Measuring Concept Completeness for Explainability in AI-Driven Network Functions

Alexander Niedermayer[1], David Monschein[1], and Oliver Waldhorst[1]

Karlsruhe University of Applied Sciences
{alexander.niedermayer,david.monschein,oliver.waldhorst}@h-ka.de

Abstract. Concept-based explanations ground ML decisions in domain concepts, assuming a concept set that captures the model's reasoning without verification. We introduce a representation-level framework for quantifying concept completeness, measuring the variance in internal representations captured by the concept subspace, yielding a global metric $\kappa^* \in [0, 1]$. For high-dimensional, noisy spaces, we add a significant-component measure τ^* via singular value decomposition and a Shapley-based importance metric. On three AI networking scenarios, we find congestion control substantially complete and the other two only moderately complete. We presented this work at the IFIP Networking 2026 conference [1].

Keywords: Concept-based Explanations, Explainable AI, Concept Completeness, AI-Driven Networking

1 Introduction

Concept-based explainability [2] projects model internals onto human-understandable concepts. The Agua framework [3] demonstrated this for AI-driven networking but left open verification of whether chosen concepts capture a model's learned representations. Prior completeness methods [2] evaluate only prediction-level reconstruction and miss cases where concepts span a narrow latent subspace. We propose a representation-level notion of concept completeness with four contributions: a global metric κ^* from variance preservation under projection; a significant-component measure τ^* via SVD; a Shapley-based importance metric; and evaluation on three AI networking scenarios.

2 Method

Given internal representations $Z = \{z_i\}_{i=1}^N$ and concept vectors $C = \{c_j\}_{j=1}^K$, we compute how much variance in Z the subspace spanned by C captures. We project each z_i onto $\text{span}(C)$ via QR decomposition ($C = QR$, orthonormal basis Q): $\hat{z}_i = z_i Q Q^\top$. The *global concept completeness* κ^* is the ratio of variance in projected representations to total variance, a value in $[0, 1]$.

For high-dimensional, noisy spaces, we apply SVD, keep the top singular directions covering τ percent of variance, and recompute completeness as $\kappa^*(\tau)$; the *reduction threshold* τ_C^* is the largest τ with $\kappa_C(\tau) \geq 1 - \varepsilon$. Concept importance uses Shapley values [4]; the method is post-hoc, model-agnostic, and needs no retraining.

3 Evaluation and Results

Synthetic validation confirms that κ^* matches ground truth under no noise and degrades smoothly with noise and dimensionality; in high-dimensional spaces ($d_z = 2048$), τ^* tracks ground truth more closely.

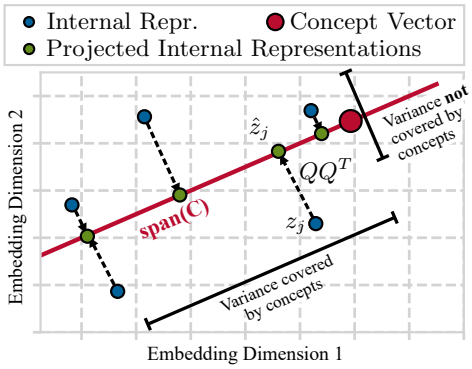


Fig. 1. Visualization of the main steps for our completeness method.

We evaluate three AI-driven networking scenarios from [3] (3096-dimensional, 5000 samples, 5 folds). Global completeness is uniformly low ($\kappa^* \in [8.7\%, 13.3\%]$), but the significant-component measure distinguishes the scenarios: *congestion control* reaches *substantial completeness* ($\tau^* \approx 76\%$), *adaptive bitrate streaming* ($\tau^* \approx 58\%$) and *DDoS detection* ($\tau^* \approx 63\%$) only *moderate completeness*, answering the open question from [3].

Concept importance rankings align with domain knowledge: for congestion control, "increasing packet loss" and "rapidly increasing latency" rank highest, matching rule-based TCP congestion control; for DDoS detection, "high request rates," "geographic irregularities," and "typical application behavior" rank highest. Computations finish in under 130 ms (global) and 49 s (with Shapley importance).

4 Conclusion

We presented a representation-level framework for quantifying concept completeness: κ^* measures the variance captured by concept subspaces, τ^* extends it to high dimensions via SVD, and Shapley values rank importance. The framework finds the congestion control concept set substantially complete and the other two only moderately complete. Future work extends the framework to concept selection and links completeness to generalization.

References

1. Niedermayer, A., et al.: Measuring concept completness for explainability in AI-driven network functions. In: IFIP Networking 2026; Lugano, Switzerland. (2026)
2. Yeh, C.K., et al.: On completeness-aware concept-based explanations in deep neural networks. In: Advances in Neural Information Processing Systems. Volume 33., Curran Associates, Inc. (2020) 20554–20565
3. Patel, S., Han, D., Narodytska, N., Jyothi, S.A.: Agua: A concept-based explainer for learning-enabled systems. In: Proceedings of the ACM SIGCOMM 2025 Conference. SIGCOMM '25, ACM (August 2025) 329–346
4. Shapley, L.S. In: A value for n-person games. Cambridge University Press (October 1958) 31–40