendix, we provide additional experimental results that supplement the main body of the paper.

A.1 THIRD-PERSON SINGULAR *-s*

Detailed Experimental Setup Using the developed methodology, we investigate the problem introduced in the motivating example. We evaluate the abstraction procedure on GPT-2 small (`gpt2`), which has 12 layers and hidden dimension $d = 768$. We consider interventions at the inputs to layers 0, 5, and 10. Due to computational constraints, we partition the 768-dimensional representation in advance into $|\mathcal{I}| = 32$ contiguous blocks of width 24, which serve as the low-level mediators. We consider two input components, $k \in \{\text{NUMBER}, \text{PERSON}\}$. For NUMBER, we use the contrasts he $\leftarrow$ they and I $\leftarrow$ we; for PERSON, we use he $\leftarrow$ I and they $\leftarrow$ we. Each contrast is combined with four templates: "Every day, {pronoun}", "At the library, {pronoun}", "After lunch, {pronoun}", and "When there is time, {pronoun}". This gives eight possible reference pairs per component. We draw $B = 8$ pairs with replacement using random seed 0, and we reuse the same pairs for all coalition evaluations. For a coalition $S \subseteq \mathcal{I}$, we run the model on the control prompt and, at every token position, replace the coordinates corresponding to mediators in S with their treatment values. The outcome is the conditional probability of the singular verb form,

$$\frac{P(\texttt{``reads''} \mid \boldsymbol{T})}{P(\texttt{``reads''} \mid \boldsymbol{T}) + P(\texttt{``read''} \mid \boldsymbol{T})}. \tag{25}$$

Coalition values are estimated by averaging this outcome over the B sampled reference pairs. For a candidate partition $\mathcal{P} = \{\mathcal{I}_1, \ldots, \mathcal{I}_J\}$, we estimate the high-level Shapley indirect effects using $N = 64$ permutations. We set $\delta = 1$ and $\lambda = 10$.

Additional Results Table 3 reports three-group abstractions for different numbers of low-level mediators, $|\mathcal{I}| \in \{4, 8, 12, 16, 24, 32\}$, at intervention layers 0, 5, and 10. We did not consider $|\mathcal{I}| > 32$ due to computational resource constraints. Abstraction quality varies with both intervention layer and mediator granularity.

At layer 5, finer mediator partitions generally yield clearer separation between grammatical number and person. In particular, for $|\mathcal{I}| = 32$, the number effect is concentrated in H_1 (0.353 of a total effect of 0.358), while the person effect is predominantly concentrated in H_2 (0.182 of 0.202). Layer 0 also exhibits strong separation at finer granularities. In contrast, at layer 10, the effects generally remain more distributed across high-level mediators, indicating weaker completeness even when the dominant mediators are distinct.

Overall, clear abstractions are obtained at layers 0 and 5 with sufficiently fine mediator partitions. However, finer granularity does not uniformly improve abstraction quality, and coarse partitions can yield either dispersed effects or a shared dominant mediator for both attributes.

Tables 4–6 report the low-level Shapley indirect effects without abstraction. These effects are distributed across many low-level components, making the roles associated with grammatical number and person less apparent than in the high-level abstractions.

Different Seed To assess robustness to random sampling, we repeat the abstraction procedure using a different random seed, as reported in Table 7. At $|\mathcal{I}| = 32$, we observe a similar qualitative pattern: layers 0 and 5 yield low non-dominant leakage with distinct dominant mediators, whereas layer 10 exhibits weaker completeness. However, the selected partitions and the results at other granularities vary across runs.

Table 3: (Third-Person Singular -*s*) Three-group abstractions across intervention layers and different numbers of low-level mediators. L denotes non-dominant leakage and D denotes distinctness.

Layer	$\lvert\mathcal{I}\rvert$	Selected abstraction	Number IEs	Dom.	Person IEs	Dom.	L	D
0	4	$H_1 = \{0, 2\}, H_2 = \{1\}, H_3 = \{3\}$	$(0.112, 0.121, \mathbf{0.125})$	H_3	$(\mathbf{0.081}, 0.059, 0.062)$	H_1	0.625	1
0	8	$H_1 = \{0, 2, 6\}, H_2 = \{1, 4, 5, 7\}, H_3 = \{3\}$	$(0.116, 0.113, \mathbf{0.128})$	H_3	$(0.030, \mathbf{0.144}, 0.028)$	H_2	0.464	1
0	12	$H_1 = \{0, 4, 6, 8\}, H_2 = \{1, 3, 10\}, H_3 = \{2, 5, 7, 9, 11\}$	$(0.122, \mathbf{0.125}, 0.111)$	H_2	$(0.062, 0.064, \mathbf{0.076})$	H_3	0.637	1
0	16	$H_1 = \{0, 2, 3, 4, 5, 7, 9, 14\}, H_2 = \{1, 6, 8, 11, 12, 13, 15\}, H_3 = \{10\}$	$(\mathbf{0.348}, -0.003, 0.013)$	H_1	$(0.074, \mathbf{0.096}, 0.031)$	H_2	0.284	1
0	24	$H_1 = \{0, 2, 3, 6, 7, 8, 14, 15, 18, 21, 23\}, H_2 = \{1, 4, 5, 9, 10, 11, 12, 13, 16, 19, 20, 22\}, H_3 = \{17\}$	$(\mathbf{0.358}, -0.008, 0.008)$	H_1	$(0.007, \mathbf{0.193}, 0.001)$	H_2	0.042	1
0	32	$H_1 = \{0, 2, 5, 6, 7, 13, 14, 19, 21, 24, 26, 27, 28, 30, 31\}, H_2 = \{1, 3, 4, 8, 9, 11, 12, 15, 16, 17, 18, 20, 22, 23, 25, 29\}, H_3 = \{10\}$	$(0.017, \mathbf{0.309}, 0.032)$	H_2	$(\mathbf{0.192}, 0.007, 0.003)$	H_1	0.093	1
5	4	$H_1 = \{0\}, H_2 = \{1\}, H_3 = \{2, 3\}$	$(0.015, 0.028, \mathbf{0.315})$	H_3	$(0.021, 0.031, \mathbf{0.150})$	H_3	0.190	0
5	8	$H_1 = \{0, 6, 7\}, H_2 = \{1\}, H_3 = \{2, 3, 4, 5\}$	$(\mathbf{0.145}, 0.094, 0.118)$	H_1	$(0.029, 0.018, \mathbf{0.154})$	H_3	0.414	1
5	12	$H_1 = \{0, 1, 7, 9, 11\}, H_2 = \{2, 3, 4, 5, 6, 8\}, H_3 = \{10\}$	$(\mathbf{0.244}, 0.063, 0.050)$	H_1	$(0.030, \mathbf{0.166}, 0.005)$	H_2	0.248	1
5	16	$H_1 = \{0, 1, 5, 6, 9, 10, 12\}, H_2 = \{2, 3, 4, 7, 8, 11, 13, 15\}, H_3 = \{14\}$	$(\mathbf{0.296}, 0.029, 0.032)$	H_1	$(0.039, \mathbf{0.161}, 0.002)$	H_2	0.186	1
5	24	$H_1 = \{0, 1, 4, 11, 14, 15, 16, 18, 19, 21, 22\}, H_2 = \{2\}, H_3 = \{3, 5, 6, 7, 8, 9, 10, 12, 13, 17, 20, 23\}$	$(\mathbf{0.354}, 0.002, 0.001)$	H_1	$(0.031, 0.003, \mathbf{0.168})$	H_3	0.088	1
5	32	$H_1 = \{0, 1, 2, 4, 13, 15, 18, 21, 24, 25, 26, 28, 29, 30\}, H_2 = \{3, 5, 6, 7, 8, 9, 10, 11, 12, 14, 16, 17, 19, 22, 23, 27, 31\}, H_3 = \{20\}$	$(\mathbf{0.353}, -0.000, 0.005)$	H_1	$(0.019, \mathbf{0.182}, 0.001)$	H_2	0.057	1
10	4	$H_1 = \{0\}, H_2 = \{1, 2\}, H_3 = \{3\}$	$(-0.002, \mathbf{0.321}, 0.039)$	H_2	$(0.029, \mathbf{0.145}, 0.028)$	H_2	0.198	0
10	8	$H_1 = \{0, 6, 7\}, H_2 = \{1, 3, 5\}, H_3 = \{2, 4\}$	$(0.118, 0.114, \mathbf{0.126})$	H_3	$(0.067, \mathbf{0.076}, 0.058)$	H_2	0.635	1
10	12	$H_1 = \{0\}, H_2 = \{1, 2, 3, 4, 9, 10, 11\}, H_3 = \{5, 6, 7, 8\}$	$(0.024, 0.166, \mathbf{0.168})$	H_3	$(0.008, \mathbf{0.118}, 0.075)$	H_2	0.473	1
10	16	$H_1 = \{0, 1, 2, 6, 7, 8, 10, 15\}, H_2 = \{3, 4, 5, 9, 11, 12, 14\}, H_3 = \{13\}$	$(\mathbf{0.249}, 0.108, 0.000)$	H_1	$(0.095, \mathbf{0.103}, 0.003)$	H_2	0.395	1
10	24	$H_1 = \{0, 1, 4, 6, 8, 10, 11, 12, 13, 15, 16, 21\}, H_2 = \{2, 3, 5, 7, 9, 14, 17, 18, 19, 20, 22\}, H_3 = \{23\}$	$(\mathbf{0.265}, 0.090, 0.003)$	H_1	$(0.074, \mathbf{0.124}, 0.004)$	H_2	0.323	1
10	32	$H_1 = \{0, 2, 4, 5, 7, 11, 13, 15, 17, 20, 21, 22, 24, 26, 27, 28, 30\}, H_2 = \{1, 6, 19, 25\}, H_3 = \{3, 8, 9, 10, 12, 14, 16, 18, 23, 29, 31\}$	$(\mathbf{0.248}, 0.019, 0.091)$	H_1	$(0.092, 0.006, \mathbf{0.104})$	H_3	0.396	1

Table 4: Low-level Shapley indirect effects without abstraction at the input to layer 0. The 768-dimensional residual representation is partitioned into $|\mathcal{I}|$ contiguous blocks. For each granularity, the estimated indirect effects sum to the corresponding total effect ($\mathrm{TE}_{\mathrm{Number}} = 0.358$ and $\mathrm{TE}_{\mathrm{Person}} = 0.202$).

$\lvert\mathcal{I}\rvert$	Block width	Number IEs	Person IEs
32	24	$M_0 : 0.005,\ M_1 : 0.006,\ M_2 : 0.004,\ M_3 : 0.016,$ $M_4 : 0.006,\ M_5 : 0.013,\ M_6 : 0.010,\ M_7 : 0.010,$ $M_8 : 0.016,\ M_9 : 0.010,\ M_{10} : 0.011,\ M_{11} : 0.013,$ $M_{12} : 0.016,\ M_{13} : 0.008,\ M_{14} : 0.017,\ M_{15} : 0.008,$ $M_{16} : 0.015,\ M_{17} : 0.025,\ M_{18} : 0.018,\ M_{19} : 0.008,$ $M_{20} : 0.006,\ M_{21} : 0.010,\ M_{22} : 0.001,\ M_{23} : 0.003,$ $M_{24} : 0.023,\ M_{25} : 0.010,\ M_{26} : 0.022,\ M_{27} : 0.003,$ $M_{28} : 0.017,\ M_{29} : 0.006,\ M_{30} : 0.019,\ M_{31} : 0.004$	$M_0 : 0.006,\ M_1 : 0.001,\ M_2 : 0.007,\ M_3 : 0.003,$ $M_4 : 0.004,\ M_5 : 0.004,\ M_6 : 0.002,\ M_7 : 0.010,$ $M_8 : 0.003,\ M_9 : 0.006,\ M_{10} : 0.006,\ M_{11} : 0.001,$ $M_{12} : 0.011,\ M_{13} : 0.004,\ M_{14} : 0.005,\ M_{15} : 0.003,$ $M_{16} : 0.011,\ M_{17} : 0.004,\ M_{18} : 0.005,\ M_{19} : 0.009,$ $M_{20} : 0.003,\ M_{21} : 0.005,\ M_{22} : 0.006,\ M_{23} : 0.003,$ $M_{24} : 0.020,\ M_{25} : 0.004,\ M_{26} : 0.018,\ M_{27} : 0.007,$ $M_{28} : 0.011,\ M_{29} : 0.006,\ M_{30} : 0.009,\ M_{31} : 0.003$
24	32	$M_0 : 0.008,\ M_1 : 0.005,\ M_2 : 0.022,\ M_3 : 0.010,$ $M_4 : 0.012,\ M_5 : 0.020,\ M_6 : 0.022,\ M_7 : 0.035,$ $M_8 : 0.012,\ M_9 : 0.009,\ M_{10} : 0.028,\ M_{11} : 0.015,$ $M_{12} : 0.022,\ M_{13} : 0.013,\ M_{14} : 0.010,\ M_{15} : 0.002,$ $M_{16} : 0.010,\ M_{17} : 0.012,\ M_{18} : 0.015,\ M_{19} : 0.010,$ $M_{20} : 0.004,\ M_{21} : 0.037,\ M_{22} : 0.010,\ M_{23} : 0.013$	$M_0 : 0.003,\ M_1 : 0.013,\ M_2 : 0.002,\ M_3 : 0.010,$ $M_4 : 0.005,\ M_5 : 0.015,\ M_6 : 0.006,\ M_7 : 0.013,$ $M_8 : 0.005,\ M_9 : 0.011,\ M_{10} : 0.006,\ M_{11} : 0.015,$ $M_{12} : 0.008,\ M_{13} : 0.008,\ M_{14} : 0.006,\ M_{15} : 0.004,$ $M_{16} : 0.007,\ M_{17} : 0.007,\ M_{18} : 0.011,\ M_{19} : 0.012,$ $M_{20} : 0.004,\ M_{21} : 0.013,\ M_{22} : 0.007,\ M_{23} : 0.010$
16	48	$M_0 : 0.017,\ M_1 : 0.028,\ M_2 : 0.015,\ M_3 : 0.025,$ $M_4 : 0.045,\ M_5 : 0.021,\ M_6 : 0.008,\ M_7 : 0.017,$ $M_8 : 0.012,\ M_9 : 0.032,\ M_{10} : 0.021,\ M_{11} : 0.003,$ $M_{12} : 0.039,\ M_{13} : 0.024,\ M_{14} : 0.042,\ M_{15} : 0.011$	$M_0 : 0.011,\ M_1 : 0.012,\ M_2 : 0.005,\ M_3 : 0.015,$ $M_4 : 0.007,\ M_5 : 0.010,\ M_6 : 0.013,\ M_7 : 0.011,$ $M_8 : 0.009,\ M_9 : 0.026,\ M_{10} : 0.005,\ M_{11} : 0.006,$ $M_{12} : 0.015,\ M_{13} : 0.020,\ M_{14} : 0.021,\ M_{15} : 0.018$
12	64	$M_0 : 0.021,\ M_1 : 0.016,\ M_2 : 0.023,\ M_3 : 0.079,$ $M_4 : 0.030,\ M_5 : 0.036,\ M_6 : 0.040,\ M_7 : 0.008,$ $M_8 : 0.011,\ M_9 : 0.060,\ M_{10} : 0.019,\ M_{11} : 0.014$	$M_0 : 0.017,\ M_1 : 0.008,\ M_2 : 0.024,\ M_3 : 0.015,$ $M_4 : 0.011,\ M_5 : 0.024,\ M_6 : 0.015,\ M_7 : 0.009,$ $M_8 : 0.018,\ M_9 : 0.029,\ M_{10} : 0.016,\ M_{11} : 0.015$
8	96	$M_0 : 0.039,\ M_1 : 0.023,\ M_2 : 0.088,\ M_3 : 0.024,$ $M_4 : 0.057,\ M_5 : 0.029,\ M_6 : 0.077,\ M_7 : 0.021$	$M_0 : 0.017,\ M_1 : 0.022,\ M_2 : 0.021,\ M_3 : 0.017,$ $M_4 : 0.034,\ M_5 : 0.017,\ M_6 : 0.047,\ M_7 : 0.027$
4	192	$M_0 : 0.043,\ M_1 : 0.146,$ $M_2 : 0.074,\ M_3 : 0.095$	$M_0 : 0.032,\ M_1 : 0.036,$ $M_2 : 0.042,\ M_3 : 0.092$

Table 5: Low-level Shapley indirect effects without abstraction at the input to layer 5. The 768-dimensional residual representation is partitioned into $|\mathcal{I}|$ contiguous blocks. For each granularity, the estimated indirect effects sum to the corresponding total effect ($\mathrm{TE}_{\mathrm{Number}} = 0.358$ and $\mathrm{TE}_{\mathrm{Person}} = 0.202$).

$\lvert\mathcal{I}\rvert$	Block width	Number IEs	Person IEs
32	24	$M_0 : 0.019,\ M_1 : 0.017,\ M_2 : 0.015,\ M_3 : 0.002,$ $M_4 : 0.009,\ M_5 : 0.008,\ M_6 : 0.011,\ M_7 : 0.012,$ $M_8 : 0.007,\ M_9 : -0.002,\ M_{10} : 0.011,\ M_{11} : 0.001,$ $M_{12} : 0.014,\ M_{13} : 0.013,\ M_{14} : 0.010,\ M_{15} : 0.005,$ $M_{16} : 0.012,\ M_{17} : 0.002,\ M_{18} : 0.009,\ M_{19} : 0.010,$ $M_{20} : 0.008,\ M_{21} : 0.018,\ M_{22} : 0.011,\ M_{23} : 0.010,$ $M_{24} : 0.039,\ M_{25} : 0.010,\ M_{26} : 0.010,\ M_{27} : 0.007,$ $M_{28} : 0.009,\ M_{29} : 0.017,\ M_{30} : 0.031,\ M_{31} : 0.006$	$M_0 : 0.005,\ M_1 : 0.004,\ M_2 : 0.004,\ M_3 : 0.004,$ $M_4 : 0.006,\ M_5 : 0.006,\ M_6 : 0.002,\ M_7 : 0.012,$ $M_8 : 0.010,\ M_9 : 0.009,\ M_{10} : 0.007,\ M_{11} : 0.003,$ $M_{12} : 0.013,\ M_{13} : 0.005,\ M_{14} : 0.007,\ M_{15} : 0.000,$ $M_{16} : 0.011,\ M_{17} : 0.007,\ M_{18} : 0.005,\ M_{19} : 0.002,$ $M_{20} : 0.002,\ M_{21} : 0.010,\ M_{22} : 0.003,\ M_{23} : 0.009,$ $M_{24} : 0.019,\ M_{25} : 0.003,\ M_{26} : 0.002,\ M_{27} : 0.012,$ $M_{28} : -0.001,\ M_{29} : 0.005,\ M_{30} : 0.007,\ M_{31} : 0.006$
24	32	$M_0 : 0.009,\ M_1 : 0.028,\ M_2 : 0.003,\ M_3 : 0.014,$ $M_4 : 0.014,\ M_5 : 0.028,\ M_6 : 0.004,\ M_7 : 0.008,$ $M_8 : 0.009,\ M_9 : 0.011,\ M_{10} : 0.027,\ M_{11} : 0.005,$ $M_{12} : 0.010,\ M_{13} : 0.003,\ M_{14} : 0.020,\ M_{15} : 0.015,$ $M_{16} : 0.030,\ M_{17} : 0.020,\ M_{18} : 0.028,\ M_{19} : 0.011,$ $M_{20} : 0.006,\ M_{21} : 0.021,\ M_{22} : 0.016,\ M_{23} : 0.019$	$M_0 : 0.000,\ M_1 : 0.006,\ M_2 : 0.003,\ M_3 : 0.007,$ $M_4 : 0.005,\ M_5 : 0.019,\ M_6 : 0.013,\ M_7 : 0.010,$ $M_8 : 0.005,\ M_9 : 0.014,\ M_{10} : 0.010,\ M_{11} : 0.003,$ $M_{12} : 0.009,\ M_{13} : 0.006,\ M_{14} : 0.006,\ M_{15} : 0.002,$ $M_{16} : 0.013,\ M_{17} : 0.020,\ M_{18} : 0.018,\ M_{19} : 0.006,$ $M_{20} : 0.008,\ M_{21} : -0.001,\ M_{22} : 0.007,\ M_{23} : 0.012$
16	48	$M_0 : 0.047,\ M_1 : 0.018,\ M_2 : 0.009,\ M_3 : 0.026,$ $M_4 : 0.003,\ M_5 : 0.017,\ M_6 : 0.020,\ M_7 : 0.013,$ $M_8 : 0.009,\ M_9 : 0.029,\ M_{10} : 0.028,\ M_{11} : 0.017,$ $M_{12} : 0.042,\ M_{13} : 0.016,\ M_{14} : 0.031,\ M_{15} : 0.033$	$M_0 : 0.006,\ M_1 : 0.009,\ M_2 : 0.009,\ M_3 : 0.014,$ $M_4 : 0.021,\ M_5 : 0.012,\ M_6 : 0.013,\ M_7 : 0.009,$ $M_8 : 0.019,\ M_9 : 0.009,\ M_{10} : 0.013,\ M_{11} : 0.011,$ $M_{12} : 0.021,\ M_{13} : 0.015,\ M_{14} : 0.006,\ M_{15} : 0.015$
12	64	$M_0 : 0.040,\ M_1 : 0.007,\ M_2 : 0.046,\ M_3 : 0.013,$ $M_4 : 0.016,\ M_5 : 0.034,\ M_6 : 0.012,\ M_7 : 0.023,$ $M_8 : 0.055,\ M_9 : 0.049,\ M_{10} : 0.025,\ M_{11} : 0.036$	$M_0 : 0.012,\ M_1 : 0.006,\ M_2 : 0.026,\ M_3 : 0.024,$ $M_4 : 0.021,\ M_5 : 0.012,\ M_6 : 0.015,\ M_7 : 0.008,$ $M_8 : 0.028,\ M_9 : 0.021,\ M_{10} : 0.010,\ M_{11} : 0.018$
8	96	$M_0 : 0.072,\ M_1 : 0.046,\ M_2 : 0.026,\ M_3 : 0.021,$ $M_4 : 0.038,\ M_5 : 0.053,\ M_6 : 0.058,\ M_7 : 0.043$	$M_0 : 0.022,\ M_1 : 0.024,\ M_2 : 0.035,\ M_3 : 0.023,$ $M_4 : 0.024,\ M_5 : 0.034,\ M_6 : 0.027,\ M_7 : 0.013$
4	192	$M_0 : 0.109,\ M_1 : 0.030,$ $M_2 : 0.096,\ M_3 : 0.123$	$M_0 : 0.037,\ M_1 : 0.056,$ $M_2 : 0.057,\ M_3 : 0.051$

Table 6: Low-level Shapley indirect effects without abstraction at the input to layer 10. The 768-dimensional residual representation is partitioned into $|\mathcal{I}|$ contiguous blocks. For each granularity, the estimated indirect effects sum to the corresponding total effect ($\mathrm{TE}_{\mathrm{Number}} = 0.358$ and $\mathrm{TE}_{\mathrm{Person}} = 0.202$).

$\lvert\mathcal{I}\rvert$	Block width	Number IEs	Person IEs
32	24	$M_0 : -0.005$, $M_1 : 0.012$, $M_2 : 0.026$, $M_3 : 0.004$, $M_4 : 0.008$, $M_5 : -0.000$, $M_6 : -0.000$, $M_7 : 0.006$, $M_8 : 0.025$, $M_9 : 0.007$, $M_{10} : 0.010$, $M_{11} : 0.005$, $M_{12} : 0.017$, $M_{13} : 0.004$, $M_{14} : 0.010$, $M_{15} : 0.009$, $M_{16} : 0.022$, $M_{17} : 0.030$, $M_{18} : 0.011$, $M_{19} : 0.011$, $M_{20} : 0.005$, $M_{21} : 0.020$, $M_{22} : 0.025$, $M_{23} : 0.009$, $M_{24} : 0.017$, $M_{25} : 0.005$, $M_{26} : 0.004$, $M_{27} : 0.003$, $M_{28} : 0.009$, $M_{29} : 0.032$, $M_{30} : 0.012$, $M_{31} : 0.005$	$M_0 : -0.001$, $M_1 : 0.009$, $M_2 : 0.008$, $M_3 : 0.006$, $M_4 : 0.009$, $M_5 : -0.002$, $M_6 : -0.002$, $M_7 : 0.014$, $M_8 : 0.013$, $M_9 : 0.005$, $M_{10} : 0.012$, $M_{11} : -0.001$, $M_{12} : 0.018$, $M_{13} : 0.004$, $M_{14} : 0.001$, $M_{15} : -0.001$, $M_{16} : 0.011$, $M_{17} : 0.011$, $M_{18} : 0.009$, $M_{19} : 0.001$, $M_{20} : 0.000$, $M_{21} : 0.010$, $M_{22} : 0.010$, $M_{23} : 0.014$, $M_{24} : 0.009$, $M_{25} : 0.001$, $M_{26} : 0.001$, $M_{27} : 0.005$, $M_{28} : 0.004$, $M_{29} : 0.018$, $M_{30} : 0.001$, $M_{31} : 0.004$
24	32	$M_0 : -0.013$, $M_1 : 0.036$, $M_2 : 0.005$, $M_3 : 0.008$, $M_4 : -0.001$, $M_5 : 0.008$, $M_6 : 0.023$, $M_7 : 0.010$, $M_8 : 0.015$, $M_9 : 0.015$, $M_{10} : 0.009$, $M_{11} : 0.018$, $M_{12} : 0.035$, $M_{13} : 0.012$, $M_{14} : 0.024$, $M_{15} : 0.008$, $M_{16} : 0.030$, $M_{17} : 0.040$, $M_{18} : 0.017$, $M_{19} : 0.013$, $M_{20} : -0.005$, $M_{21} : 0.016$, $M_{22} : 0.027$, $M_{23} : 0.009$	$M_0 : 0.001$, $M_1 : 0.010$, $M_2 : 0.009$, $M_3 : 0.008$, $M_4 : -0.008$, $M_5 : 0.018$, $M_6 : 0.011$, $M_7 : 0.015$, $M_8 : -0.000$, $M_9 : 0.021$, $M_{10} : 0.005$, $M_{11} : -0.003$, $M_{12} : 0.015$, $M_{13} : 0.005$, $M_{14} : 0.010$, $M_{15} : 0.003$, $M_{16} : 0.013$, $M_{17} : 0.030$, $M_{18} : 0.011$, $M_{19} : 0.002$, $M_{20} : 0.002$, $M_{21} : 0.005$, $M_{22} : 0.016$, $M_{23} : 0.004$
16	48	$M_0 : 0.003$, $M_1 : 0.031$, $M_2 : 0.005$, $M_3 : 0.005$, $M_4 : 0.035$, $M_5 : 0.015$, $M_6 : 0.017$, $M_7 : 0.016$, $M_8 : 0.062$, $M_9 : 0.026$, $M_{10} : 0.020$, $M_{11} : 0.033$, $M_{12} : 0.020$, $M_{13} : 0.005$, $M_{14} : 0.047$, $M_{15} : 0.018$	$M_0 : 0.006$, $M_1 : 0.015$, $M_2 : 0.008$, $M_3 : 0.011$, $M_4 : 0.020$, $M_5 : 0.009$, $M_6 : 0.021$, $M_7 : -0.001$, $M_8 : 0.026$, $M_9 : 0.011$, $M_{10} : 0.008$, $M_{11} : 0.027$, $M_{12} : 0.009$, $M_{13} : 0.006$, $M_{14} : 0.024$, $M_{15} : 0.005$
12	64	$M_0 : 0.030$, $M_1 : 0.007$, $M_2 : 0.008$, $M_3 : 0.035$, $M_4 : 0.032$, $M_5 : 0.024$, $M_6 : 0.058$, $M_7 : 0.025$, $M_8 : 0.058$, $M_9 : 0.025$, $M_{10} : 0.011$, $M_{11} : 0.043$	$M_0 : 0.013$, $M_1 : 0.016$, $M_2 : 0.013$, $M_3 : 0.027$, $M_4 : 0.022$, $M_5 : 0.002$, $M_6 : 0.026$, $M_7 : 0.010$, $M_8 : 0.033$, $M_9 : 0.010$, $M_{10} : 0.008$, $M_{11} : 0.024$
8	96	$M_0 : 0.050$, $M_1 : 0.013$, $M_2 : 0.051$, $M_3 : 0.034$, $M_4 : 0.078$, $M_5 : 0.073$, $M_6 : 0.017$, $M_7 : 0.041$	$M_0 : 0.028$, $M_1 : 0.019$, $M_2 : 0.029$, $M_3 : 0.021$, $M_4 : 0.030$, $M_5 : 0.043$, $M_6 : 0.012$, $M_7 : 0.022$
4	192	$M_0 : 0.040$, $M_1 : 0.082$, $M_2 : 0.161$, $M_3 : 0.075$	$M_0 : 0.037$, $M_1 : 0.049$, $M_2 : 0.076$, $M_3 : 0.040$

Table 7: (Third-Person Singular -*s*; different-seed run) Three-group abstractions across intervention layers and different numbers of low-level mediators. L denotes non-dominant leakage and D denotes distinctness.

Layer	$\lvert\mathcal{I}\rvert$	Selected abstraction	Number IEs	Dom.	Person IEs	Dom.	L	D
0	4	$H_1 = \{0, 1\}, H_2 = \{2\}, H_3 = \{3\}$	$(\mathbf{0.226}, 0.043, 0.043)$	H_1	$(0.133, 0.134, \mathbf{0.135})$	H_3	0.470	1
0	8	$H_1 = \{0, 1, 2, 3\}, H_2 = \{4\}, H_3 = \{5, 6, 7\}$	$(\mathbf{0.227}, 0.045, 0.040)$	H_1	$(0.133, \mathbf{0.136}, 0.133)$	H_2	0.467	1
0	12	$H_1 = \{0, 4, 5, 6, 8, 10\}, H_2 = \{1, 2, 11\}, H_3 = \{3, 7, 9\}$	$(0.097, \mathbf{0.109}, 0.106)$	H_2	$(\mathbf{0.272}, 0.067, 0.064)$	H_1	0.488	1
0	16	$H_1 = \{0, 2, 6, 9, 11, 12, 13, 15\}, H_2 = \{1\}, H_3 = \{3, 4, 5, 7, 8, 10, 14\}$	$(0.012, 0.024, \mathbf{0.276})$	H_3	$(\mathbf{0.394}, 0.006, 0.002)$	H_1	0.068	1
0	24	$H_1 = \{0, 1, 5, 7, 9, 11, 18, 20, 22\}, H_2 = \{2, 4, 6, 8, 10, 12, 13, 16, 19, 21\}, H_3 = \{3, 14, 15, 17, 23\}$	$(0.018, \mathbf{0.271}, 0.023)$	H_2	$(\mathbf{0.136}, 0.133, 0.134)$	H_1	0.396	1
0	32	$H_1 = \{0, 1, 4, 7, 12, 13, 15, 18, 19, 21, 22, 24, 26, 27, 29, 30\}, H_2 = \{2, 3, 5, 6, 8, 9, 10, 11, 14, 16, 17, 20, 23, 25, 28\}, H_3 = \{31\}$	$(-0.010, \mathbf{0.321}, 0.000)$	H_2	$(\mathbf{0.407}, -0.006, 0.001)$	H_1	0.024	1
5	4	$H_1 = \{0\}, H_2 = \{1\}, H_3 = \{2, 3\}$	$(0.018, 0.024, \mathbf{0.269})$	H_3	$(0.042, 0.048, \mathbf{0.312})$	H_3	0.180	0
5	8	$H_1 = \{0, 6, 7\}, H_2 = \{1\}, H_3 = \{2, 3, 4, 5\}$	$(\mathbf{0.130}, 0.081, 0.101)$	H_1	$(0.025, 0.025, \mathbf{0.353})$	H_3	0.353	1
5	12	$H_1 = \{0, 1, 2, 7, 10, 11\}, H_2 = \{3, 4, 6, 8, 9\}, H_3 = \{5\}$	$(\mathbf{0.279}, 0.015, 0.017)$	H_1	$(0.119, \mathbf{0.229}, 0.055)$	H_2	0.267	1
5	16	$H_1 = \{0, 1, 3, 9, 11, 12, 14\}, H_2 = \{2, 4, 6, 7, 8, 10, 13, 15\}, H_3 = \{5\}$	$(\mathbf{0.274}, 0.023, 0.015)$	H_1	$(0.057, \mathbf{0.328}, 0.018)$	H_2	0.153	1
5	24	$H_1 = \{0, 1, 2, 3, 10, 14, 15, 16, 18, 19, 21\}, H_2 = \{4, 5, 6, 7, 9, 11, 12, 13, 17, 20, 22, 23\}, H_3 = \{8\}$	$(\mathbf{0.304}, 0.006, 0.002)$	H_1	$(0.019, \mathbf{0.377}, 0.006)$	H_2	0.044	1
5	32	$H_1 = \{0, 1, 2, 3, 4, 7, 10, 13, 19, 20, 22, 24, 25, 28, 30\}, H_2 = \{5, 8, 9, 11, 12, 14, 15, 16, 17, 18, 21, 23, 26, 27, 29, 31\}, H_3 = \{6\}$	$(\mathbf{0.312}, -0.001, 0.001)$	H_1	$(0.015, \mathbf{0.387}, 0.000)$	H_2	0.023	1
10	4	$H_1 = \{0\}, H_2 = \{1, 2\}, H_3 = \{3\}$	$(0.002, \mathbf{0.280}, 0.029)$	H_2	$(0.039, \mathbf{0.332}, 0.032)$	H_2	0.137	0
10	8	$H_1 = \{0, 1, 2, 3, 4, 5\}, H_2 = \{6\}, H_3 = \{7\}$	$(\mathbf{0.311}, -0.006, 0.007)$	H_1	$(\mathbf{0.411}, -0.017, 0.008)$	H_1	0.053	0
10	12	$H_1 = \{0, 2, 4, 5, 6, 10\}, H_2 = \{1, 3, 8, 9, 11\}, H_3 = \{7\}$	$(\mathbf{0.158}, 0.127, 0.027)$	H_1	$(0.147, \mathbf{0.236}, 0.020)$	H_2	0.454	1
10	16	$H_1 = \{0, 11, 13\}, H_2 = \{1, 2\}, H_3 = \{3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15\}$	$(-0.008, -0.008, \mathbf{0.327})$	H_3	$(0.006, 0.006, \mathbf{0.390})$	H_3	0.040	0
10	24	$H_1 = \{0, 2, 3, 5, 6, 7, 9, 17, 18, 19, 20\}, H_2 = \{1, 4, 8, 11, 12, 13, 14, 15, 16, 21, 22, 23\}, H_3 = \{10\}$	$(0.042, \mathbf{0.266}, 0.003)$	H_2	$(\mathbf{0.205}, 0.191, 0.006)$	H_1	0.318	1
10	32	$H_1 = \{0, 31\}, H_2 = \{1, 3, 4, 7, 8, 9, 11, 12, 14, 17, 18, 23, 26, 28\}, H_3 = \{2, 5, 6, 10, 13, 15, 16, 19, 20, 21, 22, 24, 25, 27, 29, 30\}$	$(0.001, 0.077, \mathbf{0.234})$	H_3	$(0.004, \mathbf{0.200}, 0.198)$	H_2	0.376	1

Experiments Across Different Numbers of High-Level Mediators J Figures 3–5 show how completeness and dominant-group size evolve as the number of high-level mediators J decreases during the greedy abstraction procedure. The remaining experimental settings are the same as those used for the $|\mathcal{I}| = 32$ experiment. Across the experiments, as J gradually decreases, both the IE captured by the dominant group and dominant-group size generally increase. Moreover, as the dominant-group size increases, IE captured by the dominant group also tends to increase, indicating that concentrating more low-level mediators into the dominant high-level mediator yields a more complete abstraction. Furthermore, we observe a mild trade-off between Number and Person: as the abstraction increasingly concentrates the effect of one attribute into its dominant group, the other does not always improve simultaneously. This behavior is consistent with the intended design, in which the two attributes are encouraged to concentrate in distinct high-level mediators.

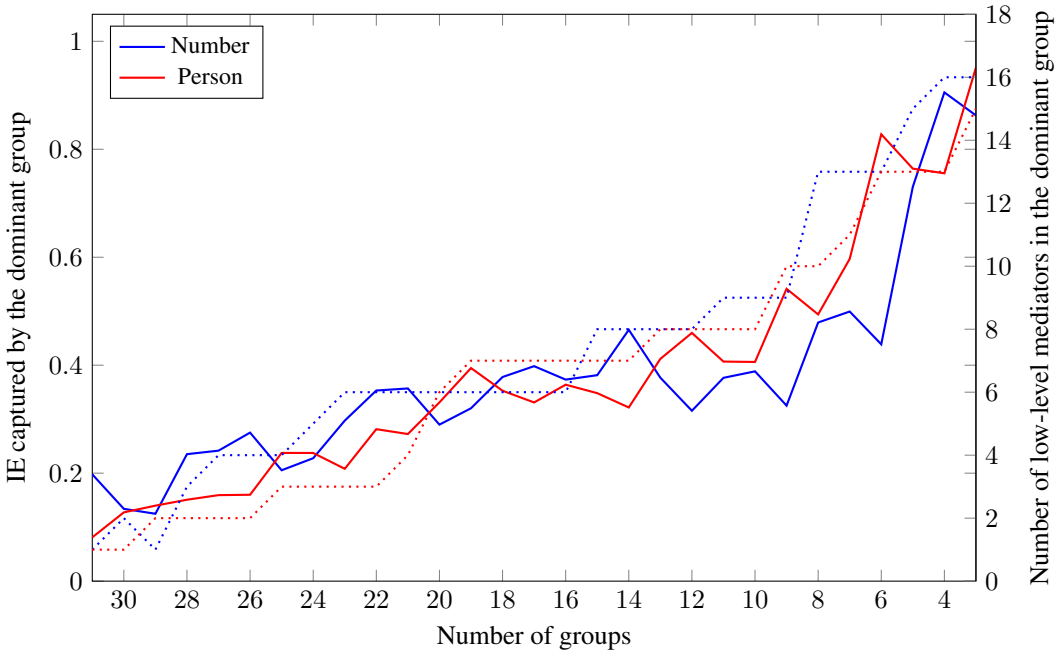


Figure 3: (Third-Person Singular -s) Evolution of IE captured by the dominant group and dominant-group size during the greedy abstraction procedure at layer 0. Solid lines show IE captured by the dominant group, while dotted lines show the number of low-level mediators in the dominant high-level mediator.

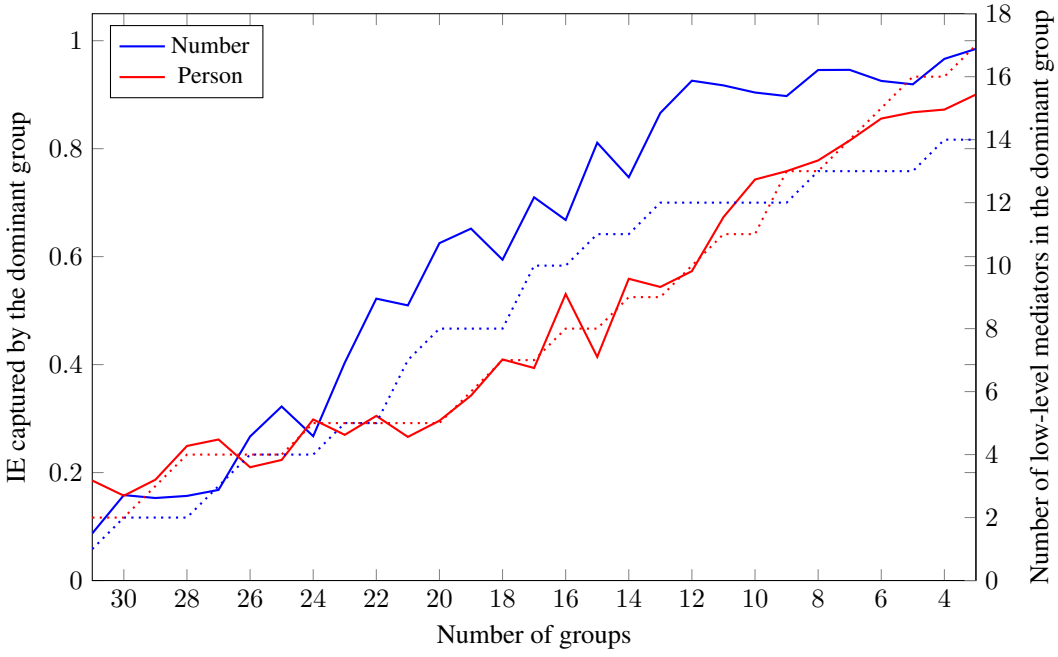


Figure 4: (Third-Person Singular -s) Evolution of IE captured by the dominant group and dominant-group size during the greedy abstraction procedure at layer 5. Solid lines show IE captured by the dominant group, while dotted lines show the number of low-level mediators in the dominant high-level mediator.

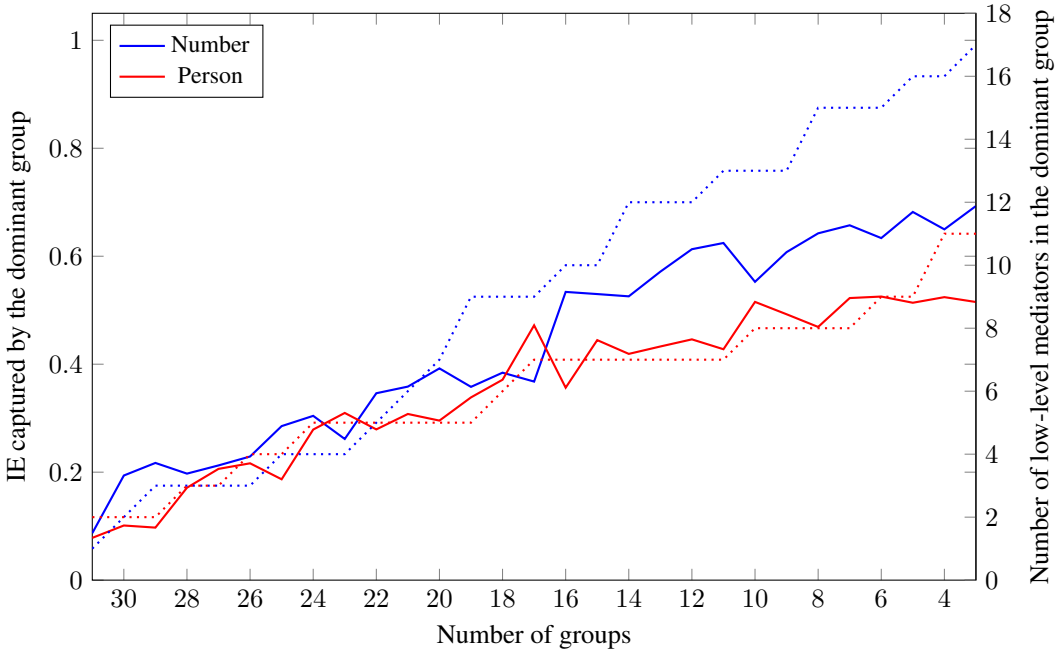


Figure 5: (Third-Person Singular -s) Evolution of IE captured by the dominant group and dominant-group size during the greedy abstraction procedure at layer 10. Solid lines show IE captured by the dominant group, while dotted lines show the number of low-level mediators in the dominant high-level mediator.

A.2 GENDER BIAS

Detailed Experimental Setup We consider prompts that combine an occupation with a gender-associated adjective, for example,

``The [adjective] [occupation] said that ...''

where both the adjective and occupation may carry male- or female-associated cues. We measure the model's relative preference for the continuations ``he'' and ``she'' using

$$\frac{P(\text{`` he''} \mid \boldsymbol{T})}{P(\text{`` he''} \mid \boldsymbol{T}) + P(\text{`` she''} \mid \boldsymbol{T})}. \tag{26}$$

We characterize each prompt using two input components: occupation and adjective. For the occupation component, we use the male-/female-associated pairs $(\texttt{engineer}, \texttt{nurse})$, $(\texttt{mechanic}, \texttt{teacher})$, $(\texttt{carpenter}, \texttt{librarian})$, $(\texttt{surgeon}, \texttt{secretary})$. For the adjective component, we use $(\texttt{bold}, \texttt{gentle})$, $(\texttt{dominant}, \texttt{caring})$, $(\texttt{rugged}, \texttt{delicate})$, $(\texttt{ambitious}, \texttt{compassionate})$. Occupation effects are averaged over all eight adjective values and four prompt templates, whereas adjective effects are averaged over all eight occupation values and the same four templates. This produces 128 reference pairs for each component, from which we sample 16 pairs. For each component, the total effect is the change in the normalized `he` preference between the female- and male-associated conditions. The computational setting is the same as in the previous experiment.

Additional Results Table 8 reports three-group abstractions for different numbers of low-level mediators, $|\mathcal{I}| \in \{4, 8, 12, 16, 24, 32\}$, at intervention layers 0, 5, and 10. We did not consider $|\mathcal{I}| > 32$ due to computational resource constraints. Abstraction quality varies with both intervention layer and mediator granularity.

At layer 5, finer mediator partitions generally yield clearer separation between occupation and adjective effects. For $|\mathcal{I}| = 32$, the occupation effect is predominantly concentrated in H_1 (0.503 of a total effect of 0.659), while the adjective effect is predominantly concentrated in H_3 (0.155 of 0.207). A similar separation is observed for $|\mathcal{I}| = 24$, where occupation and adjective effects have distinct dominant mediators, H_1 and H_2, respectively.

Layer 0 also yields distinct dominant mediators at finer granularities. For $|\mathcal{I}| = 32$, the occupation effect is strongly concentrated in H_2 (0.582 of 0.659), whereas the adjective effect is less concentrated, with 0.113 of 0.207 attributed to its dominant mediator H_1. At layer 10, the dominant mediators are distinct for $|\mathcal{I}| \geq 12$, but the adjective effect often remains split between two high-level mediators, indicating weaker completeness.

Overall, finer mediator partitions generally improve abstraction quality, although the improvement is not uniform across granularities. At finer resolutions, layers 0 and 5 yield lower non-dominant leakage than layer 10 while maintaining distinct dominant mediators.

Tables 9–11 report the low-level Shapley indirect effects without abstraction. These effects are distributed across many low-level components, making the roles associated with occupation and adjective cues less apparent than in the high-level abstractions.

Different Seed To assess robustness to random sampling, we repeat the abstraction procedure using a different random seed, as reported in Table 12. At $|\mathcal{I}| = 32$, we observe a similar qualitative pattern: occupation and adjective effects have distinct dominant mediators at all evaluated layers, with lower non-dominant leakage at layers 0 and 5 than at layer 10. However, the selected partitions, effect magnitudes, and results at other granularities vary across runs.

Table 8: (Gender Bias) Three-group abstractions across intervention layers and different numbers of low-level mediators. L denotes non-dominant leakage and D denotes distinctness.

Layer	$\lvert\mathcal{I}\rvert$	Selected abstraction	Occupation IEs	Dom.	Adjective IEs	Dom.	L	D
0	4	$H_1 = \{0\}, H_2 = \{1, 2\}, H_3 = \{3\}$	$(0.098, \mathbf{0.459}, 0.102)$	H_2	$(0.014, \mathbf{0.156}, 0.038)$	H_2	0.276	0
0	8	$H_1 = \{0, 1, 2\}, H_2 = \{3, 6, 7\}, H_3 = \{4, 5\}$	$(\mathbf{0.249}, 0.217, 0.193)$	H_1	$(0.057, \mathbf{0.084}, 0.066)$	H_2	0.609	1
0	12	$H_1 = \{0, 4, 6, 7, 9\}, H_2 = \{1, 2, 3, 5, 8, 10\}, H_3 = \{11\}$	$(0.190, \mathbf{0.450}, 0.019)$	H_2	$(\mathbf{0.111}, 0.090, 0.007)$	H_1	0.392	1
0	16	$H_1 = \{0, 5, 6, 8, 10, 11, 15\}, H_2 = \{1, 2, 4, 7, 9, 12, 13, 14\}, H_3 = \{3\}$	$(0.196, \mathbf{0.453}, 0.011)$	H_2	$(\mathbf{0.111}, 0.095, 0.002)$	H_1	0.390	1
0	24	$H_1 = \{0, 6, 8, 9, 10, 13, 14, 15, 19, 23\}, H_2 = \{1, 2, 3, 4, 7, 11, 12, 16, 17, 18, 20, 21, 22\}, H_3 = \{5\}$	$(0.125, \mathbf{0.533}, 0.001)$	H_2	$(\mathbf{0.127}, 0.078, 0.002)$	H_1	0.289	1
0	32	$H_1 = \{0, 4, 9, 10, 12, 13, 14, 17, 21, 25, 26, 31\}, H_2 = \{1, 2, 3, 5, 7, 8, 11, 15, 16, 18, 19, 20, 22, 23, 24, 27, 28, 29, 30\}, H_3 = \{6\}$	$(0.078, \mathbf{0.582}, -0.001)$	H_2	$(\mathbf{0.113}, 0.094, 0.001)$	H_1	0.288	1
5	4	$H_1 = \{0, 1\}, H_2 = \{2\}, H_3 = \{3\}$	$(\mathbf{0.337}, 0.158, 0.164)$	H_1	$(0.071, 0.064, \mathbf{0.073})$	H_3	0.568	1
5	8	$H_1 = \{0, 1, 3, 5\}, H_2 = \{2, 4, 7\}, H_3 = \{6\}$	$(\mathbf{0.285}, 0.281, 0.093)$	H_1	$(0.047, \mathbf{0.121}, 0.039)$	H_2	0.491	1
5	12	$H_1 = \{0\}, H_2 = \{1, 2, 3, 5, 6, 8, 10\}, H_3 = \{4, 7, 9, 11\}$	$(0.020, \mathbf{0.484}, 0.155)$	H_2	$(-0.001, 0.092, \mathbf{0.116})$	H_3	0.356	1
5	16	$H_1 = \{0, 2, 3, 4, 7, 11, 13\}, H_2 = \{1, 10\}, H_3 = \{5, 6, 8, 9, 12, 14, 15\}$	$(\mathbf{0.423}, 0.013, 0.224)$	H_1	$(0.071, 0.003, \mathbf{0.133})$	H_3	0.359	1
5	24	$H_1 = \{0, 1, 2, 3, 4, 7, 10, 11, 12, 17, 19, 20, 23\}, H_2 = \{5, 6, 8, 9, 13, 14, 15, 16, 18, 22\}, H_3 = \{21\}$	$(\mathbf{0.523}, 0.130, 0.007)$	H_1	$(0.064, \mathbf{0.142}, 0.001)$	H_2	0.260	1
5	32	$H_1 = \{0, 2, 3, 4, 5, 6, 9, 14, 15, 16, 20, 22, 23, 26, 27, 28, 31\}, H_2 = \{1\}, H_3 = \{7, 8, 10, 11, 12, 13, 17, 18, 19, 21, 24, 25, 29, 30\}$	$(\mathbf{0.503}, 0.005, 0.151)$	H_1	$(0.052, 0.001, \mathbf{0.155})$	H_3	0.245	1
10	4	$H_1 = \{0\}, H_2 = \{1\}, H_3 = \{2, 3\}$	$(0.151, 0.138, \mathbf{0.370})$	H_3	$(0.046, 0.041, \mathbf{0.120})$	H_3	0.430	0
10	8	$H_1 = \{0\}, H_2 = \{1, 2, 3, 4, 6, 7\}, H_3 = \{5\}$	$(0.060, \mathbf{0.501}, 0.098)$	H_2	$(0.018, \mathbf{0.148}, 0.041)$	H_2	0.263	0
10	12	$H_1 = \{0, 2, 4, 7, 9\}, H_2 = \{1, 3, 5, 8, 10, 11\}, H_3 = \{6\}$	$(0.262, \mathbf{0.379}, 0.018)$	H_2	$(\mathbf{0.104}, 0.093, 0.011)$	H_1	0.463	1
10	16	$H_1 = \{0, 3, 6, 8, 10, 12, 13\}, H_2 = \{1, 2, 4, 5, 7, 11, 14, 15\}, H_3 = \{9\}$	$(0.242, \mathbf{0.404}, 0.013)$	H_2	$(\mathbf{0.111}, 0.090, 0.007)$	H_1	0.427	1
10	24	$H_1 = \{0, 2, 3, 4, 6, 7, 8, 11, 17, 20, 21, 22, 23\}, H_2 = \{1, 5, 9, 10, 12, 13, 15, 16, 18, 19\}, H_3 = \{14\}$	$(\mathbf{0.428}, 0.222, 0.010)$	H_1	$(0.090, \mathbf{0.113}, 0.004)$	H_2	0.403	1
10	32	$H_1 = \{0, 1, 7, 12, 18, 19, 20, 21, 24, 26, 31\}, H_2 = \{2, 3, 4, 5, 6, 8, 9, 10, 11, 14, 15, 16, 22, 23, 25, 27, 28, 29, 30\}, H_3 = \{13, 17\}$	$(0.166, \mathbf{0.493}, 0.000)$	H_2	$(\mathbf{0.105}, 0.100, 0.002)$	H_1	0.372	1

Table 9: Low-level Shapley indirect effects without abstraction at the input to layer 0 for the gender-bias experiment. The 768-dimensional residual representation is partitioned into $|\mathcal{I}|$ contiguous blocks. For each granularity, the estimated indirect effects sum to the corresponding total effect ($\mathrm{TE}_{\mathrm{Occupation}} = 0.659$ and $\mathrm{TE}_{\mathrm{Adjective}} = 0.207$).

$\lvert\mathcal{I}\rvert$	Block width	Occupation IEs	Adjective IEs
32	24	$M_0 : 0.010,\ M_1 : 0.014,\ M_2 : 0.013,\ M_3 : 0.019,$ $M_4 : 0.012,\ M_5 : 0.038,\ M_6 : 0.008,\ M_7 : 0.011,$ $M_8 : 0.044,\ M_9 : 0.041,\ M_{10} : 0.007,\ M_{11} : 0.019,$ $M_{12} : 0.032,\ M_{13} : 0.005,\ M_{14} : 0.026,\ M_{15} : 0.011,$ $M_{16} : 0.030,\ M_{17} : 0.007,\ M_{18} : 0.023,\ M_{19} : 0.013,$ $M_{20} : 0.045,\ M_{21} : 0.017,\ M_{22} : 0.030,\ M_{23} : 0.038,$ $M_{24} : 0.040,\ M_{25} : 0.011,\ M_{26} : 0.006,\ M_{27} : 0.023,$ $M_{28} : 0.025,\ M_{29} : 0.013,\ M_{30} : 0.033,\ M_{31} : -0.006$	$M_0 : 0.010,\ M_1 : 0.002,\ M_2 : -0.002,\ M_3 : -0.004,$ $M_4 : 0.005,\ M_5 : 0.007,\ M_6 : 0.003,\ M_7 : 0.001,$ $M_8 : 0.014,\ M_9 : 0.016,\ M_{10} : 0.007,\ M_{11} : 0.000,$ $M_{12} : 0.015,\ M_{13} : 0.011,\ M_{14} : 0.007,\ M_{15} : 0.001,$ $M_{16} : 0.005,\ M_{17} : 0.013,\ M_{18} : 0.005,\ M_{19} : 0.004,$ $M_{20} : 0.013,\ M_{21} : 0.012,\ M_{22} : 0.010,\ M_{23} : 0.008,$ $M_{24} : 0.013,\ M_{25} : 0.007,\ M_{26} : 0.014,\ M_{27} : -0.000,$ $M_{28} : 0.003,\ M_{29} : -0.000,\ M_{30} : 0.003,\ M_{31} : 0.007$
24	32	$M_0 : 0.009,\ M_1 : 0.023,\ M_2 : 0.024,\ M_3 : 0.023,$ $M_4 : 0.034,\ M_5 : 0.011,\ M_6 : 0.050,\ M_7 : 0.050,$ $M_8 : 0.015,\ M_9 : 0.035,\ M_{10} : 0.017,\ M_{11} : 0.029,$ $M_{12} : 0.036,\ M_{13} : 0.011,\ M_{14} : 0.027,\ M_{15} : 0.047,$ $M_{16} : 0.029,\ M_{17} : 0.049,\ M_{18} : 0.049,\ M_{19} : 0.007,$ $M_{20} : 0.022,\ M_{21} : 0.034,\ M_{22} : 0.031,\ M_{23} : -0.002$	$M_0 : 0.007,\ M_1 : -0.000,\ M_2 : -0.003,\ M_3 : 0.006,$ $M_4 : 0.009,\ M_5 : 0.002,\ M_6 : 0.020,\ M_7 : 0.013,$ $M_8 : 0.006,\ M_9 : 0.017,\ M_{10} : 0.013,\ M_{11} : 0.004,$ $M_{12} : 0.010,\ M_{13} : 0.008,\ M_{14} : 0.009,\ M_{15} : 0.018,$ $M_{16} : 0.010,\ M_{17} : 0.014,\ M_{18} : 0.015,\ M_{19} : 0.023,$ $M_{20} : -0.004,\ M_{21} : 0.005,\ M_{22} : -0.002,\ M_{23} : 0.010$
16	48	$M_0 : 0.027,\ M_1 : 0.037,\ M_2 : 0.045,\ M_3 : 0.016,$ $M_4 : 0.081,\ M_5 : 0.028,\ M_6 : 0.037,\ M_7 : 0.038,$ $M_8 : 0.036,\ M_9 : 0.041,\ M_{10} : 0.062,\ M_{11} : 0.059,$ $M_{12} : 0.047,\ M_{13} : 0.029,\ M_{14} : 0.053,\ M_{15} : 0.022$	$M_0 : 0.011,\ M_1 : -0.006,\ M_2 : 0.013,\ M_3 : 0.003,$ $M_4 : 0.028,\ M_5 : 0.008,\ M_6 : 0.025,\ M_7 : 0.008,$ $M_8 : 0.020,\ M_9 : 0.009,\ M_{10} : 0.025,\ M_{11} : 0.016,$ $M_{12} : 0.019,\ M_{13} : 0.014,\ M_{14} : 0.004,\ M_{15} : 0.010$
12	64	$M_0 : 0.033,\ M_1 : 0.038,\ M_2 : 0.048,\ M_3 : 0.112,$ $M_4 : 0.042,\ M_5 : 0.042,\ M_6 : 0.047,\ M_7 : 0.070,$ $M_8 : 0.075,\ M_9 : 0.063,\ M_{10} : 0.058,\ M_{11} : 0.029$	$M_0 : 0.008,\ M_1 : 0.002,\ M_2 : 0.011,\ M_3 : 0.033,$ $M_4 : 0.023,\ M_5 : 0.017,\ M_6 : 0.018,\ M_7 : 0.027,$ $M_8 : 0.023,\ M_9 : 0.041,\ M_{10} : -0.001,\ M_{11} : 0.007$
8	96	$M_0 : 0.057,\ M_1 : 0.065,\ M_2 : 0.117,\ M_3 : 0.074,$ $M_4 : 0.078,\ M_5 : 0.129,\ M_6 : 0.088,\ M_7 : 0.051$	$M_0 : 0.005,\ M_1 : 0.015,\ M_2 : 0.039,\ M_3 : 0.033,$ $M_4 : 0.027,\ M_5 : 0.044,\ M_6 : 0.034,\ M_7 : 0.010$
4	192	$M_0 : 0.125,\ M_1 : 0.192,$ $M_2 : 0.200,\ M_3 : 0.142$	$M_0 : 0.020,\ M_1 : 0.072,$ $M_2 : 0.070,\ M_3 : 0.045$

Table 10: Low-level Shapley indirect effects without abstraction at the input to layer 5 for the gender-bias experiment. The 768-dimensional residual representation is partitioned into $|\mathcal{I}|$ contiguous blocks. For each granularity, the estimated indirect effects sum to the corresponding total effect ($\mathrm{TE}_{\mathrm{Occupation}} = 0.659$ and $\mathrm{TE}_{\mathrm{Adjective}} = 0.207$).

$\lvert\mathcal{I}\rvert$	Block width	Occupation IEs	Adjective IEs
32	24	$M_0 : 0.018,\ M_1 : 0.008,\ M_2 : -0.005,\ M_3 : 0.013,$ $M_4 : 0.034,\ M_5 : -0.002,\ M_6 : 0.039,\ M_7 : 0.005,$ $M_8 : 0.014,\ M_9 : 0.048,\ M_{10} : 0.008,\ M_{11} : 0.023,$ $M_{12} : 0.008,\ M_{13} : 0.006,\ M_{14} : 0.047,\ M_{15} : 0.045,$ $M_{16} : 0.054,\ M_{17} : 0.016,\ M_{18} : 0.020,\ M_{19} : 0.015,$ $M_{20} : 0.015,\ M_{21} : -0.010,\ M_{22} : 0.029,\ M_{23} : 0.032,$ $M_{24} : 0.043,\ M_{25} : 0.000,\ M_{26} : 0.049,\ M_{27} : 0.019,$ $M_{28} : 0.004,\ M_{29} : 0.019,\ M_{30} : 0.035,\ M_{31} : 0.012$	$M_0 : 0.002,\ M_1 : 0.002,\ M_2 : -0.005,\ M_3 : -0.001,$ $M_4 : 0.009,\ M_5 : -0.005,\ M_6 : 0.001,\ M_7 : 0.002,$ $M_8 : 0.014,\ M_9 : 0.008,\ M_{10} : 0.004,\ M_{11} : 0.013,$ $M_{12} : 0.008,\ M_{13} : 0.003,\ M_{14} : 0.009,\ M_{15} : 0.004,$ $M_{16} : 0.001,\ M_{17} : 0.019,\ M_{18} : 0.016,\ M_{19} : 0.010,$ $M_{20} : 0.007,\ M_{21} : 0.003,\ M_{22} : 0.008,\ M_{23} : 0.000,$ $M_{24} : 0.020,\ M_{25} : 0.005,\ M_{26} : 0.011,\ M_{27} : 0.004,$ $M_{28} : -0.004,\ M_{29} : 0.018,\ M_{30} : 0.019,\ M_{31} : 0.002$
24	32	$M_0 : 0.015,\ M_1 : 0.009,\ M_2 : 0.011,\ M_3 : 0.040,$ $M_4 : 0.037,\ M_5 : -0.006,\ M_6 : 0.022,\ M_7 : 0.047,$ $M_8 : 0.028,\ M_9 : 0.005,\ M_{10} : 0.041,\ M_{11} : 0.067,$ $M_{12} : 0.057,\ M_{13} : 0.025,\ M_{14} : 0.019,\ M_{15} : 0.002,$ $M_{16} : -0.008,\ M_{17} : 0.076,\ M_{18} : 0.042,\ M_{19} : 0.037,$ $M_{20} : 0.024,\ M_{21} : 0.012,\ M_{22} : 0.041,\ M_{23} : 0.015$	$M_0 : 0.000,\ M_1 : -0.001,\ M_2 : -0.002,\ M_3 : 0.005,$ $M_4 : -0.001,\ M_5 : 0.004,\ M_6 : 0.014,\ M_7 : 0.010,$ $M_8 : 0.014,\ M_9 : 0.009,\ M_{10} : 0.011,\ M_{11} : 0.006,$ $M_{12} : 0.007,\ M_{13} : 0.018,\ M_{14} : 0.022,\ M_{15} : 0.007,$ $M_{16} : 0.005,\ M_{17} : 0.005,\ M_{18} : 0.023,\ M_{19} : 0.012,$ $M_{20} : 0.005,\ M_{21} : 0.002,\ M_{22} : 0.028,\ M_{23} : 0.004$
16	48	$M_0 : 0.026,\ M_1 : 0.009,\ M_2 : 0.031,\ M_3 : 0.040,$ $M_4 : 0.062,\ M_5 : 0.029,\ M_6 : 0.013,\ M_7 : 0.097,$ $M_8 : 0.074,\ M_9 : 0.037,\ M_{10} : 0.005,\ M_{11} : 0.056,$ $M_{12} : 0.041,\ M_{13} : 0.066,\ M_{14} : 0.026,\ M_{15} : 0.047$	$M_0 : 0.004,\ M_1 : -0.006,\ M_2 : 0.004,\ M_3 : 0.004,$ $M_4 : 0.023,\ M_5 : 0.017,\ M_6 : 0.011,\ M_7 : 0.013,$ $M_8 : 0.020,\ M_9 : 0.026,\ M_{10} : 0.010,\ M_{11} : 0.008,$ $M_{12} : 0.024,\ M_{13} : 0.016,\ M_{14} : 0.013,\ M_{15} : 0.021$
12	64	$M_0 : 0.025,\ M_1 : 0.044,\ M_2 : 0.035,\ M_3 : 0.072,$ $M_4 : 0.034,\ M_5 : 0.104,\ M_6 : 0.086,\ M_7 : 0.022,$ $M_8 : 0.059,\ M_9 : 0.084,\ M_{10} : 0.037,\ M_{11} : 0.055$	$M_0 : -0.001,\ M_1 : 0.003,\ M_2 : 0.002,\ M_3 : 0.027,$ $M_4 : 0.022,\ M_5 : 0.016,\ M_6 : 0.025,\ M_7 : 0.028,$ $M_8 : 0.010,\ M_9 : 0.035,\ M_{10} : 0.008,\ M_{11} : 0.032$
8	96	$M_0 : 0.035,\ M_1 : 0.071,\ M_2 : 0.098,\ M_3 : 0.109,$ $M_4 : 0.111,\ M_5 : 0.066,\ M_6 : 0.105,\ M_7 : 0.064$	$M_0 : -0.003,\ M_1 : 0.007,\ M_2 : 0.040,\ M_3 : 0.024,$ $M_4 : 0.046,\ M_5 : 0.018,\ M_6 : 0.042,\ M_7 : 0.033$
4	192	$M_0 : 0.105,\ M_1 : 0.210,$ $M_2 : 0.169,\ M_3 : 0.175$	$M_0 : 0.003,\ M_1 : 0.065,$ $M_2 : 0.064,\ M_3 : 0.076$

Table 11: Low-level Shapley indirect effects without abstraction at the input to layer 10 for the gender-bias experiment. The 768-dimensional residual representation is partitioned into $|\mathcal{I}|$ contiguous blocks. For each granularity, the estimated indirect effects sum to the corresponding total effect ($\mathrm{TE}_{\mathrm{Occupation}} = 0.659$ and $\mathrm{TE}_{\mathrm{Adjective}} = 0.207$).

$\lvert\mathcal{I}\rvert$	Block width	Occupation IEs	Adjective IEs
32	24	$M_0 : 0.021,\ M_1 : 0.027,\ M_2 : 0.022,\ M_3 : 0.009,$ $M_4 : 0.023,\ M_5 : 0.029,\ M_6 : 0.018,\ M_7 : 0.017,$ $M_8 : 0.021,\ M_9 : 0.021,\ M_{10} : 0.011,\ M_{11} : 0.017,$ $M_{12} : 0.021,\ M_{13} : 0.002,\ M_{14} : 0.038,\ M_{15} : 0.018,$ $M_{16} : 0.021,\ M_{17} : -0.001,\ M_{18} : 0.004,\ M_{19} : 0.010,$ $M_{20} : 0.015,\ M_{21} : 0.021,\ M_{22} : 0.038,\ M_{23} : 0.051,$ $M_{24} : 0.024,\ M_{25} : 0.029,\ M_{26} : 0.028,\ M_{27} : 0.020,$ $M_{28} : 0.011,\ M_{29} : 0.020,\ M_{30} : 0.039,\ M_{31} : 0.012$	$M_0 : 0.008,\ M_1 : 0.010,\ M_2 : 0.003,\ M_3 : -0.003,$ $M_4 : 0.004,\ M_5 : 0.012,\ M_6 : 0.006,\ M_7 : 0.006,$ $M_8 : 0.008,\ M_9 : -0.001,\ M_{10} : 0.001,\ M_{11} : 0.002,$ $M_{12} : 0.016,\ M_{13} : -0.001,\ M_{14} : 0.012,\ M_{15} : 0.004,$ $M_{16} : 0.006,\ M_{17} : 0.003,\ M_{18} : 0.003,\ M_{19} : 0.004,$ $M_{20} : 0.008,\ M_{21} : 0.008,\ M_{22} : 0.011,\ M_{23} : 0.015,$ $M_{24} : 0.014,\ M_{25} : 0.006,\ M_{26} : 0.020,\ M_{27} : 0.000,$ $M_{28} : 0.004,\ M_{29} : 0.004,\ M_{30} : 0.005,\ M_{31} : 0.008$
24	32	$M_0 : 0.020,\ M_1 : 0.044,\ M_2 : 0.013,\ M_3 : 0.038,$ $M_4 : 0.037,\ M_5 : 0.015,\ M_6 : 0.026,\ M_7 : 0.023,$ $M_8 : 0.022,\ M_9 : 0.024,\ M_{10} : 0.013,\ M_{11} : 0.046,$ $M_{12} : 0.015,\ M_{13} : 0.008,\ M_{14} : 0.011,\ M_{15} : 0.020,$ $M_{16} : 0.028,\ M_{17} : 0.082,\ M_{18} : 0.036,\ M_{19} : 0.042,$ $M_{20} : 0.021,\ M_{21} : 0.016,\ M_{22} : 0.035,\ M_{23} : 0.027$	$M_0 : 0.007,\ M_1 : 0.013,\ M_2 : -0.001,\ M_3 : 0.009,$ $M_4 : 0.014,\ M_5 : 0.005,\ M_6 : 0.005,\ M_7 : 0.007,$ $M_8 : -0.002,\ M_9 : 0.014,\ M_{10} : 0.011,\ M_{11} : 0.006,$ $M_{12} : 0.007,\ M_{13} : 0.005,\ M_{14} : 0.004,\ M_{15} : 0.009,$ $M_{16} : 0.011,\ M_{17} : 0.022,\ M_{18} : 0.015,\ M_{19} : 0.023,$ $M_{20} : 0.002,\ M_{21} : 0.004,\ M_{22} : 0.010,\ M_{23} : 0.008$
16	48	$M_0 : 0.047,\ M_1 : 0.032,\ M_2 : 0.052,\ M_3 : 0.033,$ $M_4 : 0.042,\ M_5 : 0.030,\ M_6 : 0.023,\ M_7 : 0.058,$ $M_8 : 0.020,\ M_9 : 0.016,\ M_{10} : 0.036,\ M_{11} : 0.088,$ $M_{12} : 0.051,\ M_{13} : 0.047,\ M_{14} : 0.034,\ M_{15} : 0.052$	$M_0 : 0.018,\ M_1 : 0.000,\ M_2 : 0.016,\ M_3 : 0.013,$ $M_4 : 0.008,\ M_5 : 0.003,\ M_6 : 0.014,\ M_7 : 0.016,$ $M_8 : 0.009,\ M_9 : 0.007,\ M_{10} : 0.015,\ M_{11} : 0.026,$ $M_{12} : 0.020,\ M_{13} : 0.020,\ M_{14} : 0.008,\ M_{15} : 0.013$
12	64	$M_0 : 0.066,\ M_1 : 0.047,\ M_2 : 0.054,\ M_3 : 0.050,$ $M_4 : 0.047,\ M_5 : 0.055,\ M_6 : 0.023,\ M_7 : 0.030,$ $M_8 : 0.106,\ M_9 : 0.081,\ M_{10} : 0.038,\ M_{11} : 0.062$	$M_0 : 0.020,\ M_1 : 0.008,\ M_2 : 0.020,\ M_3 : 0.012,$ $M_4 : 0.012,\ M_5 : 0.017,\ M_6 : 0.012,\ M_7 : 0.013,$ $M_8 : 0.033,\ M_9 : 0.038,\ M_{10} : 0.005,\ M_{11} : 0.019$
8	96	$M_0 : 0.083,\ M_1 : 0.086,\ M_2 : 0.072,\ M_3 : 0.078,$ $M_4 : 0.034,\ M_5 : 0.129,\ M_6 : 0.103,\ M_7 : 0.075$	$M_0 : 0.018,\ M_1 : 0.028,\ M_2 : 0.010,\ M_3 : 0.030,$ $M_4 : 0.016,\ M_5 : 0.042,\ M_6 : 0.040,\ M_7 : 0.022$
4	192	$M_0 : 0.168,\ M_1 : 0.149,$ $M_2 : 0.159,\ M_3 : 0.182$	$M_0 : 0.046,\ M_1 : 0.041,$ $M_2 : 0.058,\ M_3 : 0.062$

Table 12: (Gender bias; different-seed run) Three-group abstractions across intervention layers and different numbers of low-level mediators. L denotes non-dominant leakage and D denotes distinctness.

Layer	$\lvert\mathcal{I}\rvert$	Selected abstraction	Occupation IEs	Dom.	Adjective IEs	Dom.	L	D
0	4	$H_1 = \{0, 3\}, H_2 = \{1\}, H_3 = \{2\}$	$(\mathbf{0.297}, 0.138, 0.149)$	H_1	$(0.052, \mathbf{0.063}, 0.056)$	H_2	0.563	1
0	8	$H_1 = \{0, 1, 3, 4, 6, 7\}, H_2 = \{2\}, H_3 = \{5\}$	$(\mathbf{0.476}, 0.052, 0.056)$	H_1	$(\mathbf{0.115}, 0.028, 0.028)$	H_1	0.255	0
0	12	$H_1 = \{0, 1, 3, 5, 8, 10\}, H_2 = \{2, 4, 6, 7, 9\}, H_3 = \{11\}$	$(\mathbf{0.354}, 0.206, 0.025)$	H_1	$(0.046, \mathbf{0.117}, 0.008)$	H_2	0.355	1
0	16	$H_1 = \{0, 5, 6, 8, 10, 12, 15\}, H_2 = \{1, 2, 4, 7, 9, 11, 13, 14\}, H_3 = \{3\}$	$(0.188, \mathbf{0.384}, 0.012)$	H_2	$(\mathbf{0.124}, 0.047, 0.000)$	H_1	0.309	1
0	24	$H_1 = \{0\}, H_2 = \{1, 2, 3, 7, 11, 12, 14, 16, 18, 20, 21, 22\}, H_3 = \{4, 5, 6, 8, 9, 10, 13, 15, 17, 19, 23\}$	$(0.011, \mathbf{0.391}, 0.182)$	H_2	$(-0.001, 0.004, \mathbf{0.167})$	H_3	0.180	1
0	32	$H_1 = \{0, 1, 8, 10, 11, 13, 17, 20, 21, 23, 26, 29, 31\}, H_2 = \{2, 3, 4, 5, 7, 9, 12, 14, 15, 16, 18, 19, 22, 24, 25, 27, 28, 30\}, H_3 = \{6\}$	$(0.116, \mathbf{0.468}, 0.001)$	H_2	$(\mathbf{0.125}, 0.045, 0.000)$	H_1	0.232	1
5	4	$H_1 = \{0, 1\}, H_2 = \{2\}, H_3 = \{3\}$	$(\mathbf{0.304}, 0.137, 0.144)$	H_1	$(0.059, 0.041, \mathbf{0.071})$	H_3	0.532	1
5	8	$H_1 = \{0, 3, 4, 5\}, H_2 = \{1\}, H_3 = \{2, 6, 7\}$	$(\mathbf{0.295}, 0.061, 0.228)$	H_1	$(0.057, 0.003, \mathbf{0.111})$	H_3	0.423	1
5	12	$H_1 = \{0\}, H_2 = \{1, 2, 3, 5, 6, 8, 10\}, H_3 = \{4, 7, 9, 11\}$	$(0.026, \mathbf{0.408}, 0.151)$	H_2	$(0.004, 0.047, \mathbf{0.120})$	H_3	0.300	1
5	16	$H_1 = \{0, 5, 6, 9, 12, 14, 15\}, H_2 = \{1, 3, 4, 7, 8, 10, 11, 13\}, H_3 = \{2\}$	$(0.198, \mathbf{0.374}, 0.013)$	H_2	$(\mathbf{0.131}, 0.038, 0.002)$	H_1	0.298	1
5	24	$H_1 = \{0, 6, 8, 9, 13, 14, 16, 18, 22\}, H_2 = \{1, 2, 3, 4, 7, 10, 11, 12, 15, 17, 19, 20, 21, 23\}, H_3 = \{5\}$	$(0.140, \mathbf{0.432}, 0.013)$	H_2	$(\mathbf{0.140}, 0.030, 0.001)$	H_1	0.221	1
5	32	$H_1 = \{0, 2, 3, 6, 7, 8, 9, 13, 14, 15, 16, 20, 21, 23, 26, 27, 28, 31\}, H_2 = \{1, 4, 10, 11, 12, 18, 19, 22, 24, 25, 29, 30\}, H_3 = \{5, 17\}$	$(\mathbf{0.387}, 0.189, 0.009)$	H_1	$(0.023, \mathbf{0.147}, 0.001)$	H_2	0.239	1
10	4	$H_1 = \{0, 1\}, H_2 = \{2\}, H_3 = \{3\}$	$(\mathbf{0.308}, 0.132, 0.145)$	H_1	$(0.054, 0.049, \mathbf{0.068})$	H_3	0.538	1
10	8	$H_1 = \{0, 1, 2, 5, 7\}, H_2 = \{3, 6\}, H_3 = \{4\}$	$(\mathbf{0.406}, 0.146, 0.033)$	H_1	$(0.079, \mathbf{0.083}, 0.008)$	H_2	0.410	1
10	12	$H_1 = \{0, 1, 2, 3, 5, 8, 10\}, H_2 = \{4, 6, 9, 11\}, H_3 = \{7\}$	$(\mathbf{0.389}, 0.174, 0.022)$	H_1	$(0.082, \mathbf{0.088}, 0.001)$	H_2	0.409	1
10	16	$H_1 = \{0, 1, 2, 4, 5, 7, 9, 11, 15\}, H_2 = \{3\}, H_3 = \{6, 8, 10, 12, 13, 14\}$	$(\mathbf{0.408}, 0.018, 0.159)$	H_1	$(0.065, 0.005, \mathbf{0.101})$	H_3	0.355	1
10	24	$H_1 = \{0, 1, 2, 3, 5, 6, 8, 11, 13, 14, 17, 20, 21, 23\}, H_2 = \{4, 7, 9, 10, 12, 16, 18, 19, 22\}, H_3 = \{15\}$	$(\mathbf{0.376}, 0.200, 0.008)$	H_1	$(0.030, \mathbf{0.136}, 0.004)$	H_2	0.279	1
10	32	$H_1 = \{0, 1, 2, 3, 4, 7, 8, 9, 10, 11, 13, 14, 18, 19, 21, 23, 25, 27, 29, 30, 31\}$, $H_2 = \{5, 12, 15, 16, 17, 20, 22, 24, 26, 28\}, H_3 = \{6\}$	$(\mathbf{0.437}, 0.143, 0.005)$	H_1	$(0.048, \mathbf{0.120}, 0.003)$	H_2	0.276	1

Experiments Across Different Numbers of High-Level Mediators J Figures 6–8 show how the indirect effect captured by the dominant high-level mediator and the size of the dominant group evolve as the number of high-level mediators J decreases during the greedy abstraction procedure. The remaining experimental settings are the same as those used for the $|\mathcal{I}| = 32$ experiment. Across the experiments, as J gradually decreases, both the indirect effect captured by the dominant group and the dominant-group size generally increase. Moreover, as the dominant-group size increases, the indirect effect captured by the dominant group also tends to increase, indicating that concentrating more low-level mediators into the dominant high-level mediator generally yields a more complete abstraction. Furthermore, we observe a mild trade-off between Occupation and Adjective: as the abstraction increasingly concentrates the effect of one attribute into its dominant group, the other does not always improve simultaneously. This behavior is consistent with the intended design, in which the two attributes are encouraged to concentrate in distinct high-level mediators.

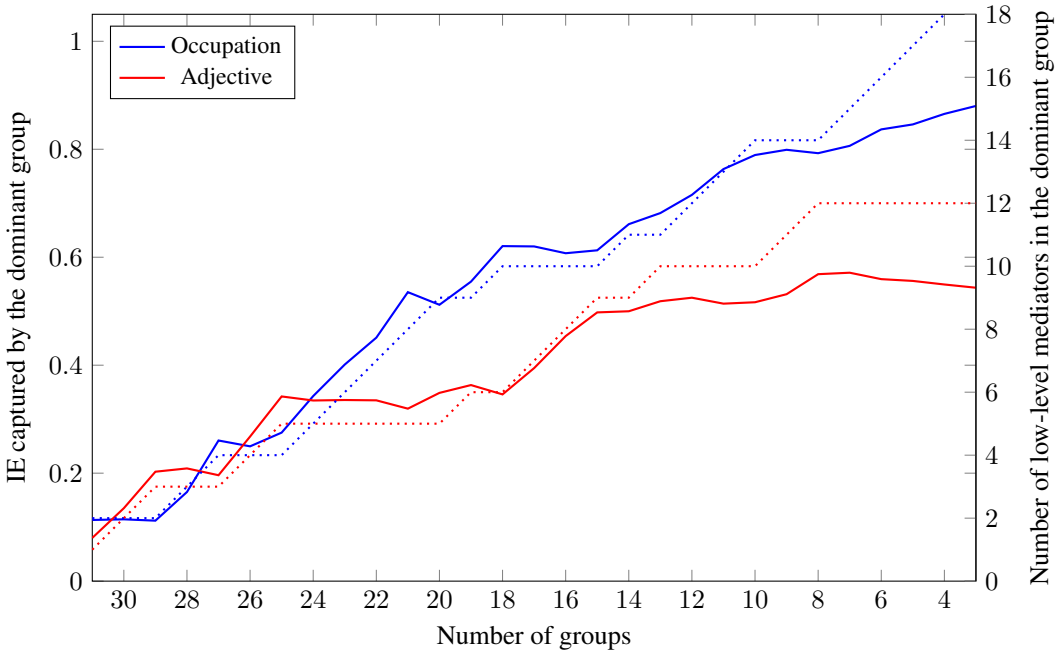


Figure 6: (Gender Bias) Evolution of IE captured by the dominant group and dominant-group size during the greedy abstraction procedure at layer 0. Solid lines show IE captured by the dominant group, while dotted lines show the number of low-level mediators in the dominant high-level mediator.

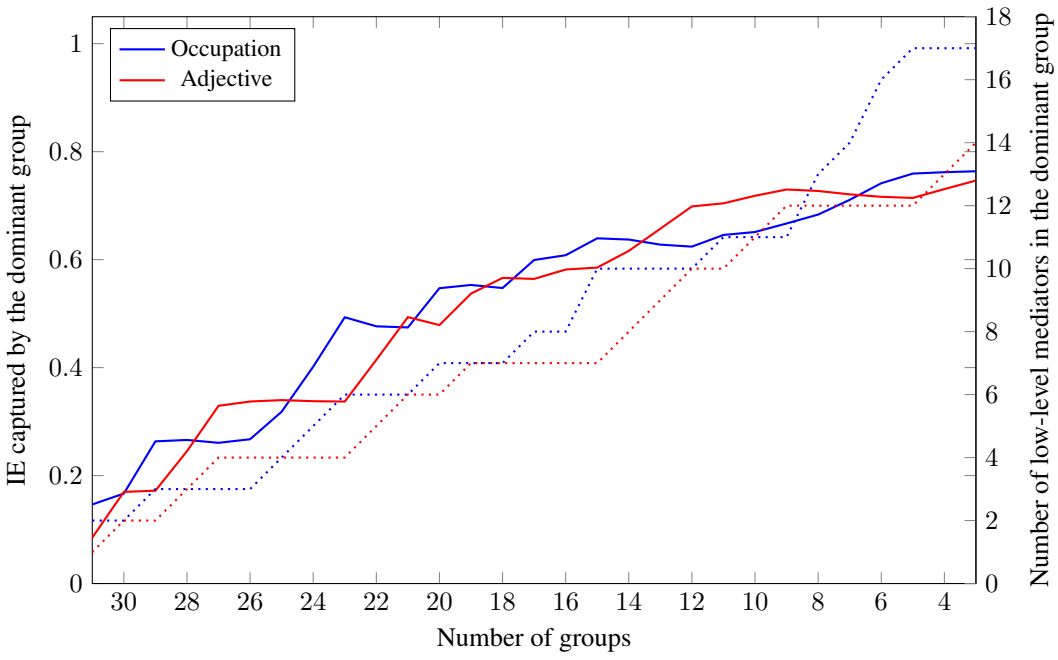


Figure 7: (Gender Bias) Evolution of IE captured by the dominant group and dominant-group size during the greedy abstraction procedure at layer 5. Solid lines show IE captured by the dominant group, while dotted lines show the number of low-level mediators in the dominant high-level mediator.

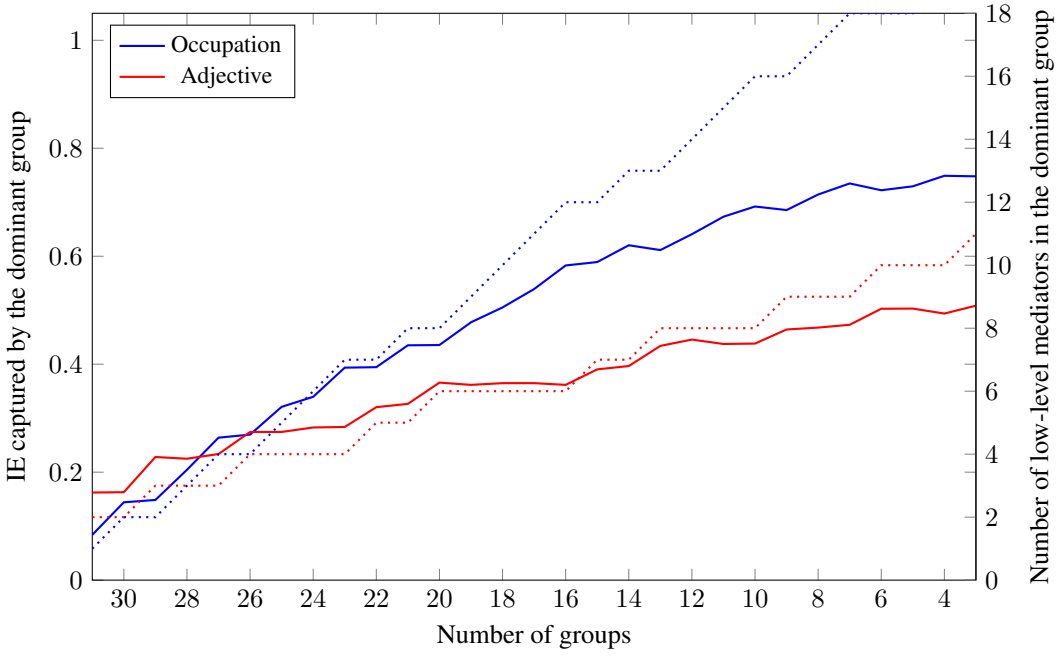


Figure 8: (Gender Bias) Evolution of IE captured by the dominant group and dominant-group size during the greedy abstraction procedure at layer 10. Solid lines show IE captured by the dominant group, while dotted lines show the number of low-level mediators in the dominant high-level mediator.

B EXPERIMENTS ON QWEN3.5-9B-BASE

B.1 THIRD-PERSON SINGULAR -*s*

As an additional model-scale experiment, we apply the abstraction procedure to Qwen3.5-9B-base. We partition the intermediate representation into $|\mathcal{I}| = 16$ low-level mediators and construct $J = 3$ high-level mediators at intervention layers 0, 10, 20, and 30. We did not consider $|\mathcal{I}| > 16$ or larger models due to computational resource constraints. Table 13 reports the selected abstractions and corresponding Shapley indirect effects.

At layer 0, the number effect is strongly concentrated in H_1, with an indirect effect of 0.368 relative to a total effect of 0.390, while the person effect is primarily concentrated in H_2, with an indirect effect of 0.138 relative to a total effect of 0.205. The resulting non-dominant leakage is $L = 0.191$, and the dominant mediators are distinct ($D = 1$).

At layer 10, the number effect is strongly concentrated in H_2, with an indirect effect of 0.358 relative to a total effect of 0.391, while the person effect is primarily concentrated in H_1, with an indirect effect of 0.126 relative to a total effect of 0.207. The resulting non-dominant leakage is $L = 0.239$, and the dominant mediators remain distinct ($D = 1$).

At layer 20, the same dominant-mediator assignment as at layer 10 is recovered: number is primarily mediated through H_2, whereas person is primarily mediated through H_1. The effects are somewhat less concentrated, yielding $L = 0.272$, while distinctness remains $D = 1$.

At layer 30, the dominant mediators remain distinct, but the mediated effects are substantially more distributed across groups. In particular, the number effect is split mainly between H_1 and H_2 (0.156 and 0.179), while the person effect is also divided between these two groups (0.094 and 0.087). Accordingly, the leakage increases to $L = 0.543$, although $D = 1$ because the two features still have different dominant mediators.

Overall, the Qwen3.5-9B-base results show clear separation between grammatical number and person at layers 0, 10, and 20, with substantially weaker completeness at layer 30.

Table 13: (Third-Person Singular -s) Selected three-group abstractions and estimated effects across intervention layers for Qwen3.5-9B-base with 16 low-level mediators. Indirect effects are reported as (IE via H_1, IE via H_2, IE via H_3).

Layer	Selected abstraction	Feature	TE	IEs	Dominant
0	$H_1 = \{0, 4, 5, 6, 7, 10, 13, 15\}$, $H_2 = \{1, 2, 3, 8, 9, 11, 12\}$, $H_3 = \{14\}$	Number	0.390	(**0.368**, 0.011, 0.011)	H_1
		Person	0.205	(0.048, **0.138**, 0.019)	H_2
10	$H_1 = \{0, 2, 3, 4, 6, 8, 9, 12\}$, $H_2 = \{1, 5, 7, 10, 11, 13, 14\}$, $H_3 = \{15\}$	Number	0.391	(0.023, **0.358**, 0.010)	H_2
		Person	0.207	(**0.126**, 0.062, 0.020)	H_1
20	$H_1 = \{0, 1, 2, 3, 8, 9, 12\}$, $H_2 = \{4, 5, 6, 7, 10, 11, 13, 14\}$, $H_3 = \{15\}$	Number	0.391	(0.068, **0.302**, 0.020)	H_2
		Person	0.207	(**0.142**, 0.053, 0.013)	H_1
30	$H_1 = \{0, 1, 3, 4, 9, 13\}$, $H_2 = \{2, 6, 7, 8, 10, 12, 14\}$, $H_3 = \{5, 11, 15\}$	Number	0.391	(0.156, **0.179**, 0.055)	H_2
		Person	0.207	(**0.094**, 0.087, 0.026)	H_1

B.2 GENDER BIAS

As an additional model-scale experiment, we apply the abstraction procedure to Qwen3.5-9B-base for occupation- and adjective-related gender effects. We partition the intermediate representation into $|\mathcal{I}| = 16$ low-level mediators and construct $J = 3$ high-level mediators at intervention layers 0, 10, 20, and 30. We did not consider $|\mathcal{I}| > 16$ or larger models due to computational resource constraints. Table 14 reports the selected abstractions and corresponding Shapley indirect effects.

At layer 0, the occupation effect is primarily concentrated in H_1, with an indirect effect of 0.380 relative to a total effect of 0.560, while the adjective effect is primarily concentrated in H_3, with

an indirect effect of 0.139 relative to a total effect of 0.200. The resulting non-dominant leakage is $L = 0.312$, and the dominant mediators are distinct ($D = 1$).

At layer 10, the occupation effect remains concentrated in H_1, with an indirect effect of 0.400 relative to a total effect of 0.560, while the adjective effect is primarily concentrated in H_2, with an indirect effect of 0.115 relative to a total effect of 0.199. The resulting leakage is $L = 0.355$, and the dominant mediators remain distinct ($D = 1$).

At layer 20, the same dominant-mediator assignment as at layer 10 is recovered: occupation is primarily mediated through H_1, whereas adjective is primarily mediated through H_2. The resulting leakage is $L = 0.327$, while distinctness remains $D = 1$.

At layer 30, the dominant mediators remain distinct, but the adjective effect is nearly evenly split between H_1 and H_2 (0.095 and 0.096). The occupation effect is also distributed across multiple groups, with indirect effects of 0.373 and 0.173 through H_1 and H_2, respectively. Accordingly, the leakage increases to $L = 0.426$, indicating weaker completeness despite $D = 1$.

Overall, the Qwen3.5-9B-base results show distinct dominant mediators for occupation and adjective effects across all evaluated layers, with weaker completeness at layer 30.

Table 14: (Gender Bias) Selected three-group abstractions and estimated effects across intervention layers for Qwen3.5-9B-base with 16 low-level mediators. Indirect effects are reported as (IE via H_1, IE via H_2, IE via H_3).

Layer	Selected abstraction	Feature	TE	IEs	Dominant
0	$H_1 = \{0, 2, 3, 5, 8, 10, 11, 15\}$, $H_2 = \{1\}$, $H_3 = \{4, 6, 7, 9, 12, 13, 14\}$	Occupation	0.560	(**0.380**, 0.034, 0.146)	H_1
		Adjective	0.200	(0.050, 0.010, **0.139**)	H_3
10	$H_1 = \{0, 3, 4, 7, 10, 11, 12, 14, 15\}$, $H_2 = \{1, 2, 5, 8, 9, 13\}$, $H_3 = \{6\}$	Occupation	0.560	(**0.400**, 0.147, 0.013)	H_1
		Adjective	0.199	(0.079, **0.115**, 0.005)	H_2
20	$H_1 = \{0, 4, 6, 7, 8, 10, 12, 14, 15\}$, $H_2 = \{1, 2, 5, 9, 11, 13\}$, $H_3 = \{3\}$	Occupation	0.560	(**0.369**, 0.188, 0.003)	H_1
		Adjective	0.200	(0.062, **0.138**, −0.000)	H_2
30	$H_1 = \{0, 5, 6, 7, 8, 9, 13, 14, 15\}$, $H_2 = \{1, 2, 4, 10, 11, 12\}$, $H_3 = \{3\}$	Occupation	0.561	(**0.373**, 0.173, 0.015)	H_1
		Adjective	0.200	(0.095, **0.096**, 0.008)	H_2