IDENTIFYING RELIABLE AND UNRELIABLE COMPONENTS OF SHAPLEY VALUES

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a method to decompose Shapley values into reliable and unreliable components. First, we prove that, given a set of AND–OR interaction patterns modeled by the DNN, evenly allocating the utility of each interaction to the input units involved in the interaction yields the Shapley value. Moreover, interaction patterns can be categorized into reliable and unreliable interactions according to their transferability across different DNNs. In this way, we define the Shapley attributions allocated from reliable interaction patterns as *reliable Shapley components*, which reflect the relatively trustworthy decision logic learned by the DNN, while the Shapley attributions allocated from unreliable interaction patterns are termed *unreliable Shapley components*. We prove that the decomposed reliable and unreliable Shapley components still satisfy the four extended Shapley axioms. Experiments demonstrate that the proposed reliable and unreliable Shapley components capture decision logic with markedly different representational qualities: (1) Reliable Shapley components are substantially more similar across LLMs and more robust to input perturbations than unreliable Shapley components. (2) Unreliable Shapley components tend to exhibit greater cancellation between positive and negative attributions, thereby contributing less to the LLM's predictions. (3) Reliable Shapley components tend to concentrate more heavily on input units relevant to the prediction than unreliable Shapley components.

1 INTRODUCTION

The interpretability of neural networks has long been a central problem in deep learning. One prominent line of research addresses this challenge using attribution methods, which quantify the contributions of different input units to DNN outputs. Existing methods include gradient-based explanations (Simonyan et al., 2013; Ancona et al., 2017), perturbation-based explanations (Ribeiro et al., 2016; Fong & Vedaldi, 2017), and game-theoretic approaches (Štrumbelj & Kononenko, 2010; Lundberg & Lee, 2017). Among them, the Shapley value (Shapley, 1953; Lundberg & Lee, 2017) is one of the most widely used attribution methods. Its four axioms, *efficiency, symmetry, dummy, and additivity*, provide a principled basis for distributing a model output among its input units.

However, the scalar score of the Shapley value cannot fully characterize the complex inference logic of a DNN. Therefore, we investigate a new scientific problem: whether we can decompose a Shapley value into a *reliable Shapley component*, which captures relatively trustworthy inference patterns, and an *unreliable Shapley component*, which captures untrustworthy inference patterns.

First, as the theoretical foundation of this study, we prove that a Shapley value can be expressed as the sum of the utilities allocated by interactions encoded by the DNN. Interaction-based explanation (Deng et al., 2022; Zhou et al., 2024; Cheng et al., 2024b) has recently emerged as a new framework for understanding DNNs' inference logic. Many studies (Li & Zhang, 2023; Ren et al., 2024) show that (1) *the prediction score of a DNN can be mathematically decomposed into numerical utilities of a small set of interaction patterns.* Each interaction captures a non-linear relation among a set of input units encoded by a DNN. Moreover, (2) *Ren et al. (2024) and Chen et al. (2024) have established the theoretical rigor of the interaction-based explanation by proving the sparsity property and the universal-matching property of interactions.*

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.

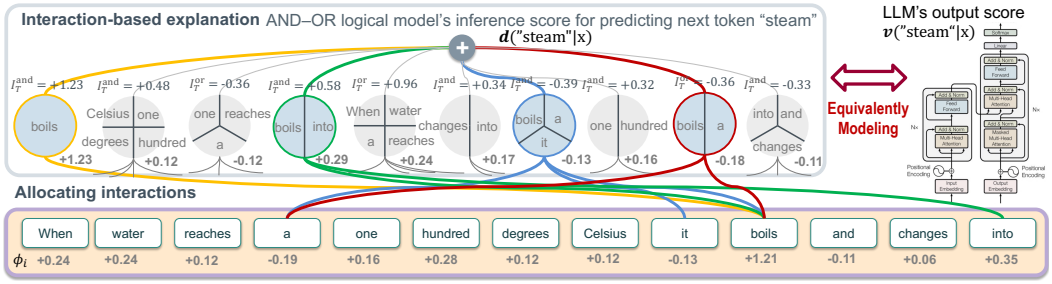


Figure 1: Illustration of recovering Shapley values by allocating AND-OR interaction utilities.

As shown in Figure 1, for the input sentence *"when water reaches ... boils and changes into"*, each interaction corresponds to a phrase pattern autonomously extracted from the input by an LLM. For example, the following two interactions $S_1 = \{boils\}$ and $S_2 = \{one, hundred\}$ are utilized to predict *steam*. The interaction S_1 encodes the relation that *boils* results in *steam*, while S_2 captures the physical context of water's boiling point. These interactions boost the prediction score for generating *steam* by 1.23 and 0.32, respectively.

In this toy example, four interactions contain the word *boils*, i.e., $S_1 = \{boils\}$, $S_3 = \{boils, into\}$, $S_4 = \{boils, a\}$, and $S_5 = \{boils, a, it\}$. We distribute the utilities of these four interactions equally among their constituent input units. Accordingly, *boils* receives the following attribution from these interactions: $\frac{1.23}{1} + \frac{0.58}{2} + \frac{-0.36}{2} + \frac{-0.39}{3} = 1.21$, which recovers its Shapley value.

Second, we can categorize interactions into reliable interactions and unreliable interactions. Accordingly, we further decompose each Shapley value into a reliable component and an unreliable component. When presented with the same prompt, different LLMs typically extract overlapping sets of interactions. Interactions commonly used by multiple LLMs for inference are considered *reliable*, whereas those exclusively used by a single LLM are considered *unreliable*. Such unreliable interactions often behave like random noise (Cheng et al., 2025; Lou et al., 2026).

In this way, we define reliable Shapley components as those that aggregate the utility allocations from reliable interactions and unreliable Shapley components as those that aggregate the utility allocations from unreliable interactions. We prove that the proposed reliable Shapley component and unreliable Shapley component also satisfy the four extended Shapley axioms, namely *efficiency*, *symmetry*, *dummy*, and *additivity*.

More importantly, reliable and unreliable Shapley components exhibit markedly different representational qualities in the following three respects:

(1) We find that reliable Shapley components are better aligned with human knowledge. Specifically, compared with unreliable Shapley components, reliable Shapley components are more concentrated on input units semantically relevant to the prediction.

(2) We find that, given the same input prompt to different LLMs, reliable Shapley components are substantially more similar across LLMs than unreliable Shapley components. Moreover, reliable Shapley components are more stable under input perturbations than unreliable Shapley components. These results suggest that reliable Shapley components capture more trustworthy attributions.

(3) We find that reliable Shapley components make a larger net contribution to the LLM's prediction. In contrast, unreliable Shapley components exhibit stronger cancellation between positive and negative attributions, thereby providing less net support for the prediction. That is, the LLM's predictions are driven primarily by the reliable Shapley components.

2 Preliminaries and Method

2.1 Preliminaries on AND-OR Interactions

Let $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ be an input sample with n input units, and let $N = \{1, 2, \ldots, n\}$ index these units. We use $v(\mathbf{x}) \in \mathbb{R}$ to denote a scalar output produced by the DNN. Several definitions

of $v(\mathbf{x})$ can be used.[1] In accordance with the convention used in prior works (Ren et al., 2024; Liu et al., 2023), we take $v(\mathbf{x})$ to be the DNN's confidence assigned to the target token $\mathbf{y}^*$:

$$v(\mathbf{x}) = f(p(\mathbf{y}^* \mid \mathbf{x})) = \log \frac{p(\mathbf{y}^* \mid \mathbf{x})}{1 - p(\mathbf{y}^* \mid \mathbf{x})}. \tag{1}$$

Using interactions to explain complex inference logic within a DNN. Chen et al. (2024) have constructed a logical model $d(\cdot)$ composed of AND-OR interaction patterns. **The universal-matching property of interactions has been proven by Chen et al. (2024) as formalized in Eq. (2), demonstrating that the logical model $d(\cdot)$ can precisely predict the DNN's outputs $v(\cdot)$ for each of the 2^n masked[2] configurations of the input $\mathbf{x}$.** This ensures the faithfulness of using interactions to explain the DNN's inference logic. *A **video** in the supplementary material provides an illustration of the AND-OR interactions.*

$$\begin{aligned} &\forall \mathbf{x}' \in \mathcal{X}, \quad |d(\mathbf{x}') - v(\mathbf{x}')| \le \epsilon, \\ &d(\mathbf{x}') = \sum_{T \in \mathcal{C}^{\text{and}}} \underbrace{I_T^{\text{and}} \cdot \delta_{\text{and}}\Big(\mathbf{x}' \text{ activates AND interaction among input units in } T\Big)}_{\text{an AND interaction between input units in } T} + \sum_{T \in \mathcal{C}^{\text{or}}} \underbrace{I_T^{\text{or}} \cdot \delta_{\text{or}}\Big(\mathbf{x}' \text{ activates OR interaction among input units in } T\Big)}_{\text{an OR interaction between input units in } T} + b, \end{aligned} \tag{2}$$

where $\mathcal{X} = \{\mathbf{x}_S \mid S \subseteq N\}$ contains all 2^n masked[2] configurations, with $\mathbf{x}_S$ retaining the input units indexed by S and masking those in $N \setminus S$. The small scalar ϵ bounds the approximation error and preserves the faithfulness of using interactions to explain the function $v(\cdot)$. The coefficients I_T^{and}, I_T^{or}, and b are learned as in Chen et al. (2024). We term $\mathcal{C}^{\text{and}}$ the set of AND interactions and $\mathcal{C}^{\text{or}}$ the set of OR interactions. The setting of ϵ is provided in Appendix A.

> As shown in Figure 1, given the prompt *"when water reaches ... boils and changes into"*, the LLM scores the target token *steam* by relying on learned phrases (interactions), such as $\{$*boils*$\}$, $\{$*boils, into*$\}$, and $\{$*a, boils, it*$\}$. These phrases need not coincide with human expectations. For example, the phrase $\{$*boils, into*$\}$ is encoded as an AND interaction. When *boils* and *into* appear together, the interaction contributes a utility of 0.58 to boost the prediction score of *steam*. The phrase $\{$*boils, a*$\}$ is encoded as an OR interaction. The presence of either word *boils* or *a* contributes a utility of -0.36, lowering the prediction score of *steam*.

Each AND interaction is activated by the binary function $\delta_{\text{and}}\left(\mathbf{x}' \text{ activates AND interaction among input units in } T\right) \in \{0, 1\}$. This function equals 1 only when every unit in T is unmasked in $\mathbf{x}'$ and equals 0 otherwise. *Each OR interaction* is activated by the binary function $\delta_{\text{or}}\left(\mathbf{x}' \text{ activates OR interaction among input units in } T\right) \in \{0, 1\}$. It equals 1 when at least one unit in T is unmasked in $\mathbf{x}'$ and equals 0 when all units in T are masked.

Interaction sparsity. Ren et al. (2024) have established that DNNs rely on only a limited number of AND–OR interactions under three smoothness conditions[3]. Corresponding theoretical analyses show that the number of interactions is bounded, and about 50 to 150 interactions are usually extracted from a single input sample. More details of interaction sparsity are provided in Appendix C.

AND-OR interactions as elementary inference patterns. As the interaction patterns in $\mathcal{C}^{\text{and}}$ and $\mathcal{C}^{\text{or}}$ can faithfully reconstruct the landscape of the function $v(\cdot)$ over numerous (2^n) masked samples, they provide a compact basis for explaining the elementary inference patterns encoded by the DNN. This perspective has been adopted in prior interaction-based explanations (Li & Zhang, 2023; Ren et al., 2024; Zhou et al., 2024). Please see Appendix B for details of extracting interactions.

2.2 Decomposing Shapley Values into AND-OR Interactions

In this subsection, we prove that the Shapley value of an input unit can be written as the sum of the interaction utilities allocated to that input unit, where each interaction utility is divided equally and allocated among all input units participating in the interaction.

[1] For regression problems, $v(\mathbf{x})$ can correspond to the scalar prediction for the target variable such as price, rating, or risk score in a financial regression model. In addition, $v(\mathbf{x})$ can be set to the pre-softmax logit corresponding to the target class.

[2] Masking an input unit refers to replacing it with a baseline value. We define the baseline of an input unit as its average value over different input samples. For NLP models, we follow (Cheng et al., 2024b; Ren et al., 2024) and use a specific embedding to mask input tokens.

[3] Ren et al. (2024) have demonstrated that interaction sparsity can be guaranteed under three common smooth-inference conditions for different DNNs on randomly masked samples; see Appendix C for details.

Shapley value. The Shapley value is one of the most widely used attribution metrics. It is characterized as the unique allocation rule satisfying the *efficiency, symmetry, dummy, and additivity* axioms (see Appendix E for details). Given a DNN v, let $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ denote the input sample with n input units, and let $N = \{1, 2, \ldots, n\}$ denote the index set of these units. The output function of the DNN is $v(\cdot)$. For $S \subseteq N$, let $\mathbf{x}_S$ denote the masked input that retains the units in S and masks[2] the units in $N \setminus S$. The Shapley value of the i-th input unit, ϕ_i[4], is given as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} \left[v(\mathbf{x}_{S\cup\{i\}}) - v(\mathbf{x}_S) \right]. \tag{3}$$

Taking AND-OR interactions as elementary components of the Shapley value. Compared with conventional Harsanyi interactions (Harsanyi, 1963), AND-OR interactions provide a more concise explanation of the inference logic of a DNN (Zhou et al., 2023). In this study, we further establish a close mathematical relationship between these AND-OR interactions and Shapley values.

Specifically, Theorem 1 proves that the Shapley value of an input unit can be expressed as the sum of the interaction utilities allocated to the input unit over all interactions that contain the input unit. Each interaction utility is divided equally among the input units participating in that interaction.

Theorem 1. *Given an AND-OR logical model $d(\cdot)$ for the DNN v satisfying $\forall S \subseteq N, |d(\mathbf{x}_S) - v(\mathbf{x}_S)| \le \epsilon$ in Eq. (2), we define the value ϕ'_i as the sum of the average allocations from all AND and OR interaction utilities that contain the i-th input unit. Then, ϕ'_i approximates the Shapley value ϕ_i with an error of at most 2ϵ:*

$$|\phi'_i - \phi_i| \le 2\epsilon \quad \textit{with} \quad \phi'_i \overset{\text{def}}{=} \sum_{T \in \mathcal{C}^{\text{and}}: T \ni i} \frac{I^{\text{and}}_T}{|T|} + \sum_{T \in \mathcal{C}^{\text{or}}: T \ni i} \frac{I^{\text{or}}_T}{|T|}. \tag{4}$$

2.3 Decomposing Reliable and Unreliable Shapley Components

Defining reliable and unreliable interactions. Once a DNN's output can be represented as the sum of AND-OR interaction utilities, the reliability of the DNN's inference logic can be interpreted as the reliability of compositional interactions in the DNN. Following (Lou et al., 2026), we evaluate the transferability of an interaction extracted by a DNN to another DNN. Interactions that are transferable across DNNs are considered reliable; otherwise, they are considered unreliable.

Definition 1 (Reliable and unreliable interactions)**.** Given an AND interaction S with utility I^{and}_S (or an OR interaction S with utility I^{or}_S) extracted from the target DNN v, we use a new DNN, namely *the reference DNN* v_{ref}, to evaluate the reliability of this interaction. We define an interaction as a *reliable interaction* if it can transfer to the reference DNN, v_{ref}. The reliability of an AND and that of an OR interaction are assessed using the following binary metrics $\mathscr{R}^{\text{and}}_S, \mathscr{R}^{\text{or}}_S \in \{0, 1\}$, respectively:

$$\begin{aligned} \mathscr{R}^{\text{and}}_S &= \mathbf{1}\big(S \in \mathcal{C}^{\text{and}}_{\text{ref}}\big) \cdot \mathbf{1}\big(I^{\text{and}}_S \cdot I^{\text{and}}_{S,\text{ref}} > 0\big), \\ \mathscr{R}^{\text{or}}_S &= \mathbf{1}(S \in \mathcal{C}^{\text{or}}_{\text{ref}}) \cdot \mathbf{1}\big(I^{\text{or}}_S \cdot I^{\text{or}}_{S,\text{ref}} > 0\big), \end{aligned} \tag{5}$$

where the AND interactions extracted from $\mathbf{x}$ by v_{ref} constitute $\mathcal{C}^{\text{and}}_{\text{ref}}$ and the OR interactions extracted from $\mathbf{x}$ by v_{ref} constitute $\mathcal{C}^{\text{or}}_{\text{ref}}$. $I^{\text{and}}_{S,\text{ref}}$ and $I^{\text{or}}_{S,\text{ref}}$ denote the AND and OR utilities, respectively, of interaction S extracted by v_{ref} from $\mathbf{x}$. The indicator function $\mathbf{1}(\cdot)$ returns 1 when its enclosed condition is satisfied, and returns 0 otherwise.

According to the above definition, reliable interactions must be shared by both the target DNN v and the reference DNN v_{ref}. Moreover, the interaction utility used by v_{ref} must have the same sign as that used by the target DNN v, *i.e.*, this interaction must either increase the prediction score or decrease it in both DNNs.

In this way, the reliable and unreliable interactions extracted from the target DNN v are given as two pairs of complementary sets, $(\mathcal{R}^{\text{and}}, \mathcal{U}^{\text{and}})$ and $(\mathcal{R}^{\text{or}}, \mathcal{U}^{\text{or}})$, with $\mathcal{R}^{\text{and}} = \{S \in \mathcal{C}^{\text{and}} : \mathscr{R}^{\text{and}}_S = 1\}$, $\mathcal{R}^{\text{or}} = \{S \in \mathcal{C}^{\text{or}} : \mathscr{R}^{\text{or}}_S = 1\}$, $\mathcal{U}^{\text{and}} = \{S \in \mathcal{C}^{\text{and}} : \mathscr{R}^{\text{and}}_S = 0\}$, and $\mathcal{U}^{\text{or}} = \{S \in \mathcal{C}^{\text{or}} : \mathscr{R}^{\text{or}}_S = 0\}$.

Noise-like behavior of unreliable interactions. Random noise typically lacks a consistent directional utility, so its positive and negative utilities tend to cancel when aggregated. In experiments,

[4] ϕ_i is shorthand for $\phi_i(v)$ when only one model (game) v is considered.

Table 1: Cosine similarities between the reliable Shapley components of two target LLMs, $\mathrm{CosSim}(\boldsymbol{\phi}_1^{\mathcal{R}}, \boldsymbol{\phi}_2^{\mathcal{R}})$, and cosine similarities between the unreliable Shapley components of two target LLMs, $\mathrm{CosSim}(\boldsymbol{\phi}_1^{\mathcal{U}}, \boldsymbol{\phi}_2^{\mathcal{U}})$. Reliable Shapley components exhibit substantially greater similarity.

Target LLMs		finance		WikiText	
v_1	v_2	*Reliable*	*Unreliable*	*Reliable*	*Unreliable*
Llama-3-8B	Gemma-3-12B	**0.75**	-0.03	**0.87**	0.10
Qwen2.5-14B	Qwen3-8B	**0.70**	0.08	**0.85**	-0.14
Llama-7B	Qwen2-7B	**0.72**	-0.02	**0.81**	-0.04
Qwen2.5-1.5B	Qwen2.5-7B	**0.68**	-0.12	**0.78**	0.00
GLM-Z1-9B	Phi-4	**0.82**	-0.09	**0.73**	0.04
Falcon-7B	DeepSeek-R1-8B	**0.75**	-0.10	**0.79**	0.07
Llama-2-7B	Qwen-7B	**0.75**	-0.12	**0.83**	0.02
Qwen2.5-14B	DeepSeek-R1-8B	**0.64**	0.05	**0.77**	0.01
Llama-3-8B	Phi-4	**0.87**	0.01	**0.72**	-0.06
Qwen2.5-7B	Qwen3-8B	**0.81**	-0.15	**0.85**	0.00
Llama-2-7B	Qwen2-7B	**0.74**	-0.12	**0.81**	0.11
Llama-7B	Qwen-7B	**0.75**	-0.02	**0.86**	-0.12
Gemma-3-12B	GLM-Z1-9B	**0.81**	-0.02	**0.83**	-0.11
Qwen2.5-1.5B	Qwen2.5-14B	**0.68**	0.14	**0.79**	-0.02

following (Cheng et al., 2025; Lou et al., 2026), we observe that unreliable interactions exhibit analogous behavior: their positive and negative utilities largely cancel each other, leaving only a small net utility relative to the total utility. By contrast, reliable interactions usually exhibit substantially weaker cancellation in experiments. In this sense, unreliable interactions behave like noise. Please see Appendix I for experiments comparing reliable and unreliable interactions.

Decomposing Shapley values into reliable and unreliable Shapley components. The identification of reliable and unreliable interactions enables us to further decompose Shapley values into reliable Shapley components (*i.e.,* those exclusively determined by reliable interactions) and unreliable Shapley components (*i.e.,* those exclusively determined by unreliable interactions).

Specifically, let $\boldsymbol{\phi}^{\mathcal{R}}(v) = \left[\phi_1^{\mathcal{R}}, \ldots, \phi_n^{\mathcal{R}}\right]^T$ and $\boldsymbol{\phi}^{\mathcal{U}}(v) = \left[\phi_1^{\mathcal{U}}, \ldots, \phi_n^{\mathcal{U}}\right]^T$ denote the vectors of reliable and unreliable Shapley components, respectively. Then, according to Theorem 1, we define the reliable Shapley component $\phi_i^{\mathcal{R}}$ of the i-th input unit as the aggregate of the utilities allocated to the i-th input unit from reliable interactions, and we define the unreliable Shapley component $\phi_i^{\mathcal{U}}$ as the aggregate utilities allocated to the i-th input unit from unreliable interactions.

$$\phi_i^{\mathcal{R}} = \sum_{\text{type}\in\{\text{and,or}\}} \sum_{S\in\mathcal{R}^{\text{type}}:S\ni i} \frac{I_S^{\text{type}}}{|S|}, \phi_i^{\mathcal{U}} = \sum_{\text{type}\in\{\text{and,or}\}} \sum_{S\in\mathcal{U}^{\text{type}}:S\ni i} \frac{I_S^{\text{type}}}{|S|}. \tag{6}$$

Mathematical properties of reliable and unreliable Shapley components. We prove that the reliable and unreliable Shapley components still satisfy the extended versions of the efficiency, symmetry, dummy, and additivity axioms, together with the decomposition property. Because these Shapley components are derived from the logical model $d(\cdot)$, which provides a sufficiently accurate approximation to the original network function $v(\cdot)$ in Eq. (2), the extended axioms differ slightly from their conventional forms. Detailed proofs and discussions are provided in Appendix F.

Theorem 2 (Efficiency)**.** *For the function v with a logical model d, we have $\sum_{i\in N}\left(\phi_i^{\mathcal{R}} + \phi_i^{\mathcal{U}}\right) = d'(\mathbf{x})$ and $\left|\sum_{i\in N}\left(\phi_i^{\mathcal{R}} + \phi_i^{\mathcal{U}}\right) - v'(\mathbf{x})\right| \le 2\epsilon$.*[5]

Theorem 3 (Symmetry)**.** *For two distinct input units $i, j \in N$, if $\forall S \subseteq N \setminus \{i, j\}$, $v(\mathbf{x}_{S\cup\{i\}}) = v(\mathbf{x}_{S\cup\{j\}})$, then we have $\phi_i^{\mathcal{R}} + \phi_i^{\mathcal{U}} = \phi_j^{\mathcal{R}} + \phi_j^{\mathcal{U}}$. Moreover, if the reference model v_{ref} satisfies $\forall S \subseteq N \setminus \{i, j\}$, $v_{\text{ref}}(\mathbf{x}_{S\cup\{i\}}) = v_{\text{ref}}(\mathbf{x}_{S\cup\{j\}})$, then we can obtain $\phi_i^{\mathcal{R}} = \phi_j^{\mathcal{R}}$ and $\phi_i^{\mathcal{U}} = \phi_j^{\mathcal{U}}$.*

Theorem 4 (Dummy)**.** *For an input unit $i \in N$, if $\forall S \subseteq N \setminus \{i\}$, $v(\mathbf{x}_{S\cup\{i\}}) = v(\mathbf{x}_S)$, then we can obtain $\phi_i^{\mathcal{R}} = 0$ and $\phi_i^{\mathcal{U}} = 0$.*

[5]The efficiency axiom is often simplified by setting the empty-coalition value to zero. Accordingly, we define $d'(\mathbf{x}) = d(\mathbf{x}) - d(\mathbf{x}_\emptyset)$ and $v'(\mathbf{x}) = v(\mathbf{x}) - v(\mathbf{x}_\emptyset)$.

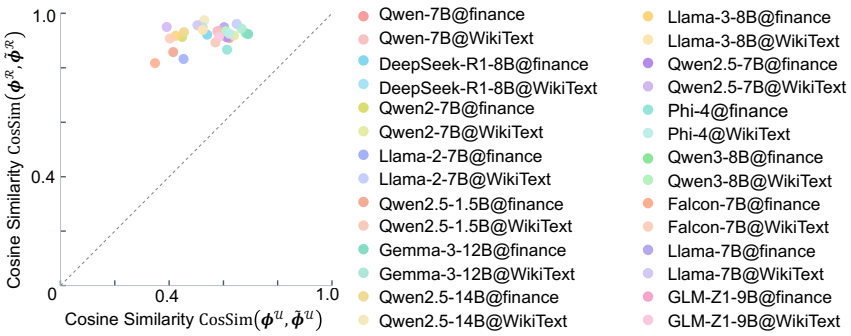


Figure 2: Perturbation stability of the Shapley components, measured by the cosine similarity between the Shapley component extracted from the original input and that extracted from the perturbed input. The similarity measured on reliable Shapley components, $\mathrm{CosSim}(\boldsymbol{\phi}^{\mathcal{R}}, \tilde{\boldsymbol{\phi}}^{\mathcal{R}})$, is substantially higher than that measured on unreliable Shapley components, $\mathrm{CosSim}(\boldsymbol{\phi}^{\mathcal{U}}, \tilde{\boldsymbol{\phi}}^{\mathcal{U}})$.

Theorem 5 (Additivity). *Given two functions v and u, let $v + u$ denote a new function obtained by summing the outputs of the two functions. Then, we have $\forall i \in N$, $\left|\left[\phi_i^{\mathcal{R}}(v+u) + \phi_i^{\mathcal{U}}(v+u)\right] - \left[\phi_i^{\mathcal{R}}(v) + \phi_i^{\mathcal{U}}(v)\right] - \left[\phi_i^{\mathcal{R}}(u) + \phi_i^{\mathcal{U}}(u)\right]\right| \leq 6\epsilon$.*

Theorem 6 (Decomposition). *For each input unit $i \in N$, the reliable and unreliable Shapley components approximate the Shapley value ϕ_i with a bounded error given by $\left|\phi_i^{\mathcal{R}} + \phi_i^{\mathcal{U}} - \phi_i\right| \leq 2\epsilon$.*

3 Experiments

In this section, we compare the representation quality of reliable and unreliable Shapley components from the following perspectives: (1) the similarity of Shapley components across different models; (2) perturbation stability of Shapley components; and (3) the cancellation between positive and negative attributions of Shapley components. We further examine whether the reliable Shapley component concentrates on prediction-relevant input units.

We conduct experiments on 14 LLMs: Phi-4 (Abdin et al., 2024); Gemma-3-12B (Gemma Team et al., 2025); Qwen-7B (Bai et al., 2023); Qwen2-7B (Yang et al., 2024); Qwen2.5-1.5B, Qwen2.5-7B, Qwen2.5-14B (Qwen Team, 2024); Qwen3-8B (Yang et al., 2025); GLM-Z1-9B-0414 (abbreviated as GLM-Z1-9B) (Team GLM et al., 2024); DeepSeek-R1-Distill-Llama-8B (abbreviated as DeepSeek-R1-8B) (DeepSeek-AI, 2025); Llama-7B (Touvron et al., 2023a); Falcon-7B (Almazrouei et al., 2023); Llama-3-8B (Grattafiori et al., 2024); and Llama-2-7B (Touvron et al., 2023b). These LLMs are evaluated on the WikiText dataset (Merity et al., 2016) and the AdaptLLM finance dataset (hereafter, *finance*) (Cheng et al., 2024a). More experimental details are provided in Appendix G.

3.1 Do Different Models Encode Similar Reliable Shapley Components?

In this subsection, we evaluate whether the reliable Shapley components generalize better across different LLMs than unreliable Shapley components, *i.e.*, whether different LLMs usually encode similar reliable Shapley components. Specifically, we examine whether the reliable Shapley components captured by one target LLM are also captured by other target LLMs. Given two target LLMs v_1 and v_2, we compute the cosine similarity between the reliable Shapley components $\boldsymbol{\phi}_1^{\mathcal{R}}$ and $\boldsymbol{\phi}_2^{\mathcal{R}}$, which are extracted by v_1 and v_2, respectively. Similarly, we compute the cosine similarity between the unreliable Shapley components $\boldsymbol{\phi}_1^{\mathcal{U}}$ and $\boldsymbol{\phi}_2^{\mathcal{U}}$ for comparison:

$$\mathrm{CosSim}\big(\boldsymbol{\phi}_1^{\mathcal{R}}, \boldsymbol{\phi}_2^{\mathcal{R}}\big) = \frac{(\boldsymbol{\phi}_1^{\mathcal{R}})^T \boldsymbol{\phi}_2^{\mathcal{R}}}{\|\boldsymbol{\phi}_1^{\mathcal{R}}\|_2 \|\boldsymbol{\phi}_2^{\mathcal{R}}\|_2}, \quad \mathrm{CosSim}\big(\boldsymbol{\phi}_1^{\mathcal{U}}, \boldsymbol{\phi}_2^{\mathcal{U}}\big) = \frac{(\boldsymbol{\phi}_1^{\mathcal{U}})^T \boldsymbol{\phi}_2^{\mathcal{U}}}{\|\boldsymbol{\phi}_1^{\mathcal{U}}\|_2 \|\boldsymbol{\phi}_2^{\mathcal{U}}\|_2}, \tag{7}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm. Notably, the Shapley components $(\boldsymbol{\phi}_1^{\mathcal{R}}, \boldsymbol{\phi}_1^{\mathcal{U}})$ for v_1 and the Shapley components $(\boldsymbol{\phi}_2^{\mathcal{R}}, \boldsymbol{\phi}_2^{\mathcal{U}})$ for v_2 are extracted using two distinct reference LLMs, which avoids inflating the estimated similarity through the use of an identical reference LLM.

Table 2: We report the uncancelled-attribution ratios for the reliable and unreliable Shapley components, $\zeta(\boldsymbol{\phi}^{\mathcal{R}})$ and $\zeta(\boldsymbol{\phi}^{\mathcal{U}})$. The reliable Shapley component's contribution ratio, $\eta^{\mathcal{R}}$, is also reported.

	Uncancelled-attribution ratio $\zeta(\boldsymbol{\psi})$				Contribution ratio $\eta^{\mathcal{R}}$	
	finance		WikiText		finance	WikiText
Model	*Reliable*	*Unreliable*	*Reliable*	*Unreliable*	*Reliable*	*Reliable*
DeepSeek-R1-8B	**72.1%**	43.9%	**76.0%**	48.6%	75.9%	75.4%
Qwen2.5-7B	**74.8%**	44.4%	**82.1%**	52.1%	80.0%	66.7%
Llama-3-8B	**70.6%**	53.3%	**81.5%**	51.2%	76.3%	79.7%
Gemma-3-12B	**70.7%**	50.5%	**79.9%**	53.7%	66.7%	72.9%
Qwen3-8B	**72.6%**	49.3%	**79.5%**	45.2%	76.7%	81.2%
Falcon-7B	**66.7%**	53.0%	**78.7%**	43.0%	74.0%	75.7%
Llama-7B	**75.7%**	60.5%	**79.8%**	45.6%	79.2%	81.5%
Qwen-7B	**69.8%**	47.1%	**79.4%**	44.4%	77.4%	81.4%
Phi-4 (14B)	**71.9%**	47.4%	**75.3%**	50.0%	77.7%	73.0%
Qwen2.5-14B	**74.3%**	50.9%	**77.8%**	49.0%	74.7%	75.7%
Llama-2-7B	**72.3%**	45.3%	**77.9%**	48.8%	74.8%	75.2%
Qwen2-7B	**63.3%**	48.2%	**77.8%**	43.1%	77.7%	81.2%
GLM-Z1-9B	**66.5%**	47.2%	**80.6%**	58.1%	76.0%	81.0%
Qwen2.5-1.5B	**74.4%**	59.3%	**84.7%**	46.7%	67.6%	79.5%

Table 1 reports the cosine similarity between the reliable Shapley components extracted from two target LLMs. For comparison, it also reports the cosine similarity between the unreliable Shapley components. We find that the reliable Shapley components attain significantly higher similarities across different LLMs, with cosine similarities ranging from 0.64 to 0.87, compared with -0.15 to 0.14 for the unreliable Shapley components. These results suggest that the reliable Shapley components represent common attributions shared across different LLMs, whereas the unreliable components reflect model-specific attributions. Note that these similarities are measured between two target DNNs and are distinct from the interaction reliability defined between a target DNN and its corresponding reference DNN.

3.2 Perturbation Stability of Shapley Components

In this subsection, we test the stability of the reliable and unreliable Shapley components under small perturbations in the input embeddings. For each input sample $\mathbf{x}$, let $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]^T$ be its embedding matrix, where $\mathbf{e}_i$ denotes the embedding of the i-th token. We create the perturbed input embedding $\tilde{\mathbf{E}}$ by adding Gaussian noise to each embedding vector, $\tilde{\mathbf{e}}_i = \mathbf{e}_i + \boldsymbol{\delta}_i$, $\boldsymbol{\delta}_i \sim \mathcal{N}\big(\mathbf{0}, 0.25\sigma^2\mathbf{I}\big)$, where σ^2 is the variance computed over all entries of the embedding matrix $\mathbf{E}$. We then extract the reliable and unreliable Shapley components $\boldsymbol{\phi}^{\mathcal{R}}$ and $\boldsymbol{\phi}^{\mathcal{U}}$ from the original input $\mathbf{E}$. We also extract Shapley components $\tilde{\boldsymbol{\phi}}^{\mathcal{R}}$ and $\tilde{\boldsymbol{\phi}}^{\mathcal{U}}$ from the perturbed input $\tilde{\mathbf{E}}$. We quantify perturbation stability by reporting the cosine similarities $\mathrm{CosSim}(\boldsymbol{\phi}^{\mathcal{R}}, \tilde{\boldsymbol{\phi}}^{\mathcal{R}})$ and $\mathrm{CosSim}(\boldsymbol{\phi}^{\mathcal{U}}, \tilde{\boldsymbol{\phi}}^{\mathcal{U}})$, respectively.

Figure 2 compares the perturbation stability of the reliable and unreliable Shapley components. The reliable Shapley components attain higher perturbation stability, with cosine similarity values ranging from 0.82 to 0.97, whereas the unreliable Shapley components exhibit lower perturbation stability, with cosine similarity values ranging from 0.35 to 0.69. These results indicate that the attributions of the unreliable Shapley components are significantly more sensitive to input perturbations than those of the reliable Shapley components.

3.3 Positive-Negative Cancellation of Shapley Components

In this subsection, we examine whether the positive and negative elements cancel each other within the vector of a Shapley component ($\boldsymbol{\phi}^{\mathcal{R}}$ or $\boldsymbol{\phi}^{\mathcal{U}}$). Strong cancellation can cause a Shapley component to make a small net contribution to the LLM's output. Therefore, the reliable Shapley components are expected to exhibit weak cancellation and increase the LLM's output. We use the following metric, $\zeta(\boldsymbol{\psi}) \in [0, 1]$, to quantify the uncancelled-attribution ratio of a Shapley component:

$$\zeta(\boldsymbol{\psi}) = \big|\textstyle\sum_{i\in N} \psi_i\big| \big/ \big(\textstyle\sum_{i\in N} |\psi_i|\big), \qquad \boldsymbol{\psi} \in \big\{\boldsymbol{\phi}^{\mathcal{R}}, \boldsymbol{\phi}^{\mathcal{U}}\big\}. \tag{8}$$

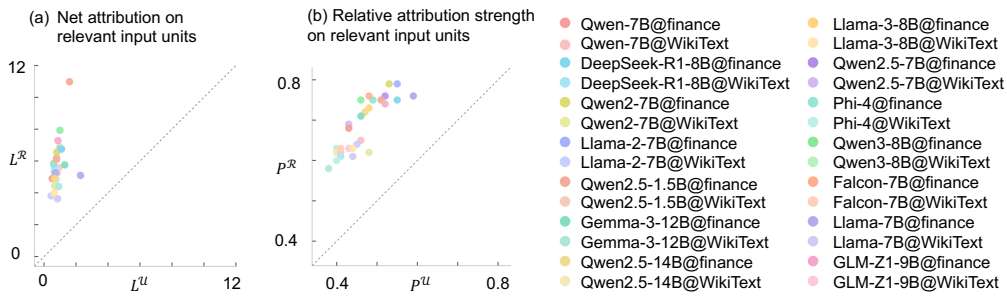


Figure 3: (a) Comparing the reliable Shapley component's net attribution on relevant input units, $L^{\mathcal{R}}$, with the unreliable Shapley component's net attribution on relevant input units, $L^{\mathcal{U}}$. (b) Comparing the reliable Shapley component's relative attribution strength on relevant input units, $P^{\mathcal{R}}$, with the unreliable Shapley component's relative attribution strength on relevant input units, $P^{\mathcal{U}}$.

According to the efficiency axiom, we further quantify the proportion of the LLM's prediction score attributable to the reliable Shapley component, as follows:

$$\eta^{\mathcal{R}} = |\textstyle\sum_{i\in N} \phi_i^{\mathcal{R}}| \big/ \left(|\textstyle\sum_{i\in N} \phi_i^{\mathcal{R}}| + |\textstyle\sum_{i\in N} \phi_i^{\mathcal{U}}|\right). \tag{9}$$

Table 2 compares the uncancelled-attribution ratio of the reliable Shapley component with that of the unreliable Shapley component. The reliable Shapley component has a higher uncancelled-attribution ratio, ranging from 63.3% to 84.7%, whereas the ratio of the unreliable Shapley component ranges from 43.0% to 60.5%. Table 2 also reports the contribution ratios of the reliable Shapley components, which range from 66.7% to 81.5%. These results indicate that the output score of an LLM is primarily driven by contributions from reliable Shapley components.

3.4 Do Reliable Shapley Components Focus on Prediction-Relevant Units?

In this subsection, we further ask whether the reliable Shapley components are more concentrated on the *relevant input units* than the unreliable Shapley components. To this end, we manually categorize all input units into a set of *relevant input units* denoted by $\mathcal{H} \subseteq N$, whose semantic meanings are closely related to the prediction result, and a set of *irrelevant input units* in $N \setminus \mathcal{H}$.

Specifically, we evaluate the semantic relevance of each Shapley component $c \in \{\mathcal{R}, \mathcal{U}\}$ from two perspectives. The first metric measures the net attribution of the Shapley component c on relevant input units as follows:

$$L^c = |\sum_{i\in\mathcal{H}} \phi_i^c|. \tag{10}$$

The second metric is the relative attribution strength on relevant input units, which measures the proportion of the Shapley component's absolute attributions on relevant input units:

$$P^c = \textstyle\sum_{i\in\mathcal{H}} |\phi_i^c| \big/ \sum_{i\in N} |\phi_i^c|. \tag{11}$$

We conduct experiments across the aforementioned LLMs on the finance and WikiText datasets. Figure 3 shows that reliable Shapley components generally yield higher net attribution to relevant input units, with net attribution ($L^{\mathcal{R}}$) ranging from 3.63 to 10.98, compared with the net attribution of unreliable Shapley components ($L^{\mathcal{U}}$) ranging from 0.44 to 2.28. Reliable components also concentrate more on relevant input units, with relative attribution on relevant input units ranging from 0.58 to 0.79 ($P^{\mathcal{R}}$), compared with that of unreliable Shapley components ($P^{\mathcal{U}}$) ranging from 0.38 to 0.59. Table 4 in Appendix H further reports the proportion of test samples in which the reliable Shapley component exceeds the unreliable Shapley component under each metric. Specifically, in 91.2% of the test samples, reliable Shapley components yield greater net attribution to relevant input units than unreliable Shapley components (*i.e.*, $L^{\mathcal{R}} > L^{\mathcal{U}}$). Moreover, in 81.8% of the test samples, reliable Shapley components concentrate more on relevant input units (*i.e.*, $P^{\mathcal{R}} > P^{\mathcal{U}}$).

These results indicate that reliable Shapley components focus more on semantically relevant input units and assign stronger attributions to these input units than unreliable Shapley components.

Case studies. We show reliable and unreliable Shapley components for several input samples.

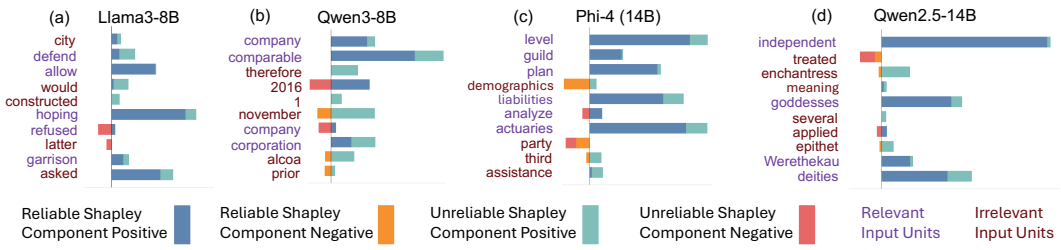


Figure 4: Reliable and unreliable Shapley components for four input samples. (a) Prediction of the token *reinforcements* by Llama-3-8B. (b) Prediction of the token *financial* by Qwen3-8B. (c) Prediction of the token *funding* by Llama-3-8B. (d) Prediction of the token *deity* by Qwen2.5-14B.

> **Llama-3-8B:** target token *reinforcements*.
> **Input prompt:** *Madero asked the commander of the city's garrison to surrender but the latter refused, hoping that the fortifications he had constructed would allow him to defend the city until.*

In Figure 4(a), the reliable Shapley component emphasizes *allow*, *hoping*, *garrison*, and *asked*. Most of these highly attributed input units are relevant input units, including *garrison*, *hoping* and *allow*. By contrast, the unreliable Shapley component assigns small attributions to both irrelevant input units (e.g., *would*, *constructed*, and *latter*) and relevant input units (e.g., *defend* and *hoping*).

> **Qwen3-8B:** target token *financial*.
> **Input prompt:** *prior to the separation of Alcoa Corporation from the company on November 1, 2016, and therefore are not comparable to the company's post-separation.*

In Figure 4(b), the reliable Shapley component focuses on *comparable*, *2016*, *company*, and *corporation*. Most of these highly attributed input units are relevant input units (e.g., *comparable*, *company*, and *corporation*). By contrast, the unreliable Shapley component focuses mainly on irrelevant input units (e.g., *November*, *therefore*, and *Alcoa*).

> **Llama-3-8B:** target token *funding*.
> **Input prompt:** *with the assistance of third party actuaries, we analyze the liabilities and demographics of each plan, which help guide the level of.*

In Figure 4(c), the reliable Shapley component focuses on *level*, *actuaries*, *liabilities*, and *plan*. All of these highly attributed input units are relevant input units. By contrast, the unreliable Shapley component distributes attribution across both relevant input units (e.g., *actuaries* and *liabilities*) and irrelevant input units (e.g., *demographics*, *third*, and *assistance*).

4 Conclusion

In this paper, we explore a new scientific question: whether the attributions of a DNN can be decomposed into relatively reliable and unreliable components. To this end, we prove that the Shapley value of an input unit can be expressed as the sum of the average allocations of all AND-OR interaction utilities involving that unit. Building on this, we decompose each Shapley value into a reliable Shapley component induced by reliable interactions and an unreliable Shapley component induced by unreliable interactions. We prove that the reliable and unreliable Shapley components still satisfy the extended efficiency, symmetry, dummy, and additivity axioms.

We conduct experiments on the finance and WikiText datasets using 14 LLMs. Our results show that (1) different LLMs yield similar reliable Shapley components, whereas unreliable Shapley components vary substantially across LLMs. (2) We also find that reliable Shapley components are more robust to embedding perturbations than unreliable Shapley components. Furthermore, (3) unreliable Shapley components exhibit stronger cancellation between positive and negative attributions, resulting in smaller net contributions to the LLMs' predictions. (4) Compared to unreliable Shapley components, reliable Shapley components concentrate more strongly on input units relevant to the prediction results. Together, these findings demonstrate that the proposed reliable and unreliable Shapley components provide a principled way to assess the reliability of Shapley values.

AI USE STATEMENT

Generative AI tools were used solely to edit the manuscript for grammar, wording, and readability. They were not used to formulate research ideas, hypotheses, mathematical claims, or proofs; to design the methodology or experiments; to implement the proposed method; to process or analyze data; or to interpret results or generate scientific claims. All AI-assisted edits were reviewed and approved by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu et al. Phi-4 technical report. 2024. URL https://arxiv.org/abs/2412.08905.

Ebtesam Almazrouei, Hamza Alobeidli, Asma Alshamsi, Alessandro Cappelli, Ruxandra-Aimée Cojocaru, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of open language models. *ArXiv*, abs/2311.16867, 2023. URL https://api.semanticscholar.org/CorpusID:265466629.

Hongjun An, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, et al. Ai flow: Perspectives, scenarios, and approaches. *Vicinagearth*, 3(1):1, 2026.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin et al. Qwen technical report. 2023. URL https://arxiv.org/abs/2309.16609.

Landon Butler, Abhineet Agarwal, Justin Singh Kang, Yigit Efe Erginbas, Bin Yu, and Kannan Ramchandran. Proxyspex: Inference-efficient interpretability via sparse feature interactions in llms. *ArXiv*, abs/2505.17495, 2025. URL https://api.semanticscholar.org/CorpusID:278886628.

Lu Chen, Siyu Lou, Benhao Huang, and Quanshi Zhang. Defining and extracting generalizable interaction primitives from DNNs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=OCqyFVFNeF.

Lu Chen, Yuxuan Huang, Yixing Li, Dongrui Liu, Qihan Ren, Shuai Zhao, Kun Kuang, Zilong Zheng, and Quanshi Zhang. Can LLMs reason soundly in law? auditing inference patterns for legal judgment. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=5T0BXtJxzN.

Xiangyu Chen, Jixiang Luo, Jingyu Xu, Fangqiu Yi, Chi Zhang, and Xuelong Li. Generative video compression: Towards 0.01% compression rate for video transmission. *Vicinagearth*, 3(1):7, 2026b. doi: 10.1007/s44336-026-00035-2.

Daixuan Cheng, Shaohan Huang, and Furu Wei. Adapting large language models via reading comprehension. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=y886UXPEZ0.

Lei Cheng, Junpeng Zhang, Qihan Ren, and Quanshi Zhang. Revisiting Generalization Power of a DNN in Terms of Symbolic Interactions. *arXiv e-prints*, art. arXiv:2502.10162, February 2025. doi: 10.48550/arXiv.2502.10162.

Xu Cheng, Lei Cheng, Zhaoran Peng, Yang Xu, Tian Han, and Quanshi Zhang. Layerwise change of knowledge in neural networks. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=7zEoinErzQ.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Huiqi Deng, Qihan Ren, Hao Zhang, and Quanshi Zhang. Discovering and explaining the representation bottleneck of dnns. *International Conference on Learning Representations*, 2022.

Yuankai Fan, Qizhen Weng, and Xuelong Li. Computation-bandwidth-memory trade-offs: A unified paradigm for ai infrastructure, 2025. URL https://arxiv.org/abs/2601.11577.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill et al. Gemma 3 technical report. 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

John C. Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963. doi: 10.2307/2525487.

Justin Singh Kang, Yigit Efe Erginbas, Landon Butler, Ramtin Pedarsani, and Kannan Ramchandran. Learning to understand: Identifying interactions via the möbius transform. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 46160–46202. Curran Associates, Inc., 2024. doi: 10.52202/079017-1467. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/520b379123d16e41f85472e766846486-Paper-Conference.pdf.

Justin Singh Kang, Landon Butler, Abhineet Agarwal, Yigit Efe Erginbas, Ramtin Pedarsani, Bin Yu, and Kannan Ramchandran. SPEX: Scaling feature interaction explanations for LLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28878–28903. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kang25a.html.

Mingjie Li and Quanshi Zhang. Does a neural network really encode symbolic concepts? In *International conference on machine learning*, pp. 20452–20469. PMLR, 2023.

Xuelong Li. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6): 8708–8714, 2024. doi: 10.1109/TNNLS.2022.3224577.

Dongrui Liu, Huiqi Deng, Xu Cheng, Qihan Ren, Kangrui Wang, and Quanshi Zhang. Towards the difficulty for a deep neural network to learn concepts of different complexities. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=mZ3hnyL9bS.

Siyu Lou, Yao Yan, Yuntian Chen, and Quanshi Zhang. Cross-llm consistency in inference: Evidence from shared interactions. *ArXiv*, abs/2606.08129, 2026. URL https://api.semanticscholar.org/CorpusID:289032246.

Dakuan Lu, Jiaqi Zhang, Cheng Yuan, Jiawei Shao, and Xuelong Li. The law of multi-model collaboration: Scaling limits of model ensembling for large language models, 2025. URL https://arxiv.org/abs/2512.23340.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *ArXiv*, abs/1609.07843, 2016. URL https://api.semanticscholar.org/CorpusID:16299141.

Qwen Team. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115. Official Qwen2.5 model family release blog: https://qwenlm.github.io/blog/qwen2.5/.

Jie Ren, Die Zhang, Yisen Wang, Lu Chen, Zhanpeng Zhou, Yiting Chen, Xu Cheng, Xin Wang, Meng Zhou, Jie Shi, and Quanshi Zhang. Towards a unified game-theoretic view of adversarial perturbations and robustness. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=fMaIxda5Y6K.

Jie Ren, Mingjie Li, Qirui Chen, Huiqi Deng, and Quanshi Zhang. Defining and quantifying the emergence of sparse concepts in dnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20280–20289, 2023a.

Jie Ren, Zhanpeng Zhou, Qirui Chen, and Quanshi Zhang. Can we faithfully represent absence states to compute shapley values on a DNN? In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=YV8tP7bW6Kt.

Qihan Ren, Huiqi Deng, Yunuo Chen, Siyu Lou, and Quanshi Zhang. Bayesian neural networks avoid encoding complex and perturbation-sensitive concepts. In *International Conference on Machine Learning*, pp. 28889–28913. PMLR, 2023c.

Qihan Ren, Jiayang Gao, Wen Shen, and Quanshi Zhang. Where we have arrived in proving the emergence of sparse symbolic concepts in ai models. *International Conference on Learning Representations*, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Jiawei Shao and Xuelong Li. Ai flow at the network edge. *IEEE Network*, 40(1):330–336, 2025. doi: 10.1109/MNET.2025.3541208.

Lloyd S Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, Princeton, 1953.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Huan Song, Qingfei Zhao, Ting Long, Shuyu Tian, Hongjun An, Jiawei Shao, and Xuelong Li. Theoretical foundations of scaling law in familial models, 2025. URL https://arxiv.org/abs/2512.23407.

Erik Štrumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(1):1–18, 2010. URL http://jmlr.org/papers/v11/strumbelj10a.html.

SymTrust AI. SymTrust AI, 2026. URL https://symtrustai.com/en.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. 2024. URL https://arxiv.org/abs/2406.12793.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023a. URL https://api.semanticscholar.org/CorpusID:257219404.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen et al. Llama 2: Open foundation and fine-tuned chat models. 2023b. URL https://arxiv.org/abs/2307.09288.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang et al. Qwen2 technical report. 2024. URL https://arxiv.org/abs/2407.10671.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Cheng Yuan, Jiawei Shao, and Xuelong Li. Information capacity: Evaluating the efficiency of large language models via text compression, 2025. URL https://arxiv.org/abs/2511.08066.

Cheng Yuan, Zhenyu Jia, Jiawei Shao, and Xuelong Li. Enhancing neural video compression of static scenes with positive-incentive noise, 2026. URL https://arxiv.org/abs/2603.06095.

Hao Zhang, Sen Li, Yinchao Ma, Mingjie Li, Yichen Xie, and Quanshi Zhang. Interpreting and boosting dropout from a game-theoretic view. *International Conference on Learning Representations*, 2021.

Hongyuan Zhang, Sida Huang, Yubin Guo, and Xuelong Li. Variational positive-incentive noise: How noise benefits models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8313–8320, 2025. doi: 10.1109/TPAMI.2025.3575295.

Huilin Zhou, Huijie Tang, Mingjie Li, Hao Zhang, Zhenyu Liu, and Quanshi Zhang. Explaining how a neural network play the go game and let people learn. *arXiv preprint arXiv:2310.09838*, 2023.

Huilin Zhou, Hao Zhang, Huiqi Deng, Dongrui Liu, Wen Shen, Shih-Han Chan, and Quanshi Zhang. Explaining generalization power of a dnn using interactive concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17105–17113, 2024.

APPENDIX

A UNIVERSAL MATCHING PROPERTY

The universal matching property of the AND-OR logical model is proven by Chen et al.. Let $\mathcal{X} = \{\mathbf{x}_S \mid S \subseteq N\}$ contain all 2^n masked versions of an input sample $\mathbf{x}$, where $\mathbf{x}_S$ retains the input units in S and masks the units in $N \setminus S$. Masking substitutes an input unit with its baseline value; for language models, we use the masking embedding scheme in Cheng et al. (2024b); Ren et al. (2024).

Theorem 7 (Universal matching property). *Given a DNN v and an input $\mathbf{x}$, let*

$$I_T^{\text{and}} = \sum_{L \subseteq T} (-1)^{|T|-|L|} v_L^{\text{and}},$$
$$I_T^{\text{or}} = -\sum_{L \subseteq T} (-1)^{|T|-|L|} v_{N \setminus L}^{\text{or}},$$

where $v_L^{\text{and}} = 0.5v(\mathbf{x}_L) + \gamma_L$ and $v_L^{\text{or}} = 0.5v(\mathbf{x}_L) - \gamma_L$, and let $b = v(\mathbf{x}_\varnothing)$. Then the resulting logical model d satisfies

$$\forall \mathbf{x}' \in \mathcal{X}, \qquad d(\mathbf{x}') = v(\mathbf{x}'). \tag{12}$$

This result establishes exact fidelity before procedure in section B. In practice, we remove small output noise to obtain a sparse representation. We write

$$v(\mathbf{x}_S) = z_S^{\text{and}} + z_S^{\text{or}} + \epsilon_S,$$

where $z_S^{\text{and}} = 0.5(v(\mathbf{x}_S) - \epsilon_S) + \gamma_S$ and $z_S^{\text{or}} = 0.5(v(\mathbf{x}_S) - \epsilon_S) - \gamma_S$. The parameters $\{\gamma_S\}$ and $\{\epsilon_S\}$ are optimized in a LASSO-like objective, with $\epsilon_S \in [-\rho, \rho]$ and $\rho = 0.01|v(\mathbf{x}) - v(\mathbf{x}_\varnothing)|$. This small bounded correction suppresses numerical noise while preserving the model-output structure used to extract interactions. Thus, we have $\epsilon \geq \rho$.

B INTERACTION EXTRACTION PROCEDURE AND COMPUTATIONAL DETAILS

For an input with n input units, exhaustive extraction evaluates 2^n masked inputs. We therefore follow prior interaction studies (Li & Zhang, 2023; Liu et al., 2023; Ren et al., 2023b) and analyze ten semantically meaningful input units at a time. For text, stop words, punctuation, and non-alphanumeric symbols are removed before selecting ten content-bearing words or phrases. First, evaluate $v(\mathbf{x}_S)$ for every $S \subseteq N$. Second, optimize $\{\gamma_S, \epsilon_S\}$ to minimize

$$\|\mathbf{I}_{\text{and}}\|_1 + \|\mathbf{I}_{\text{or}}\|_1,$$

where $\mathbf{I}_{\text{and}}, \mathbf{I}_{\text{or}} \in \mathbb{R}^{2^n}$ vectorize all AND and OR utilities. For each model–dataset setting, we fix the salience threshold before extraction as the mean 2% full-input/empty-input output gap over its evaluation samples $\mathcal{D}$,

$$\tau = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} 0.02 \left| v(\mathbf{x}_N) - v(\mathbf{x}_\varnothing) \right|.$$

Third, set the bias to $b = v(\mathbf{x}_\varnothing)$ and define $\mathcal{C}^{\text{and}}$ and $\mathcal{C}^{\text{or}}$ as the AND and OR interactions whose absolute utilities exceed this fixed τ.

Scalable interaction extraction and practical use. Exactly computing interaction can become prohibitively costly as the number of input units increases. Prior work has addressed this scalability challenge from several directions, including approximate interaction estimators (Kang et al., 2024; 2025; Butler et al., 2025) and procedures that reduce the cost of extracting interactions (Li & Zhang, 2023; Ren et al., 2023b); Appendix B provides further details. Such computational advances have also supported the use of interaction-based explanations in practical systems such as SymTrust (SymTrust AI, 2026).

The LLM experiments were run on NVIDIA A800 GPUs. Extracting the interaction representation of one input required approximately 55 seconds; evaluating cross-model reliability for 100 samples required 12–16 seconds after interaction utilities had been extracted.

Algorithm 1 Extract Interactions

Input: Trained DNN $v(\cdot)$; input sample $\mathbf{x}$ with input units indexed by $N = \{1, \ldots, n\}$.
Constants assumed: fixed model–dataset threshold $\tau = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} 0.02 |v(\mathbf{x}_N) - v(\mathbf{x}_\varnothing)|$.
Output: Interaction utilities $I_S^{\text{and}}, I_S^{\text{or}}$ for all $S \subseteq N$; bias b; salient interaction sets $\mathcal{C}^{\text{and}}, \mathcal{C}^{\text{or}}$.
for each $S \subseteq N$ **do**
    Generate masked input $\mathbf{x}_S$ and the corresponding model output $v(\mathbf{x}_S)$.
end for
Determine $I_S^{\text{and}}, I_S^{\text{or}}$ for all $S \subseteq N$ by solving $\min_{\{\epsilon_S, \gamma_S\}} \sum_{S \subseteq N} \left(|I_S^{\text{and}}| + |I_S^{\text{or}}| \right)$.
Determine the bias term b as the network's output for the empty set of input units: $b \leftarrow v(\emptyset)$.
$\mathcal{C}^{\text{and}} \leftarrow \{S \subseteq N : \ |I_S^{\text{and}}| > \tau\}$.
$\mathcal{C}^{\text{or}} \leftarrow \{S \subseteq N : \ |I_S^{\text{or}}| > \tau\}$.
return $I_S^{\text{and}}, I_S^{\text{or}}$, bias b, and sets $\mathcal{C}^{\text{and}}, \mathcal{C}^{\text{or}}$.

C Interaction Sparsity

Ren et al. have proven that under the three following conditions, the inference logic of a DNN can be represented by a small set of interactions with non-negligible utility, while the other overwhelming majority contribute almost nothing. Their analysis rests on the following three conditions.

Here, sparsity means that most interaction utilities have magnitudes close to zero, which allows the model output v to be accurately approximated by a small subset of salient nonzero interactions, while satisfying the approximation-error bound ϵ. If no near-zero interactions are discarded and all interaction terms are used for fitting, then the neural network output can be exactly recovered (see Appendix A for details).

Bounded interaction order. There is an $M < n$ for which interactions involving more than M input units have zero effect. For a sufficiently smooth output function, a stronger sufficient condition is that every mixed derivative whose total order exceeds M vanishes. Thus, the model does not require arbitrarily high-order dependencies to describe its response around the masking baseline.

Average monotonicity under unmasking. Retaining more input units should not reduce the average output:

For a fixed input, define the average output above the empty-input baseline at masking level m as

$$\bar{u}(m) = \mathbb{E}_{S \subseteq N : |S| = m} \left[v(\mathbf{x}_S) - v(\mathbf{x}_\varnothing) \right].$$

for $m' \leq m$,

$$\bar{u}(m') \leq \bar{u}(m).$$

This captures the usual behavior of a model that remains meaningful on partially masked inputs and gains evidence as relevant content is restored. It excludes oscillatory responses, such as a dependence determined only by whether the number of retained units is odd or even.

Controlled degradation under masking. The average output should decline gradually rather than collapse when units are hidden. More precisely, for some $p > 0$ and any $m' \leq m$,

$$\bar{u}(m') \geq \left(\frac{m'}{m} \right)^p \bar{u}(m).$$

The exponent p quantifies the permitted sensitivity to occlusion: a moderate value rules out a model whose score is sustained only by a large collection of fragile, high-order effects.

Under these conditions, the aggregate utility at each interaction order is polynomially bounded in n. Consequently, for a salience threshold τ, the number of retained order-k interactions scales polynomially (up to the degree-k cancellation factor) rather than combinatorially with $\binom{n}{k}$. Except in the exceptional case of exponentially precise cancellation between positive and negative utilities, this yields far fewer salient interactions than the 2^n possible subsets.

D Related Work

Attribution methods. Attribution methods seek to explain a DNN's prediction by quantifying how much each input unit contributes to the model output. Prior approaches differ in the evidence they use to assign such importance. Gradient-based techniques derive importance from the model's local sensitivity to each input (Simonyan et al., 2013; Ancona et al., 2017), whereas perturbation-based techniques measure how the prediction changes when selected inputs are modified or removed (Ribeiro et al., 2016; Fong & Vedaldi, 2017). Game-theoretic formulations instead regard input units as players and allocate the prediction according to their marginal effects across different coalitions (Štrumbelj & Kononenko, 2010; Lundberg & Lee, 2017). A prominent instance is the Shapley value (Shapley, 1953; Lundberg & Lee, 2017), whose allocation is uniquely determined by the requirements of efficiency, symmetry, dummy, and additivity.

Interaction-based explanation. *From model outputs to sparse interactions.* A large number of studies (Deng et al., 2022; Li & Zhang, 2023; Ren et al., 2023a; 2024) suggest that the inference process of a DNN can be replicated by a collection of AND–OR interactions among input variables. Ren et al. (2023a) represent these interactions using a sparse graphical model and further compress this representation into an AND–OR graph. Building on this formulation, Ren et al. (2024) establish sufficient conditions under which only a small number of interactions are needed to reproduce the model's prediction scores on randomly masked samples. Empirically, Li & Zhang (2023) show that interactions extracted from well-trained networks tend to be sparse, discriminative, and transferable across samples. Chen et al. (2024) further introduce an approach for identifying generalizable interaction primitives shared by networks. Collectively, these studies establish a theoretical foundation for leveraging interactions to explain DNN behavior. *Interactions as probes of model representation.* Beyond reproducing predictions, interaction statistics provide quantitative diagnostics of the representations learned by DNNs. Extending the analysis to intermediate representations, Cheng et al. (2024b) show how interactions emerge and disappear across layers. At the level of overall representation, Deng et al. (2022) reveal a bottleneck in which networks favor very low- and very high-order interactions. For mean-field variational Bayesian neural networks, Ren et al. (2023c) prove and empirically verify a weaker tendency to encode complex interactions than that observed in standard DNNs. Regarding training mechanisms, Zhang et al. (2021) show that dropout suppresses interaction strength in a manner associated with reduced overfitting. In adversarial settings, attacks have been found to preferentially alter high-order interactions, whereas the robustness acquired through adversarial training has been attributed mainly to category-specific low-order interactions (Ren et al., 2021). Generalization-focused analyses further report that better-generalizing networks learn simple interactions earlier and encode fewer complex ones (Zhou et al., 2024), while Liu et al. (2023) explain why interactions involving more variables are intrinsically more difficult for a DNN to learn. Finally, Chen et al. (2026a) apply interaction-based explanations to legal LLMs, showing that seemingly correct judgments may still be supported by misleading or irrelevant inference patterns.

AI Flow. AI Flow supports distributed inference by coordinating models and resources across devices, edge nodes, and cloud servers (Shao & Li, 2025; An et al., 2026), shifting communication from raw-data delivery to task-oriented intelligence. Recent implementations explore generative reconstruction and scene-specific priors to reduce video-transmission overhead (Chen et al., 2026b; Yuan et al., 2026). Fan et al. (2025) view compute, communication capacity, and storage as interdependent resources for efficient AI infrastructure. Related studies (Li, 2024; Zhang et al., 2025) investigate beneficial noise that improves model representations. Information Capacity evaluates LLM efficiency through compression gains with respect to inference cost (Yuan et al., 2025). Song et al. (2025) investigate the scaling behavior of familial models. Lu et al. (2025) investigate the scaling behavior of collaborative models. As AI systems become increasingly capable and distributed, understanding the internal representations that support models' behavior is also becoming essential; interaction-based analysis offers a principled way to characterize the representations of AI models.

E Four Axioms of the Shapley Value

For a function $v : 2^N \to \mathbb{R}$, the Shapley value $\phi_i(v)$ is the unique allocation rule satisfying the following four axioms. These properties specify how the total function utility is distributed among input units.

Efficiency. The attributions recover the total utility above the empty-coalition baseline:

$$\sum_{i \in N} \phi_i(v) = v(N) - v(\varnothing). \tag{13}$$

Symmetry. If two input units have the same marginal contribution to every coalition that excludes them, then they receive the same attribution:

$$[\forall S \subseteq N \setminus \{i, j\},\ v(S \cup \{i\}) = v(S \cup \{j\})] \Longrightarrow \phi_i(v) = \phi_j(v). \tag{14}$$

Dummy. An input unit that never changes the function output receives zero attribution:

$$[\forall S \subseteq N \setminus \{i\},\ v(S \cup \{i\}) = v(S)] \Longrightarrow \phi_i(v) = 0. \tag{15}$$

Additivity. For any two functions u and v, the attribution of their sum equals the sum of their attributions:

$$\phi_i(u + v) = \phi_i(u) + \phi_i(v), \qquad i \in N. \tag{16}$$

F Proofs of Theoretical Results

F.1 Statement

Let $N = \{1, \ldots, n\}$ be the set of input units, and let $v : 2^N \to \mathbb{R}$ be the model output function. The reference-model output function on the same masked samples is denoted by $v_{\text{ref}} : 2^N \to \mathbb{R}$; it is used to determine whether an interaction extracted from v is reliable or unreliable. Let d denote the AND–OR logical model of v. We retain $\phi_i^{\mathcal{R}}(v)$ and $\phi_i^{\mathcal{U}}(v)$ for the two components reported for the original model.

For convenience, all games are first zero-normalized. We continue to use the same symbols for the zero-normalized games, so $v(\varnothing) = 0$ throughout and the empty-coalition baseline can be omitted. The interaction-extraction procedure constructs candidate decompositions of an approximation to this function. A candidate residual $e : 2^N \to \mathbb{R}$ and a candidate split vector $\boldsymbol{\gamma} : 2^N \to \mathbb{R}$ satisfy $e(\varnothing) = \boldsymbol{\gamma}(\varnothing) = 0$ and $|e(S)| \le \epsilon$ for every $S \subseteq N$. They induce the two candidate component games

$$v_{e,\boldsymbol{\gamma}}^{\text{and}}(S) = \tfrac{1}{2}\big(v(S) - e(S)\big) + \boldsymbol{\gamma}(S), \qquad v_{e,\boldsymbol{\gamma}}^{\text{or}}(S) = \tfrac{1}{2}\big(v(S) - e(S)\big) - \boldsymbol{\gamma}(S). \tag{17}$$

Thus $v_{e,\boldsymbol{\gamma}}^{\text{and}} + v_{e,\boldsymbol{\gamma}}^{\text{or}} = v - e$. The interactions of a candidate pair are defined from its component games by

$$I_{T,e,\boldsymbol{\gamma}}^{\text{and}} = \sum_{L \subseteq T} (-1)^{|T|-|L|} v_{e,\boldsymbol{\gamma}}^{\text{and}}(L), \tag{18}$$

$$I_{T,e,\boldsymbol{\gamma}}^{\text{or}} = -\sum_{L \subseteq T} (-1)^{|T|-|L|} v_{e,\boldsymbol{\gamma}}^{\text{or}}(N \setminus L), \tag{19}$$

for every nonempty $T \subseteq N$. The sparse extraction objective is the total absolute interaction strength:

$$\mathcal{L}_N(e, \boldsymbol{\gamma}; v) = \sum_{\varnothing \neq T \subseteq N} \left| I_{T,e,\boldsymbol{\gamma}}^{\text{and}} \right| + \sum_{\varnothing \neq T \subseteq N} \left| I_{T,e,\boldsymbol{\gamma}}^{\text{or}} \right|. \tag{20}$$

The sparse objective may admit multiple global minimizers. We fix a global minimizer $(e_v^\star, \boldsymbol{\gamma}_v^\star)$ of this objective over the feasible set above. For the symmetry and dummy results, this minimizer is chosen to preserve the corresponding symmetry or dummy structure; Propositions 17 and 18 show that such a choice is always possible.

$$\begin{aligned} v^{\text{and}} &:= v_{e_v^\star, \boldsymbol{\gamma}_v^\star}^{\text{and}}, \qquad v^{\text{or}} := v_{e_v^\star, \boldsymbol{\gamma}_v^\star}^{\text{or}}, \\ I_{T,v}^{\text{type}} &:= I_{T,e_v^\star, \boldsymbol{\gamma}_v^\star}^{\text{type}}, \quad \text{type} \in \{\text{and}, \text{or}\}. \end{aligned} \tag{21}$$

For each $\text{type} \in \{\text{and}, \text{or}\}$, let $\mathcal{C}_v^{\text{type}}$ denote the interaction set. The reliable and unreliable interaction sets partition this set:

$$\mathcal{C}_v^{\text{type}} = \mathcal{R}_v^{\text{type}} \dot{\cup} \mathcal{U}_v^{\text{type}}, \qquad \text{type} \in \{\text{and}, \text{or}\}. \tag{22}$$

For use in the proofs, the reliable and unreliable Shapley components are defined by

$$\phi_i^{\mathcal{R}}(v) = \sum_{\text{type}\in\{\text{and},\text{or}\}} \sum_{T\in\mathcal{R}_v^{\text{type}}:T\ni i} \frac{I_{T,v}^{\text{type}}}{|T|}, \tag{23}$$

$$\phi_i^{\mathcal{U}}(v) = \sum_{\text{type}\in\{\text{and},\text{or}\}} \sum_{T\in\mathcal{U}_v^{\text{type}}:T\ni i} \frac{I_{T,v}^{\text{type}}}{|T|}. \tag{24}$$

Theorem 8 (AND-OR interaction allocation)**.** *Given an AND-OR logical model $d(\cdot)$ for the DNN v satisfying $\forall S\subseteq N$, $|d(\mathbf{x}_S)-v(\mathbf{x}_S)|\le\epsilon$, we define the value ϕ_i' as the sum of the average allocations from all AND and OR interaction utilities that contain the i-th input unit. Then, ϕ_i' approximates the Shapley value ϕ_i with an error of at most 2ϵ:*

$$|\phi_i'-\phi_i|\le 2\epsilon \quad \textit{with} \quad \phi_i' \overset{\text{def}}{=} \sum_{T\in\mathcal{C}_{\text{and}}:T\ni i} \frac{I_T^{\text{and}}}{|T|} + \sum_{T\in\mathcal{C}_{\text{or}}:T\ni i} \frac{I_T^{\text{or}}}{|T|}. \tag{25}$$

Theorem 9 (Decomposition)**.** *For each input unit $i\in N$, the reliable and unreliable Shapley components approximate the Shapley value ϕ_i with a bounded error given by $\left|\phi_i^{\mathcal{R}}+\phi_i^{\mathcal{U}}-\phi_i\right|\le 2\epsilon$.*

Relation to the conventional Shapley axioms. The following four axioms retain the same underlying principles as their conventional counterparts, but are stated for the decomposed attribution. Efficiency and additivity apply to the sum $\phi_i^{\mathcal{R}}+\phi_i^{\mathcal{U}}$. Symmetry first applies to this sum and, when the reference model has the same symmetry, further applies to each component. The dummy axiom requires both components to vanish. Thus, the differences concern the form of the allocation rather than the principles expressed by the four axioms.

Theorem 10 (Efficiency)**.** *For the function v with a logical model d, we have $\sum_{i\in N}\left(\phi_i^{\mathcal{R}}+\phi_i^{\mathcal{U}}\right)=d(\mathbf{x})-d(\mathbf{x}_\emptyset)$ and $\left|\sum_{i\in N}\left(\phi_i^{\mathcal{R}}+\phi_i^{\mathcal{U}}\right)+v(\mathbf{x}_\emptyset)-v(\mathbf{x})\right|\le 2\epsilon$.*

Theorem 11 (Symmetry)**.** *For two distinct input units $i,j\in N$, if $\forall S\subseteq N\setminus\{i,j\}$, $v(\mathbf{x}_{S\cup\{i\}})=v(\mathbf{x}_{S\cup\{j\}})$, then we have $\phi_i^{\mathcal{R}}+\phi_i^{\mathcal{U}}=\phi_j^{\mathcal{R}}+\phi_j^{\mathcal{U}}$. Moreover, if the reference model v_{ref} satisfies $\forall S\subseteq N\setminus\{i,j\}$, $v_{\text{ref}}(\mathbf{x}_{S\cup\{i\}})=v_{\text{ref}}(\mathbf{x}_{S\cup\{j\}})$, then we can obtain $\phi_i^{\mathcal{R}}=\phi_j^{\mathcal{R}}$ and $\phi_i^{\mathcal{U}}=\phi_j^{\mathcal{U}}$.*

Theorem 12 (Dummy)**.** *For an input unit $i\in N$, if $\forall S\subseteq N\setminus\{i\}$, $v(\mathbf{x}_{S\cup\{i\}})=v(\mathbf{x}_S)$, then we can obtain $\phi_i^{\mathcal{R}}=0$ and $\phi_i^{\mathcal{U}}=0$.*

Theorem 13 (Additivity)**.** *Given two functions v and u, let $v+u$ denote a new function obtained by summing the outputs of the two functions. Then, we have $\forall i\in N$,*

$$\left|\phi_i^{\mathcal{R}}(v+u)+\phi_i^{\mathcal{U}}(v+u)-\phi_i^{\mathcal{R}}(v)-\phi_i^{\mathcal{U}}(v)-\phi_i^{\mathcal{R}}(u)-\phi_i^{\mathcal{U}}(u)\right|\le 6\epsilon. \tag{26}$$

F.2 Proof

F.3 Shapley values of AND and OR basis games

For a game $v:2^N\to\mathbb{R}$, the Shapley value of player i can be written in random-permutation form as

$$\phi_i(v)=\frac{1}{n!}\sum_{\pi\in\Pi(N)}\left[v(P_i^\pi\cup\{i\})-v(P_i^\pi)\right], \tag{27}$$

where $\Pi(N)$ is the set of all permutations of N, and P_i^π is the set of players appearing before i in permutation π.

Lemma 14 (AND basis)**.** *For a nonempty $T\subseteq N$, define $u_T^{\text{and}}(S)=\mathbf{1}\{T\subseteq S\}$. Then*

$$\phi_i(u_T^{\text{and}})=\begin{cases}1/|T|, & i\in T,\\ 0, & i\notin T.\end{cases} \tag{28}$$

Proof. If $i\notin T$, adding i cannot change whether $T\subseteq S$, so the marginal contribution is zero for every coalition. If $i\in T$, the marginal contribution in equation 27 equals one exactly when all other members of T appear before i, namely when i is the last member of T in a uniformly random permutation. Each member of T is equally likely to be last, so the probability is $1/|T|$. □

Lemma 15 (OR basis)**.** *For a nonempty $T \subseteq N$, define $u_T^{\mathrm{or}}(S) = \mathbf{1}\{T \cap S \neq \varnothing\}$. Then*

$$\phi_i(u_T^{\mathrm{or}}) = \begin{cases} 1/|T|, & i \in T, \\ 0, & i \notin T. \end{cases} \tag{29}$$

Proof. If $i \notin T$, adding i cannot change whether S intersects T. If $i \in T$, the marginal contribution equals one exactly when no member of T appears before i, namely when i is the first member of T in a uniformly random permutation. This event has probability $1/|T|$. □

F.4 Proof of the AND-OR interaction allocation theorem

Proof of Theorem 8. In the notation of the theorem, take $\mathcal{C}_{\mathrm{and}} = \mathcal{C}_v^{\mathrm{and}}$ and $\mathcal{C}_{\mathrm{or}} = \mathcal{C}_v^{\mathrm{or}}$. The Shapley operator is linear:

$$\begin{aligned} \phi_i(av + bh) &= \frac{1}{n!} \sum_{\pi \in \Pi(N)} \big[(av + bh)(P_i^\pi \cup \{i\}) \\ &\qquad - (av + bh)(P_i^\pi)\big] \\ &= a\phi_i(v) + b\phi_i(h). \end{aligned} \tag{30}$$

Write $u_T^{\mathrm{and}}(S) = \mathbf{1}\{T \subseteq S\}$ and $u_T^{\mathrm{or}}(S) = \mathbf{1}\{T \cap S \neq \varnothing\}$. Then the assumed AND-OR representation can be written as

$$d = \sum_{T \in \mathcal{C}_v^{\mathrm{and}}} I_{T,v}^{\mathrm{and}} u_T^{\mathrm{and}} + \sum_{T \in \mathcal{C}_v^{\mathrm{or}}} I_{T,v}^{\mathrm{or}} u_T^{\mathrm{or}}. \tag{31}$$

Since N is finite, repeated application of equation 30 gives

$$\begin{aligned} \phi_i(d) &= \sum_{T \in \mathcal{C}_v^{\mathrm{and}}} I_{T,v}^{\mathrm{and}} \phi_i(u_T^{\mathrm{and}}) + \sum_{T \in \mathcal{C}_v^{\mathrm{or}}} I_{T,v}^{\mathrm{or}} \phi_i(u_T^{\mathrm{or}}) \\ &= \sum_{T \in \mathcal{C}_v^{\mathrm{and}} : T \ni i} \frac{I_{T,v}^{\mathrm{and}}}{|T|} + \sum_{T \in \mathcal{C}_v^{\mathrm{or}} : T \ni i} \frac{I_{T,v}^{\mathrm{or}}}{|T|}, \end{aligned} \tag{32}$$

where the second equality uses Lemmas 14 and 15. Hence

$$\begin{aligned} \phi_i(d) &= \sum_{T \in \mathcal{C}_v^{\mathrm{and}} : T \ni i} \frac{I_{T,v}^{\mathrm{and}}}{|T|} + \sum_{T \in \mathcal{C}_v^{\mathrm{or}} : T \ni i} \frac{I_{T,v}^{\mathrm{or}}}{|T|} \\ &= \sum_{\mathrm{type} \in \{\mathrm{and},\mathrm{or}\}} \sum_{T \in \mathcal{C}_v^{\mathrm{type}} : T \ni i} \frac{I_{T,v}^{\mathrm{type}}}{|T|}, \end{aligned} \tag{33}$$

which is exactly the allocation identity in Eq. (25) under the identification above. The right-hand side is precisely the quantity denoted by ϕ_i' in the theorem statement. Therefore,

$$\phi_i' = \phi_i(d). \tag{34}$$

Combining this equality with the uniform-error argument in Section F.6 gives $|\phi_i' - \phi_i| = |\phi_i(d) - \phi_i| \leq 2\epsilon$. This proves Theorem 8. □

F.5 Decomposition of the Shapley value

Starting directly from the definitions in equation 23 and equation 24, we have

$$\begin{aligned} \phi_i^{\mathcal{R}}(v) + \phi_i^{\mathcal{U}}(v) &= \sum_{\mathrm{type} \in \{\mathrm{and},\mathrm{or}\}} \sum_{T \in \mathcal{R}_v^{\mathrm{type}} : T \ni i} \frac{I_{T,v}^{\mathrm{type}}}{|T|} \\ &\quad + \sum_{\mathrm{type} \in \{\mathrm{and},\mathrm{or}\}} \sum_{T \in \mathcal{U}_v^{\mathrm{type}} : T \ni i} \frac{I_{T,v}^{\mathrm{type}}}{|T|} \\ &= \sum_{\mathrm{type} \in \{\mathrm{and},\mathrm{or}\}} \sum_{T \in \mathcal{C}_v^{\mathrm{type}} : T \ni i} \frac{I_{T,v}^{\mathrm{type}}}{|T|} \\ &= \phi_i(d). \end{aligned} \tag{35}$$

Here the second equality follows because every retained interaction is classified into exactly one of the reliable and unreliable sets, and the last equality follows from the allocation identity equation 32. This proves the exact decomposition asserted in Theorem 9.

F.6 Error bound induced by universal matching

We now relate the exact decomposition of the logical model d to the Shapley value of the original game v. Define its approximation residual by $r_v(S) = d(\mathbf{x}_S) - v(\mathbf{x}_S)$. By the uniform approximation property of d, for every $S \subseteq N$,

$$d(\mathbf{x}_S) - v(\mathbf{x}_S) = r_v(S), \qquad |r_v(S)| \le \epsilon. \tag{36}$$

Let

$$w_S = \frac{|S|!(n-|S|-1)!}{n!}, \qquad S \subseteq N \setminus \{i\}, \tag{37}$$

denote the Shapley weight. The weights are nonnegative and sum to one, because

$$\begin{aligned} \sum_{S \subseteq N\setminus\{i\}} w_S &= \sum_{k=0}^{n-1} \binom{n-1}{k} \frac{k!(n-k-1)!}{n!} \\ &= \sum_{k=0}^{n-1} \frac{1}{n} = 1. \end{aligned} \tag{38}$$

Using the exact interaction decomposition $\phi_i^{\mathcal{R}} + \phi_i^{\mathcal{U}} = \phi_i(d)$ and the definition of the Shapley value, we obtain

$$\begin{aligned} &\left|\phi_i^{\mathcal{R}} + \phi_i^{\mathcal{U}} - \phi_i(v)\right| \\ &= |\phi_i(d) - \phi_i(v)| \\ &= \left| \sum_{S \subseteq N\setminus\{i\}} w_S \left[r_v(S \cup \{i\}) - r_v(S) \right] \right| \\ &\le \sum_{S \subseteq N\setminus\{i\}} w_S \left(|r_v(S \cup \{i\})| + |r_v(S)| \right) \\ &\le \sum_{S \subseteq N\setminus\{i\}} w_S (2\epsilon) = 2\epsilon. \end{aligned} \tag{39}$$

Thus, the reliable and unreliable components approximate the original Shapley value with an error of at most 2ϵ. This proves Theorem 9.

F.7 Efficiency

Summing both sides of equation 35 over all players and using the efficiency of the ordinary Shapley value gives

$$\sum_{i \in N} \left(\phi_i^{\mathcal{R}}(v) + \phi_i^{\mathcal{U}}(v) \right) = \sum_{i \in N} \phi_i(d) = d(\mathbf{x}_N) - d(\mathbf{x}_\emptyset). \tag{40}$$

This proves the first claim of Theorem 10. For the comparison with the original game, use the residual $r_v = d - v$ from Eq. 36. Then

$$\begin{aligned} &\left| \sum_{i \in N} \left(\phi_i^{\mathcal{R}}(v) + \phi_i^{\mathcal{U}}(v) \right) + v(\mathbf{x}_\emptyset) - v(\mathbf{x}_N) \right| \\ &= |r_v(N) - r_v(\emptyset)| \le |r_v(N)| + |r_v(\emptyset)| \le 2\epsilon. \end{aligned} \tag{41}$$

This proves the second claim of Theorem 10.

F.8 STRUCTURAL PROPERTIES OF SPARSE EXTRACTION

Proposition 16 (Relabeling equivariance)**.** *Let $\sigma : N \to N$ be a permutation. Relabel a game, a candidate residual, and a candidate split by $v^\sigma(S) = v(\sigma S)$, $e^\sigma(S) = e(\sigma S)$, and $\boldsymbol{\gamma}^\sigma(S) = \boldsymbol{\gamma}(\sigma S)$. Then*

$$\mathcal{L}_N(e^\sigma, \boldsymbol{\gamma}^\sigma; v^\sigma) = \mathcal{L}_N(e, \boldsymbol{\gamma}; v). \tag{42}$$

Proof. The relabeled pair remains feasible because relabeling preserves the empty coalition and merely permutes the constraints $|e(S)| \le \epsilon$. In each alternating sum defining an interaction, substitute $K = \sigma L$. For the OR coefficient, also use $\sigma(N \setminus L) = N \setminus \sigma L$. Thus the interaction indexed by T in the relabeled game is the original interaction indexed by σT. As $T \mapsto \sigma T$ permutes the nonempty subsets of N, the two sums of absolute interaction strengths in equation 20 are unchanged. □

Proposition 17 (Symmetric optimal split)**.** *Let σ be a transposition. If $v(\sigma S) = v(S)$ for every $S \subseteq N$, then v admits a globally optimal pair $(e_v^\star, \boldsymbol{\gamma}_v^\star)$ satisfying*

$$e_v^\star(\sigma S) = e_v^\star(S), \qquad \boldsymbol{\gamma}_v^\star(\sigma S) = \boldsymbol{\gamma}_v^\star(S), \qquad \forall S \subseteq N. \tag{43}$$

Proof. Let $(e, \boldsymbol{\gamma})$ be any global minimizer. By Proposition 16 and the symmetry of v, $(e^\sigma, \boldsymbol{\gamma}^\sigma)$ is also a global minimizer. Both the feasible set and the objective are jointly convex in $(e, \boldsymbol{\gamma})$, so

$$(e_v^\star, \boldsymbol{\gamma}_v^\star) = \tfrac{1}{2}\left((e, \boldsymbol{\gamma}) + (e^\sigma, \boldsymbol{\gamma}^\sigma)\right)$$

is again a global minimizer. Since σ is a transposition, both entries of this average are invariant under σ. □

Proposition 18 (Dummy preservation for sparse extraction)**.** *Let $N_{-i} = N \setminus \{i\}$, and suppose that*

$$v(A \cup \{i\}) = v(A), \qquad \forall A \subseteq N_{-i}. \tag{44}$$

For any feasible pair $(e, \boldsymbol{\gamma})$, define $(e', \boldsymbol{\gamma}')$ by copying its values on coalitions that do not contain i:

$$\begin{aligned} e'(A) &= e(A), \quad e'(A \cup \{i\}) = e(A), \\ \boldsymbol{\gamma}'(A) &= \boldsymbol{\gamma}(A), \quad \boldsymbol{\gamma}'(A \cup \{i\}) = \boldsymbol{\gamma}(A), \qquad A \subseteq N_{-i}. \end{aligned} \tag{45}$$

Then $(e', \boldsymbol{\gamma}')$ is feasible and

$$\mathcal{L}_N(e', \boldsymbol{\gamma}'; v) \le \mathcal{L}_N(e, \boldsymbol{\gamma}; v). \tag{46}$$

Consequently, v admits a globally optimal pair $(e_v^\star, \boldsymbol{\gamma}_v^\star)$ satisfying

$$e_v^\star(A \cup \{i\}) = e_v^\star(A), \qquad \boldsymbol{\gamma}_v^\star(A \cup \{i\}) = \boldsymbol{\gamma}_v^\star(A), \qquad A \subseteq N_{-i}. \tag{47}$$

Hence both component games induced by $(e_v^\star, \boldsymbol{\gamma}_v^\star)$ are dummy in i.

Proof. The pair $(e', \boldsymbol{\gamma}')$ is feasible because it preserves the zero values at the empty coalition and copies values that already satisfy $|e(A)| \le \epsilon$. The dummy assumption and equation 45 imply, for either interaction type,

$$v^{\text{type}}_{e', \boldsymbol{\gamma}'}(A \cup \{i\}) = v^{\text{type}}_{e', \boldsymbol{\gamma}'}(A), \qquad A \subseteq N_{-i}. \tag{48}$$

Direct expansion of the interaction transforms therefore gives, for every nonempty $T \subseteq N_{-i}$,

$$I^{\text{and}}_{T, e', \boldsymbol{\gamma}'} = I^{\text{and}}_{T, e, \boldsymbol{\gamma}}, \qquad I^{\text{and}}_{T \cup \{i\}, e', \boldsymbol{\gamma}'} = 0, \tag{49}$$

$$I^{\text{or}}_{T, e', \boldsymbol{\gamma}'} = I^{\text{or}}_{T, e, \boldsymbol{\gamma}} + I^{\text{or}}_{T \cup \{i\}, e, \boldsymbol{\gamma}}, \qquad I^{\text{or}}_{T \cup \{i\}, e', \boldsymbol{\gamma}'} = 0. \tag{50}$$

The first identity in each line follows because the relevant component values on coalitions not containing i are unchanged. The identities for interactions containing i follow by pairing L with $L \cup \{i\}$ and using equation 48. The same calculation with $T = \varnothing$ gives $I^{\text{or}}_{\{i\}, e', \boldsymbol{\gamma}'} = 0$.

Using these identities to group the objective by the pairs $T, T \cup \{i\}$ yields

$$\begin{aligned} \mathcal{L}_N(e', \boldsymbol{\gamma}'; v) &= \sum_{\varnothing \ne T \subseteq N_{-i}} \left| I^{\text{and}}_{T, e, \boldsymbol{\gamma}} \right| \\ &\quad + \sum_{\varnothing \ne T \subseteq N_{-i}} \left| I^{\text{or}}_{T, e, \boldsymbol{\gamma}} + I^{\text{or}}_{T \cup \{i\}, e, \boldsymbol{\gamma}} \right| \qquad (51) \\ &\le \mathcal{L}_N(e, \boldsymbol{\gamma}; v), \qquad (52) \end{aligned}$$

where the inequality is the triangle inequality applied to each OR pair, together with the removal of the AND interactions containing i. This proves equation 46. Apply the construction to any globally optimal pair. The resulting pair cannot have a smaller objective value and is therefore also globally optimal. We select this pair as $(e_v^\star, \boldsymbol{\gamma}_v^\star)$. Its construction gives equation 47, and equation 48 shows that both selected component games are dummy in i. □

F.9 SYMMETRY

Assume the symmetry condition on i and j stated in Theorem 11, and let σ exchange i and j while fixing all other players. The premise implies $v(\sigma S) = v(S)$ for every $S \subseteq N$. Select the invariant global minimizer supplied by Proposition 17. For either $\text{type} \in \{\text{and}, \text{or}\}$, the corresponding component game obeys

$$\begin{aligned} v^{\text{type}}(\sigma S) &= \tfrac{1}{2}\big(v(\sigma S) - e_v^\star(\sigma S)\big) \pm \boldsymbol{\gamma}_v^\star(\sigma S) \\ &= \tfrac{1}{2}\big(v(S) - e_v^\star(S)\big) \pm \boldsymbol{\gamma}_v^\star(S) = v^{\text{type}}(S), \qquad \forall S \subseteq N, \end{aligned} \tag{53}$$

where $+$ is used for and and $-$ for or. Substitution into the interaction definitions gives

$$I^{\text{type}}_{\sigma T, v} = I^{\text{type}}_{T, v}, \qquad \text{type} \in \{\text{and}, \text{or}\}. \tag{54}$$

Accordingly, equation 54 implies

$$T \in \mathcal{C}_v^{\text{type}} \iff \sigma T \in \mathcal{C}_v^{\text{type}}. \tag{55}$$

Using the partition in equation 22, changing variables from T to $K = \sigma T$, and noting that $j \in T$ if and only if $i \in K$, we obtain

$$\begin{aligned} \phi_j^{\mathcal{R}}(v) + \phi_j^{\mathcal{U}}(v) &= \sum_{\text{type}\in\{\text{and},\text{or}\}} \sum_{T \in \mathcal{C}_v^{\text{type}} : T \ni j} \frac{I^{\text{type}}_{T,v}}{|T|} \\ &= \sum_{\text{type}\in\{\text{and},\text{or}\}} \sum_{K \in \mathcal{C}_v^{\text{type}} : K \ni i} \frac{I^{\text{type}}_{\sigma K,v}}{|K|} \\ &= \phi_i^{\mathcal{R}}(v) + \phi_i^{\mathcal{U}}(v). \end{aligned} \tag{56}$$

For the component-wise statement, the two symmetry premises in the theorem allow Proposition 17 to be applied to both v and v_{ref}. We select global minimizers satisfying

$$\begin{aligned} e_v^\star(\sigma S) = e_v^\star(S), \qquad \boldsymbol{\gamma}_v^\star(\sigma S) = \boldsymbol{\gamma}_v^\star(S), \\ e_{v_{\text{ref}}}^\star(\sigma S) = e_{v_{\text{ref}}}^\star(S), \quad \boldsymbol{\gamma}_{v_{\text{ref}}}^\star(\sigma S) = \boldsymbol{\gamma}_{v_{\text{ref}}}^\star(S), \qquad \forall S \subseteq N. \end{aligned} \tag{57}$$

For $w \in \{v, v_{\text{ref}}\}$ and either $\text{type} \in \{\text{and}, \text{or}\}$, the corresponding component game obeys

$$\begin{aligned} w^{\text{type}}(\sigma S) &= \tfrac{1}{2}\big(w(\sigma S) - e_w^\star(\sigma S)\big) \pm \boldsymbol{\gamma}_w^\star(\sigma S) \\ &= \tfrac{1}{2}\big(w(S) - e_w^\star(S)\big) \pm \boldsymbol{\gamma}_w^\star(S) = w^{\text{type}}(S), \qquad \forall S \subseteq N, \end{aligned} \tag{58}$$

where $+$ is used for and and $-$ for or. Thus the selected AND–OR component games of both v and v_{ref} are invariant under σ. Substituting this invariance directly into the two interaction definitions gives

$$\begin{aligned} I^{\text{type}}_{\sigma T, v} &= I^{\text{type}}_{T, v}, \\ I^{\text{type}}_{\sigma T, v_{\text{ref}}} &= I^{\text{type}}_{T, v_{\text{ref}}}, \qquad \text{type} \in \{\text{and}, \text{or}\}. \end{aligned} \tag{59}$$

As relabeling both functions by σ gives σT the same label that T had before relabeling, then we have

$$T \in \mathcal{X}_v^{\text{type}} \iff \sigma T \in \mathcal{X}_v^{\text{type}}, \qquad \mathcal{X} \in \{\mathcal{R}, \mathcal{U}\}. \tag{60}$$

Changing variables from T to $K = \sigma T$ in the component allocation formula now gives, for either $\mathcal{X} \in \{\mathcal{R}, \mathcal{U}\}$,

$$\begin{aligned} \phi_j^{\mathcal{X}}(v) &= \sum_{\text{type}\in\{\text{and},\text{or}\}} \sum_{T \in \mathcal{X}_v^{\text{type}} : T \ni j} \frac{I^{\text{type}}_{T,v}}{|T|} \\ &= \sum_{\text{type}\in\{\text{and},\text{or}\}} \sum_{K \in \mathcal{X}_v^{\text{type}} : K \ni i} \frac{I^{\text{type}}_{\sigma K,v}}{|K|} = \phi_i^{\mathcal{X}}(v). \end{aligned} \tag{61}$$

This proves Theorem 11.

F.10 DUMMY

If i is dummy in v, select the globally optimal pair supplied by Proposition 18. Both extracted component games are then dummy in i, namely

$$\begin{aligned} v^{\mathrm{and}}(A \cup \{i\}) &= v^{\mathrm{and}}(A), \\ v^{\mathrm{or}}(A \cup \{i\}) &= v^{\mathrm{or}}(A), \qquad \forall A \subseteq N \setminus \{i\}. \end{aligned} \tag{62}$$

Fix $T \ni i$ and put $T' = T \setminus \{i\}$. Pairing $L \subseteq T'$ with $L \cup \{i\}$ in the AND transform gives

$$\begin{aligned} I^{\mathrm{and}}_{T,v} = \sum_{L \subseteq T'} \Big[&(-1)^{|T|-|L|} v^{\mathrm{and}}(L) \\ &+(-1)^{|T|-|L|-1} v^{\mathrm{and}}(L \cup \{i\}) \Big] = 0. \end{aligned} \tag{63}$$

For the OR transform, the analogous pairing gives

$$\begin{aligned} I^{\mathrm{or}}_{T,v} = -\sum_{L \subseteq T'} \Big[&(-1)^{|T|-|L|} v^{\mathrm{or}}(N \setminus L) \\ &+(-1)^{|T|-|L|-1} v^{\mathrm{or}}\big(N \setminus (L \cup \{i\})\big) \Big] = 0, \end{aligned} \tag{64}$$

because $N \setminus L$ and $N \setminus (L \cup \{i\})$ differ only by i. Thus every interaction containing i is zero. Hence both $\phi_i^{\mathcal{R}}$ and $\phi_i^{\mathcal{U}}$ vanish. This proves Theorem 12.

F.11 ADDITIVITY

Applying equation 35 separately to $v + u$, v, and u, and then using the additivity of the Shapley value for the original functions, we obtain

$$\begin{aligned} &\big|\phi_i^{\mathcal{R}}(v+u) + \phi_i^{\mathcal{U}}(v+u) - \phi_i^{\mathcal{R}}(v) - \phi_i^{\mathcal{U}}(v) - \phi_i^{\mathcal{R}}(u) - \phi_i^{\mathcal{U}}(u)\big| \\ &= |\phi_i(d_{v+u}) - \phi_i(d_v) - \phi_i(d_u)| \\ &\le |\phi_i(d_{v+u}) - \phi_i(v+u)| + |\phi_i(d_v) - \phi_i(v)| \\ &\quad + |\phi_i(d_u) - \phi_i(u)| + |\phi_i(v+u) - \phi_i(v) - \phi_i(u)| \\ &\le 2\epsilon + 2\epsilon + 2\epsilon + 0 = 6\epsilon. \end{aligned} \tag{65}$$

This proves Theorem 13.

G EXPERIMENTAL DETAILS

We use the LLMs, datasets, and evaluation metrics described in the main paper. To evaluate the decomposition under contrasting linguistic conditions, we use the AdaptLLM finance tasks Cheng et al. (2024a), where predictions often depend on specialized terminology, entities, and relations, and WikiText Merity et al. (2016), which introduces broader topical and stylistic variation through encyclopedic prose. Evaluating both settings allows us to examine whether the reliable/unreliable decomposition remains informative under different lexical and contextual conditions.

Hardware and software. Experiments ran on Ubuntu 20.04.1 LTS (Linux kernel 5.15.0-139-generic, x86_64) with two Intel Xeon Platinum 8352V CPUs at 2.10 GHz (36 physical cores and two hardware threads per CPU; 72 physical cores and 144 logical threads in total) and approximately 540 GB RAM. The machine had three NVIDIA A800 80GB PCIe GPUs, each with 81,920 MiB memory (approximately 240 GB in total), NVIDIA driver 570.211.01, and CUDA runtime 12.8 supported by the driver. The software versions were Python 3.10.20, PyTorch 2.5.1, and Transformers 4.55.4.

For the perturbation-stability evaluation, Gaussian noise is added directly to the embedding of each input sample. Let $E(\mathbf{x}) \in \mathbb{R}^{\ell \times d}$ be the embedding matrix of sample $\mathbf{x}$. The perturbed embedding is

$$\begin{aligned} \widetilde{E}(\mathbf{x}) &= E(\mathbf{x}) + \varepsilon, & \varepsilon_{ij} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_{\mathbf{x}}^2), \\ \sigma_{\mathbf{x}} &= 0.5\ \mathrm{Std}\Big(\{E(\mathbf{x})_{ij}\}_{i,j}\Big). \end{aligned}$$

Table 3: Model-level metrics for the human-annotated relevance analysis.

	Finance				WikiText			
Model	$L^{\mathcal{R}}$	$L^{\mathcal{U}}$	$P^{\mathcal{R}}$	$P^{\mathcal{U}}$	$L^{\mathcal{R}}$	$L^{\mathcal{U}}$	$P^{\mathcal{R}}$	$P^{\mathcal{U}}$
DeepSeek-R1-Distill-Llama-8B	6.75	1.09	0.75	0.55	5.51	0.64	0.61	0.41
Qwen2.5-7B	6.16	0.78	0.76	0.52	6.76	1.00	0.69	0.43
Llama-3-8B	4.90	0.66	0.73	0.48	3.98	0.62	0.63	0.44
Gemma-3-12B	5.76	1.29	0.71	0.46	4.40	0.92	0.58	0.38
Qwen3-8B	7.93	0.99	0.75	0.46	6.90	0.98	0.63	0.40
Falcon-7B	6.11	0.79	0.76	0.48	5.24	0.83	0.63	0.41
Llama-7B	5.10	2.28	0.76	0.59	3.63	0.84	0.61	0.44
Qwen-7B	4.90	0.51	0.68	0.43	5.30	0.65	0.65	0.46
Phi-4	5.81	0.60	0.75	0.49	4.84	0.75	0.60	0.40
Qwen2.5-14B	5.86	0.59	0.72	0.47	4.37	0.85	0.62	0.40
Llama-2-7B	5.27	0.72	0.79	0.55	3.81	0.44	0.64	0.45
Qwen2-7B	6.56	0.80	0.79	0.53	4.44	0.66	0.62	0.48
GLM-Z1-9B-0414	7.28	0.87	0.74	0.52	5.55	0.86	0.63	0.43
Qwen2.5-1.5B	10.98	1.59	0.75	0.51	5.58	0.92	0.62	0.41

Table 4: Proportions for which the reliable component exceeds the unreliable component in the human-annotated relevance analysis.

Dataset	$L^{\mathcal{R}} > L^{\mathcal{U}}$	$P^{\mathcal{R}} > P^{\mathcal{U}}$
Finance	92.1%	86.6%
WikiText	90.3%	76.9%
Combined	91.2%	81.8%

H Human-Annotated Relevance of Shapley Components

Table 3 reports the model-level values of the four relevance metrics. For each model, we aggregate all model-pair entries in which it appears as either Model1 or Model2, and report the mean of $L^{\mathcal{R}}$, $L^{\mathcal{U}}$, $P^{\mathcal{R}}$, and $P^{\mathcal{U}}$ separately for the finance and WikiText datasets. Table 4 reports the proportions of test samples for which the reliable component exceeds the unreliable component, namely $L^{\mathcal{R}} > L^{\mathcal{U}}$ and $P^{\mathcal{R}} > P^{\mathcal{U}}$.

Case description of Figure 4(d).

> **Qwen2.5-14B:** target token *deity*.
> **Input prompt:** *Some epithets eventually became separate deities, as with Werethekau, an epithet applied to several goddesses meaning "great enchantress", which came to be treated as an independent.*

In Figure 4(d), the reliable Shapley component focuses on *independent*, *goddesses*, *deities*, and *Werethekau*. Most of these highly attributed input units are relevant input units (e.g., *independent*, *goddesses*, and *Werethekau*). By contrast, the unreliable Shapley component focuses mainly on irrelevant input units (e.g., *treated*, *enchantress*, and *meaning*).

I Analysis of Noise-Like Unreliable Interactions

Following the interaction-order and cancellation analyses in Lou et al. (2026); Cheng et al. (2025), we examine whether unreliable interactions exhibit a noise-like utility pattern. Specifically, we consider two empirical characteristics: how their positive and negative utilities are distributed over interaction orders, and how much signed utility remains after these utilities are aggregated.

Let $c \in \{\mathcal{R}, \mathcal{U}\}$ denote either the reliable or the unreliable interaction subset. For an input sample $\mathbf{x}$, we define the positive utility mass at order m as $M^{+}_{c,m}(\mathbf{x}) = \sum_{\text{type}\in\{\text{and},\text{or}\}} \sum_{S\in c^{\text{type}}(\mathbf{x}):\,|S|=m} [I^{\text{type}}_S(\mathbf{x})]_+$ and the magnitude of the negative utility mass as $M^{-}_{c,m}(\mathbf{x}) = \sum_{\text{type}\in\{\text{and},\text{or}\}} \sum_{S\in c^{\text{type}}(\mathbf{x}):\,|S|=m} [-I^{\text{type}}_S(\mathbf{x})]_+$, where $[z]_+ = \max(z, 0)$. Thus, both

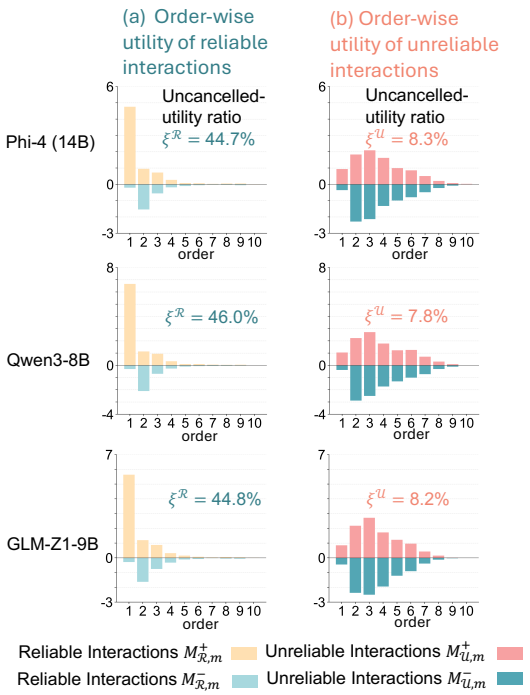


Figure 5: (a) Distributions of reliable interaction utilities, with positive and negative utilities shown separately. (b) Corresponding distributions for unreliable interactions. Each panel also reports the uncancelled-utility ratio ($\xi^{\mathcal{R}}$ or $\xi^{\mathcal{U}}$) of the interactions shown.

$M^{+}_{c,m}(\mathbf{x})$ and $M^{-}_{c,m}(\mathbf{x})$ are nonnegative: the former collects positive interaction utilities, whereas the latter records the magnitude of negative interaction utilities. Interaction order is defined as the number of participating input units, $\mathrm{order}(S) = |S|$ (Ren et al., 2024). The order profiles in Figure 5 visualize $M^{+}_{c,m}$ above the horizontal axis and $-M^{-}_{c,m}$ below it.

For each sample, we compute the fraction of absolute interaction utility that remains after signed aggregation:

$$r_c(\mathbf{x}) = \frac{\left|\sum_{m=1}^{n} \left(M^{+}_{c,m}(\mathbf{x}) - M^{-}_{c,m}(\mathbf{x})\right)\right|}{\sum_{m=1}^{n} \left(M^{+}_{c,m}(\mathbf{x}) + M^{-}_{c,m}(\mathbf{x})\right)}.$$

The reported uncancelled-utility ratio is the sample average

$$\xi^{c} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} r_c(\mathbf{x}), \qquad c \in \{\mathcal{R}, \mathcal{U}\}.$$

A small value of ξ^{c} indicates that positive and negative interaction utilities largely offset each other within individual samples, leaving little net signed utility relative to their total absolute magnitude.

Figure 5 shows a clear contrast between the two interaction subsets. The unreliable interactions retain only 7.8%–8.3% of their absolute utility after signed aggregation, whereas the reliable interactions retain 44.7%–46.0%. Thus, unreliable interactions exhibit substantially stronger positive-negative cancellation. Their order profiles are also often spindle-shaped. Together, these observations are consistent with the noise-like behavior reported in prior interaction analyses (Lou et al., 2026; Cheng et al., 2025).

The model-wide results in Figure 6 show the same separation. For every evaluated LLM, the sample-averaged uncancelled-utility ratio is lower for unreliable interactions (6.2%–10.8%) than for reliable interactions (39.3%–50.5%). Hence, the pronounced cancellation of unreliable interaction utilities occurs across the evaluated architectures rather than being driven by a single model. This cross-model consistency provides further evidence for their noise-like character.

The spindle-shaped order distribution in Figure 5 need not peak exactly at $n/2$. Although n input units are available for a sample, not every unit necessarily participates in the model computation

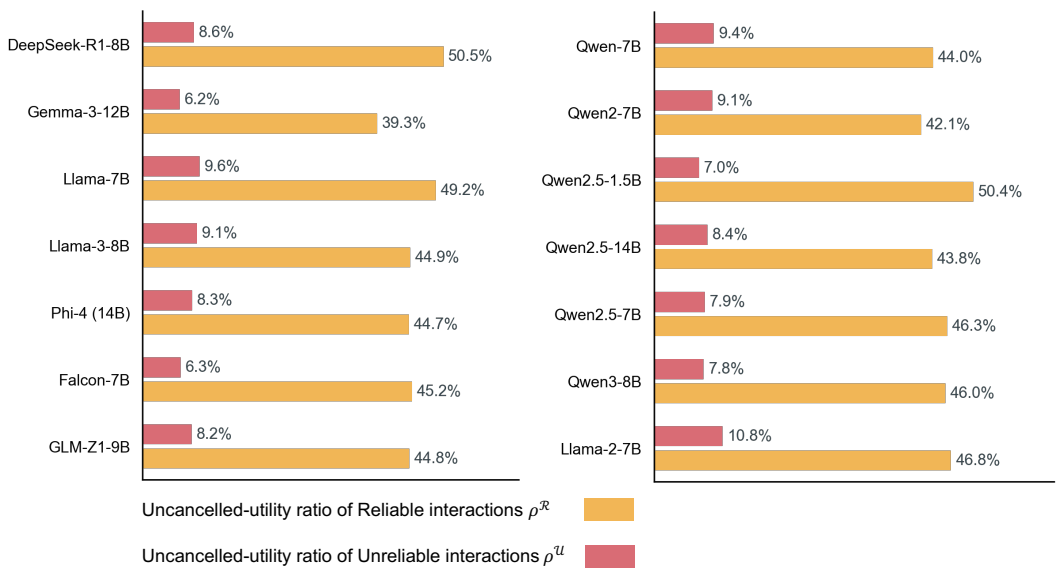


Figure 6: Sample-averaged uncancelled-utility ratios of reliable and unreliable interactions across LLMs.

relevant to the prediction. The effective interaction distribution can therefore be understood as operating on only $n' \leq n$ input units. Its central mass is consequently governed by the smaller effective number of units; when plotted against the original order axis based on n, the peak appears to the left of $n/2$. Variation in the active units across samples can further broaden this left-shifted peak.