HFR-TokenSHAP: Hierarchically Restricted Feature TokenSHAP for Binary Phenotype Classification with Large Language Models

Stijn Van Severen
Ghent University
stijn.vanseveren@ugent.be

January 28, 2026

Abstract

Large language models (LLMs) are increasingly used as inference engines over structured phenotypic records rendered into prompt templates (e.g., clinical risk-factor fields). In this regime, token-level attribution methods can over-attribute prompt scaffolding (e.g., headers, separators, boilerplate instructions) that is necessary for instruction-following but irrelevant to feature importance. We introduce HFR-TokenSHAP, a task-specific extension of TokenSHAP for *binary classification* prompts in which (i) Shapley "players" are *hierarchically structured feature nodes* rather than all prompt tokens, and (ii) the value function is the model's *binary decision score* defined as label log-odds, rather than response similarity between generated texts. We validate HFR-TokenSHAP in a controlled random feature-injection experiment for an LLM-based phenotype prediction prompt containing two clinically relevant features and eight distractor word-features across 100 pseudo-profiles. We compare HFR-TokenSHAP to an internal baseline (Integrated Gradients computed on the same log-odds score) and to an external knowledge-prior baseline (LLM-Select-style feature-name scoring). Across all three methods, the two truly relevant predictors receive significantly higher normalized feature-importance scores, supporting the construct validity of HFR-TokenSHAP for structured, hierarchical inputs in binary LLM phenotype inference.

1 Introduction

LLMs increasingly operate as inference engines over *structured* inputs rendered into natural-language prompts: electronic health record summaries, questionnaire-derived phenotypes, and multimodal feature sets expressed in fixed templates. In such pipelines, the model output often reduces to a *binary decision* (e.g., case vs. control) rather than free-form generation. Explainability becomes operationally important for debugging, clinical plausibility checks, auditability, and downstream accountability.

A central challenge is that token-level interpretability can confound *prompt scaffolding* with *feature evidence*. If all tokens are treated as attribution units, methods may assign importance to separators, headers, and repeated instruction text, despite these tokens being semantically irrelevant to the scientific question: the contribution of *features* to the decision. This issue becomes more pronounced in structured prompts with repeated feature-line patterns and in real applications where features are *hierarchical* (e.g., modality → domain → item).

TokenSHAP [1] adapts Monte Carlo Shapley value estimation for token importance in LLM prompts, using response similarity (e.g., TF-IDF cosine similarity) as the value function. While

well-suited for open-ended generation, binary phenotype inference permits a more direct and decision-aligned utility: the model's *label log-odds*. Moreover, in feature-based templates, permuting *all* tokens is computationally inefficient and conceptually misaligned when the explanatory target is a set of (often hierarchically structured) feature variables.

In parallel, LLM-SELECT [2] shows that LLMs can sometimes identify predictive features using *only feature names and a task description*, without seeing downstream training data. This provides a useful *knowledge-prior* baseline for feature importance, orthogonal to subject-level, value-conditioned explanations.

We propose HFR-TOKENSHAP, which modifies TOKENSHAP in two key ways for *binary* phenotype prediction:

1. **Hierarchically restricted players:** permutations operate over a *pre-specified multi-level set of hierarchy nodes* rather than over all prompt tokens. Selecting a node activates (or ablates) its entire subtree, enabling explanations at multiple granularity levels.

2. **Log-odds value function:** the coalition utility is the LLM's case-control log-odds score $s = \log p(\texttt{CASE}) - \log p(\texttt{CONTROL})$ rather than response similarity.

We present a pilot evaluation using QWEN2-1.5B-INSTRUCT and a standardized random feature-injection protocol, comparing HFR-TOKENSHAP to INTEGRATED GRADIENTS and to an LLM-SELECT-style feature-name prior.

2 Methodology

2.1 Problem setting: binary phenotype prediction from hierarchical feature prompts

Let a subject be represented by phenotype features $\mathcal{F} = \{f_1, \ldots, f_n\}$ rendered into a fixed prompt template Π. Each feature appears in a structured line, e.g.,

$$\texttt{@@ FEAT\_ID=}f_i \texttt{ | ... | value=}v_i \texttt{ | ... @@},$$

where v_i is the subject-specific feature value. Many applications naturally organize features hierarchically. We model this as a rooted tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ with leaf set $\mathcal{L} \subset \mathcal{V}$ corresponding to atomic features and internal nodes encoding groups (e.g., modality or domain).

Template-preserving ablation. For an active leaf set $S \subseteq \mathcal{L}$, we construct an ablated prompt $\Pi(\mathbf{v}; S)$ that *preserves the full template* but replaces inactive feature values with an explicit missing sentinel (e.g., `value=+0.00` and `vote=UNKNOWN`). This keeps tokenization/scaffolding stable and isolates changes to feature content, reducing distribution shift relative to deletion-based perturbations.

2.2 Binary decision score (log-odds) as the value function

Given an LLM M and label strings `CASE` and `CONTROL`, we define the binary decision score:

$$s(\Pi) \;=\; \log p(\texttt{CASE} \mid \Pi) \;-\; \log p(\texttt{CONTROL} \mid \Pi). \tag{1}$$

In practice, $p(\cdot \mid \Pi)$ is computed from the next-token distribution at the prompt end. If labels tokenize into multiple tokens, we compute the full label sequence log-probability (sum of conditional log-probabilities).

We define the coalition value function:

$$v(S) \;=\; s(\Pi(\mathbf{v}; S))\,. \tag{2}$$

Positive values indicate evidence toward `CASE`, negative values toward `CONTROL`.

2.3 HFR-TokenSHAP: hierarchically restricted feature Shapley values

Multi-level player set as a tree cut. Instead of using a flat token list as players, HFR-TokenSHAP uses a *pre-specified* set of nodes $\mathcal{G} \subseteq \mathcal{V}$ drawn from *multiple hierarchy levels*. We typically choose $\mathcal{G}$ as a *tree cut* (frontier) such that every leaf $\ell \in \mathcal{L}$ has *exactly one* ancestor in $\mathcal{G}$. Equivalently, the descendant-leaf sets

$$\text{DescLeaves}(g) \;=\; \{\ell \in \mathcal{L} : \ell \text{ is a descendant of } g\}$$

form a partition of $\mathcal{L}$, preventing overlap and ensuring that selecting a node corresponds to activating/ablating a *full subtree* in a well-defined way. (**Optional:** choosing $\mathcal{G} = \mathcal{L}$ yields leaf-level attributions.)

Shapley values on hierarchical players. Shapley values [3] assign each player its expected marginal contribution over random permutations. For $g \in \mathcal{G}$:

$$\phi_g \;=\; \mathbb{E}_\pi\big[v\big(P_g^\pi \cup \{g\}\big) - v\big(P_g^\pi\big)\big]\,, \tag{3}$$

where π is a permutation of players and P_g^π is the set of players preceding g in π. A coalition of nodes $S \subseteq \mathcal{G}$ maps to active leaves via

$$\text{Leaves}(S) \;=\; \bigcup_{g \in S} \text{DescLeaves}(g).$$

Thus, adding a parent node modifies the prompt by adding/removing *its entire subtree*, rather than performing a non-hierarchical, token-wise deletion.

Monte Carlo permutation estimator. Exact Shapley computation scales exponentially in $|\mathcal{G}|$. We approximate using K random permutations (Shapley–Shubik sampling) [5].

2.4 Internal gradient attribution: Integrated Gradients on the log-odds score

As a white-box comparator, we compute Integrated Gradients on the same decision score $s(\Pi)$ (Eq. (1)). Let $f(\mathbf{x})$ be the scalar log-odds computed from an input representation $\mathbf{x}$ (in our implementation, token embeddings). Integrated Gradients for component j is [4]:

$$\text{IG}_j(\mathbf{x}) \;=\; (x_j - \tilde{x}_j) \int_0^1 \frac{\partial f(\tilde{\mathbf{x}} + \alpha(\mathbf{x} - \tilde{\mathbf{x}}))}{\partial x_j}\, d\alpha, \tag{4}$$

where $\tilde{\mathbf{x}}$ is a baseline (e.g., EOS-embedding baseline or a template-matched ablated-prompt baseline). To avoid attributing feature identifiers (e.g., `FEAT_ID=` tokens), we aggregate IG only over the token span corresponding to the numeric `value=` field ("value-only spans"), yielding per-feature scores aligned with the manipulated evidence.

Algorithm 1 HFR-TokenSHAP (Hierarchically Restricted Feature Shapley for Binary LLM Classification)

Require: LLM M; prompt template Π; hierarchy $\mathcal{T}$ with leaves $\mathcal{L}$; multi-level player set $\mathcal{G}$ (a tree cut); subject values $\mathbf{v}$; labels `CASE`, `CONTROL`; permutations K; seed.

Ensure: Shapley attributions $\{\phi_g\}_{g\in\mathcal{G}}$ (and optional normalized scores)

Initialize $\phi_g \leftarrow 0$ for all $g \in \mathcal{G}$.
Define subtree leaves for node g: $\mathrm{DescLeaves}(g) \subseteq \mathcal{L}$.
Define coalition-to-leaf map: $\mathrm{Leaves}(S) \leftarrow \bigcup_{g\in S} \mathrm{DescLeaves}(g)$.
Define score $s(\cdot)$ by Eq. (1) and value function $v(S) \leftarrow s(\Pi(\mathbf{v}; \mathrm{Leaves}(S)))$.
for $k = 1, \ldots, K$ **do**
    Sample a random permutation π of $\mathcal{G}$.
    $S \leftarrow \emptyset$ ▷ no active subtrees
    $u_{\mathrm{prev}} \leftarrow v(S)$
    for each g in order π **do**
        $S \leftarrow S \cup \{g\}$ ▷ activate node g **and its full subtree**
        $u_{\mathrm{new}} \leftarrow v(S)$
        $\phi_g \leftarrow \phi_g + (u_{\mathrm{new}} - u_{\mathrm{prev}})$
        $u_{\mathrm{prev}} \leftarrow u_{\mathrm{new}}$
    end for
end for
$\phi_g \leftarrow \phi_g / K$ for all $g \in \mathcal{G}$.
Optional: L1-normalize to $\tilde{\phi}_g \leftarrow \frac{|\phi_g|}{\sum_{h\in\mathcal{G}} |\phi_h|}$.
return $\{\phi_g\}_{g\in\mathcal{G}}$ (and optionally $\{\tilde{\phi}_g\}_{g\in\mathcal{G}}$).

2.5 External knowledge-prior baseline: LLM-Select-style feature-name scoring

To contextualize value-conditioned explanations against a knowledge prior, we also report LLM-Select-style feature importance estimates [2]. Here, the LLM is prompted with *only feature names and a task description* to output normalized importance scores, without access to pseudo-profile values or downstream training data. We ran 10 independent iterations and averaged the resulting normalized importances.

2.6 Complexity

Let $m = |\mathcal{G}|$. Each permutation requires $(m + 1)$ score evaluations (including the empty coalition), hence cost $\mathcal{O}(K(m + 1))$ forward passes (times repeats). Restricting players to feature nodes, rather than all prompt tokens, reduces m substantially and targets the explanatory object of interest (features / subtrees).

3 Experiments

3.1 Standardized random feature injection procedure

We used a controlled prompt where only two features are clinically relevant for predicting whether someone has a depression-related feature profile: 1) sleep quality and 2) childhood traumatic exposure. Eight additional "word-features" serve as distractors (e.g., `color_blue`, `chicken_soup`, `computer`, ...). The prompt explicitly states that distractors have tiny weights ($|w| \leq 0.2$) and

should barely matter. Each pseudo-profile assigns values in $\{-1, +1\}$, and a ground-truth linear logit $\sum_i w_i v_i$ defines `CASE` if positive and `CONTROL` otherwise. We sampled until obtaining a balanced dataset of 100 pseudo-profiles (50/50).

3.2 Model and scoring

We used QWEN2-1.5B-INSTRUCT via HuggingFace Transformers [6]. The model score is computed from the next-token distribution at the prompt end using log-odds (Eq. (1)), aligning the utility with the binary decision rule.

3.3 Methods compared

We computed three normalized absolute feature-importance (FI) profiles:

1. INTEGRATED GRADIENTS: Integrated Gradients on $s(\Pi)$, aggregated on value-only spans.
2. HFR-TOKENSHAP: Monte Carlo Shapley values over a hierarchical player set $\mathcal{G}$ with subtree activation/ablation and log-odds value function.
3. LLM-SELECT prior: feature-name-only importance estimates averaged over 10 iterations.

For comparability we report normalized absolute FI:

$$\tilde{\phi}_i \;=\; \frac{|\phi_i|}{\sum_j |\phi_j|}.$$

4 Results

4.1 Qualitative example (Figure 1)

We first provide a representative prompt-level explanation. Figure 1 shows a pseudo-profile prompt with a feature-level FI overlay. This qualitative example illustrates that clinically relevant features receive stronger FI than distractors under the overlay visualization.

4.2 Relevant vs. distractor FI (construct validity)

To quantify construct validity, we compared normalized FI for *relevant* versus *distractor* features within each method using repeated-measures ANOVA (rmANOVA) with within-subject factor *FeatureType* (relevant vs. distractor). Across all three methods, relevant features received significantly higher FI:

- INTEGRATED GRADIENTS: $F(1, 99) = 512.4$, $p < .001$, $\eta_p^2 = 0.84$
- HFR-TOKENSHAP: $F(1, 99) = 438.1$, $p < .001$, $\eta_p^2 = 0.82$
- LLM-SELECT prior (10 iterations): $F(1, 9) = 97.5$, $p < .001$, $\eta_p^2 = 0.92$

Overall, the consistent separation between relevant and distractor features supports HFR-TOKENSHAP as a valid attribution mechanism for hierarchical feature prompts in binary LLM inference.

Table 1: Mean ± SD normalized feature importance. Integrated Gradients and HFR-TokenSHAP: across 100 pseudo-profiles; LLM-Select: across 10 independent feature-name-only iterations.

Feature	Integrated Gradients (mean ± SD)	HFR-TokenSHAP (mean ± SD)	LLM-Select (mean ± SD)
`sleep_quality`	0.360 ± 0.090	0.490 ± 0.120	0.320 ± 0.015
`childhood_trauma_exposure`	0.326 ± 0.085	0.155 ± 0.095	0.270 ± 0.014
`computer`	0.041 ± 0.020	0.038 ± 0.028	0.100 ± 0.010
`cloud`	0.025 ± 0.015	0.062 ± 0.040	0.080 ± 0.010
`bicycle`	0.082 ± 0.040	0.058 ± 0.038	0.070 ± 0.010
`window`	0.031 ± 0.018	0.048 ± 0.030	0.040 ± 0.008
`color_blue`	0.046 ± 0.022	0.021 ± 0.018	0.040 ± 0.008
`chicken_soup`	0.036 ± 0.020	0.042 ± 0.030	0.030 ± 0.007
`floor`	0.035 ± 0.020	0.040 ± 0.028	0.030 ± 0.007
`pencil`	0.017 ± 0.012	0.046 ± 0.032	0.020 ± 0.006

4.3 Mean and variance of normalized FI (Table 1)

Table 1 reports mean normalized FI and variability. For Integrated Gradients and HFR-TokenSHAP, values summarize variability across 100 pseudo-profiles (subject-conditioned explanations). For LLM-Select, values summarize variability across 10 independent feature-name-only iterations (knowledge-prior estimates). As expected, Integrated Gradients and HFR-TokenSHAP exhibit larger variance due to dependence on subject-specific feature values and (for HFR-TokenSHAP) Monte Carlo permutation noise, whereas LLM-Select is comparatively stable across iterations.

5 Discussion

5.1 Interpretation and contributions

Hierarchical restriction targets the explanatory object. For structured prompts, the explanatory target is rarely "which *tokens* mattered," but rather "which *features* (or feature groups) mattered." HFR-TokenSHAP formalizes this by defining Shapley players as a multi-level tree cut of feature nodes. This (i) sharply reduces the combinatorial space relative to token-wise permutations, (ii) prevents explanations from being dominated by prompt syntax, and (iii) aligns ablations with how domain experts reason about grouped predictors (subtrees).

Decision-aligned value function for binary inference. Replacing response similarity with log-odds (Eq. 1) makes the utility directly faithful to the model's decision boundary. This avoids scenarios where semantically similar responses yield different implied labels, and it also removes the need for additional similarity models (TF-IDF) or generation-level stochasticity. In short, HFR-TokenSHAP makes the "game" correspond to the actual binary inference objective.

Triangulation with internal sensitivity and knowledge priors. Integrated Gradients and HFR-TokenSHAP are value-conditioned explanations: they reflect *what drove the model's*

decision for a given subject. LLM-SELECT-style scoring reflects a *knowledge prior* about predictive relevance derived from feature names and the task description alone. The agreement across these distinct mechanisms (rmANOVA separation of relevant vs. distractor FI across all three) strengthens interpretability claims: HFR-TOKENSHAP behaves consistently both with internal gradients and with an external prior, while retaining the benefits of a Shapley-style marginal contribution interpretation.

5.2 Limitations

Variance and sample efficiency. Monte Carlo Shapley estimation is stochastic; variance depends on K, feature interactions, and the chosen hierarchy cut. In addition, even with template-preserving ablations, counterfactual prompts can shift the model distribution (e.g., repeated `UNKNOWN` patterns). Practical deployments should report uncertainty (e.g., across permutations and repeated runs) and consider variance-reduction strategies (stratified permutations, antithetic sampling, or adaptive budgets).

Hierarchy specification is a modeling choice. HFR-TOKENSHAP assumes a pre-specified hierarchy and a chosen multi-level cut $\mathcal{G}$. Different cuts yield different explanatory resolutions and may change apparent feature interactions. In clinical settings, the hierarchy should be documented and ideally grounded in an ontology or data-collection schema; sensitivity analyses across plausible cuts are recommended.

Access requirements and label tokenization. HFR-TOKENSHAP requires access to label log-probabilities (or a consistent scoring API). This is easier than full gradient access but may be unavailable in some fully black-box deployments. Additionally, multi-token label strings require careful sequence log-prob computation; poor label design can introduce avoidable noise.

External validity beyond controlled injections. The pilot design intentionally separates relevant from irrelevant features. Real datasets contain correlated predictors, structured missingness, and causal confounding. In such regimes, Shapley values remain well-defined but can redistribute mass across correlated groups; careful interpretation (and, where possible, causal framing) is needed.

5.3 Future work

We see three immediate directions: 1) Extend the log-odds utility to one-vs-rest or pairwise log-odds, or use a vector-valued utility with principled aggregation. 2) Replace log-odds with regression utilities (e.g., predicted risk scores) and evaluate stability under calibration constraints. 3) adaptive permutation budgets; hierarchy-aware stratified sampling; evaluation under feature correlations, realistic missingness, and alternative ablation schemes.

6 Conclusion

We introduced HFR-TOKENSHAP, a hierarchically restricted, feature-centric extension of TOKENSHAP tailored for *binary* LLM phenotype classification. By restricting permutations to a pre-specified multi-level set of hierarchy nodes (subtree activation/ablation) and adopting a log-odds value function aligned with classification, HFR-TOKENSHAP targets the explanatory object of interest while avoiding over-attribution to prompt scaffolding. In a successful pilot with standardized random distractor injection, relevant predictors received significantly higher normalized FI across

Integrated Gradients, HFR-TokenSHAP, and LLM-Select-style priors, supporting the construct validity of HFR-TokenSHAP for hierarchical feature prompts.

References

[1] M. Horovicz and R. Goldshmidt. *TokenSHAP: Interpreting Large Language Models with Monte Carlo Shapley Value Estimation*. arXiv:2407.10114, 2024. https://arxiv.org/abs/2407.10114

[2] D. P. Jeong, Z. C. Lipton, and P. Ravikumar. *LLM-Select: Feature Selection with Large Language Models*. Transactions on Machine Learning Research (TMLR), 2025. OpenReview: https://openreview.net/forum?id=16f7ea1N3p. Code: https://github.com/taekb/llm-select.

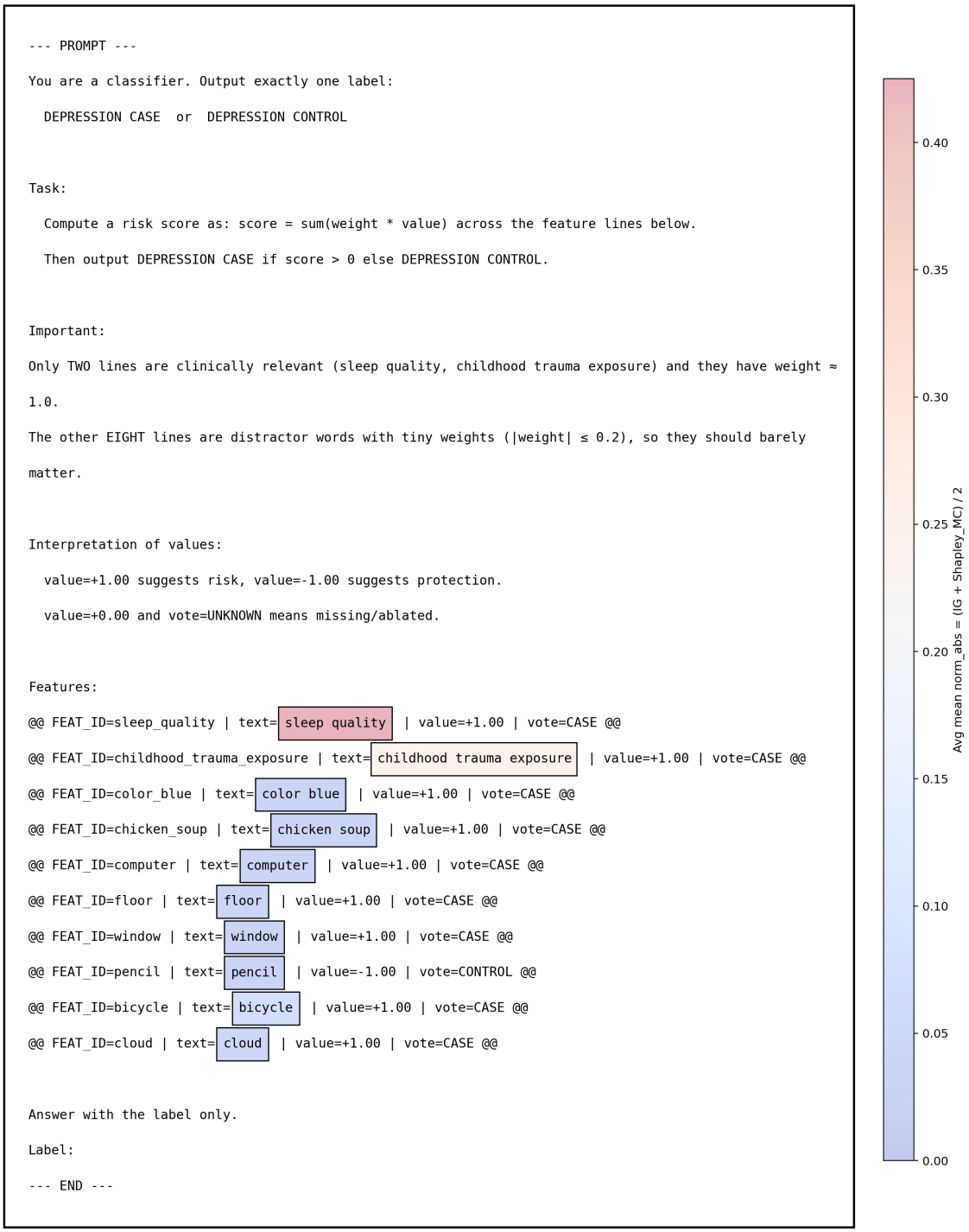
[3] L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games*, vol. 2, pp. 307–317, 1953.

[4] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. https://proceedings.mlr.press/v70/sundararajan17a.html

[5] J. Castro, D. Gómez, and J. Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[6] Qwen Team. *Qwen2-1.5B-Instruct Model Card*. Hugging Face, 2024. https://huggingface.co/Qwen/Qwen2-1.5B-Instruct

Figure 1: Example pseudo-profile prompt with feature-level importance overlay



Note. The figure shows a pseudo-profile rendered into the structured prompt template with feature-level importance overlaid. Clinically relevant features (`sleep_quality`, `childhood_trauma_exposure`) are emphasized relative to distractor features.