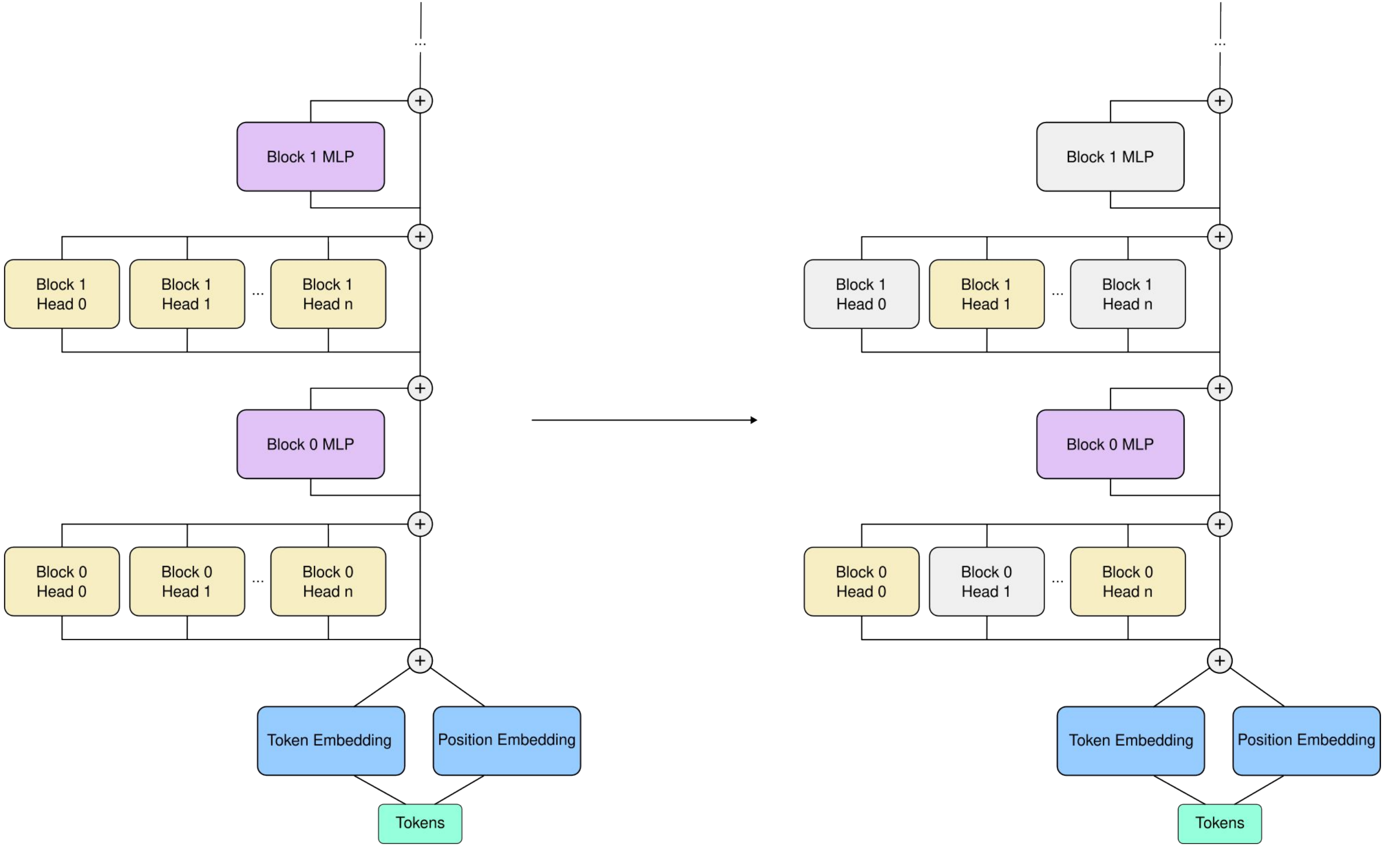
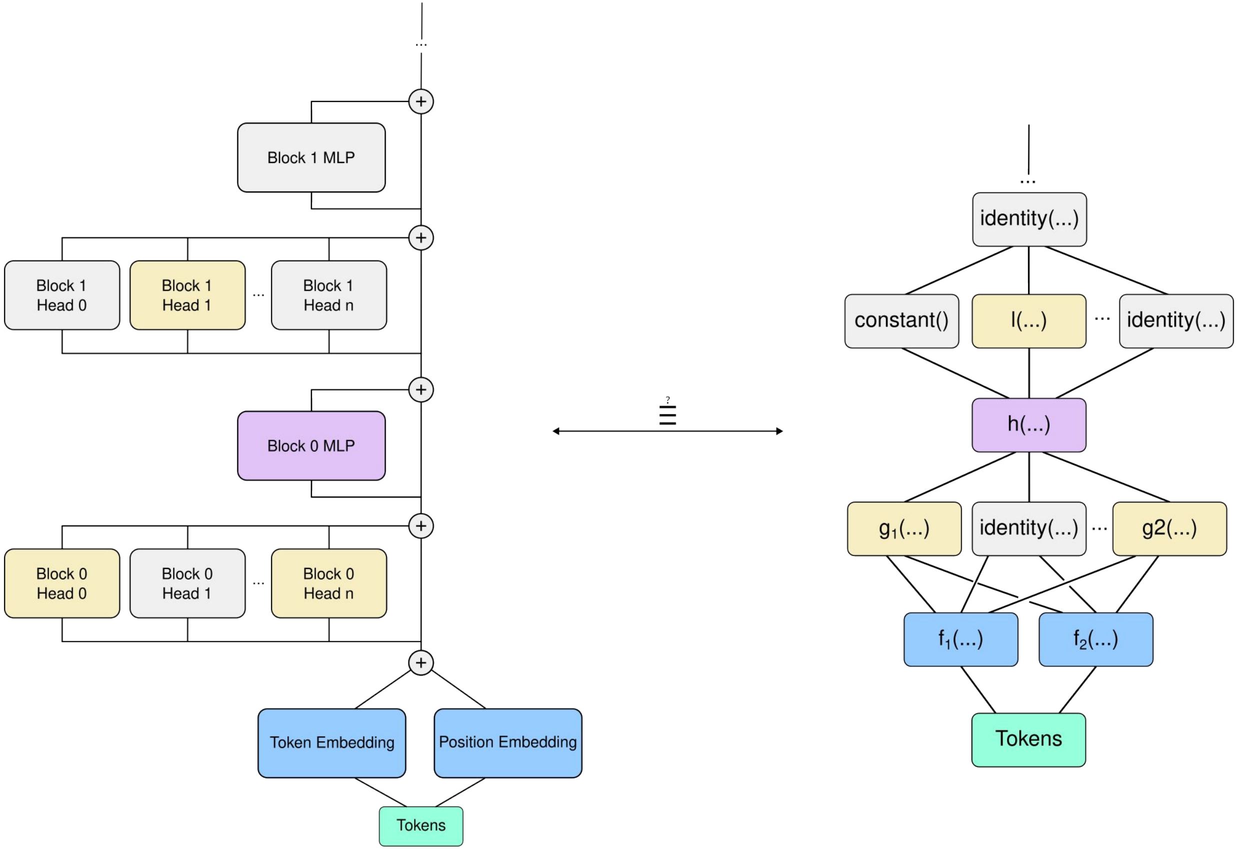
Validating Mechanistic Interpretations: An Axiomatic Approach

Nils Palumbo, Ravi Mangal, Zifan Wang,
Saranya Vijayakumar, Corina Pǎsǎreanu, Somesh Jha

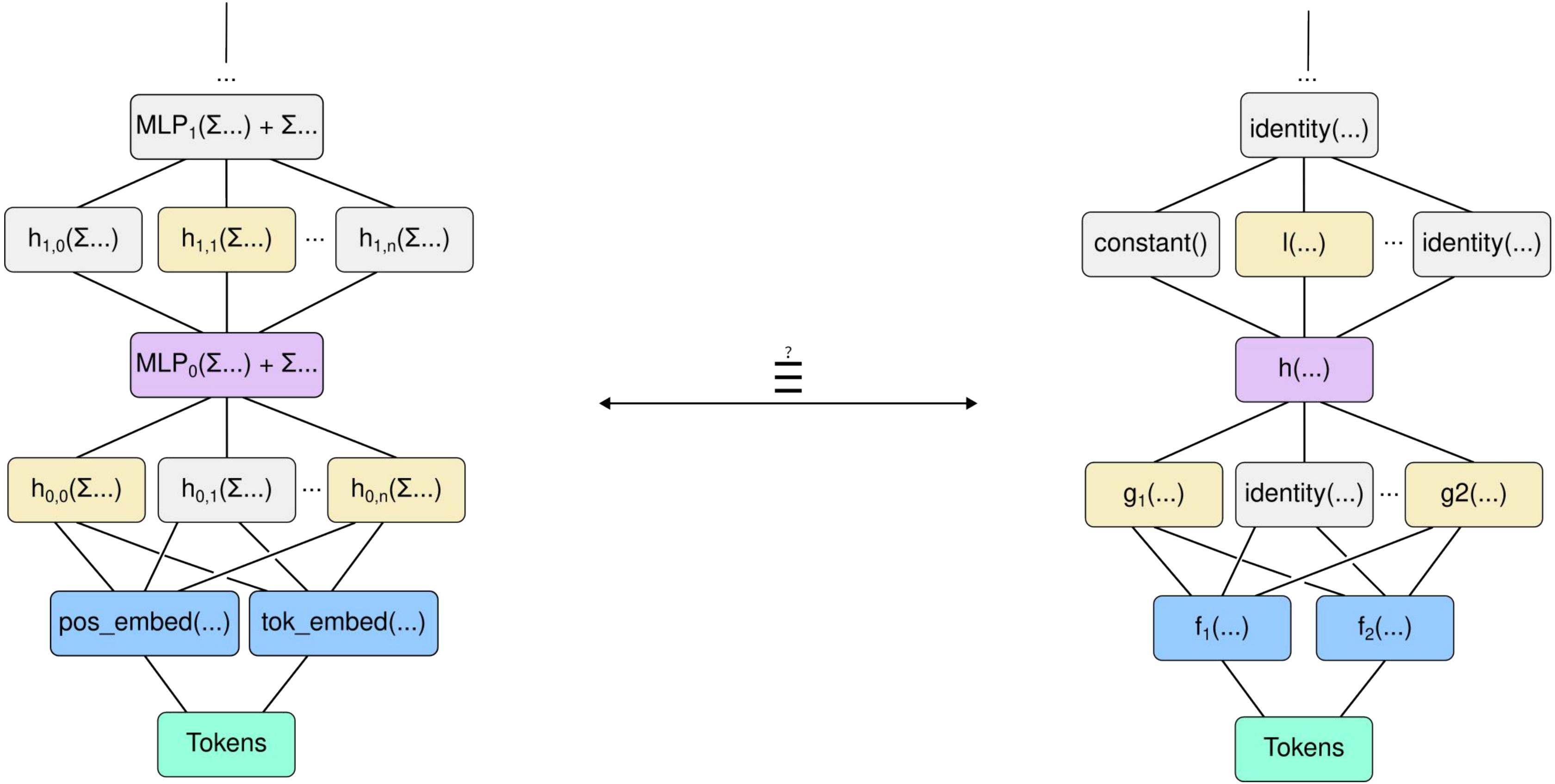
Mechanistic Interpretations: Extracting a Circuit



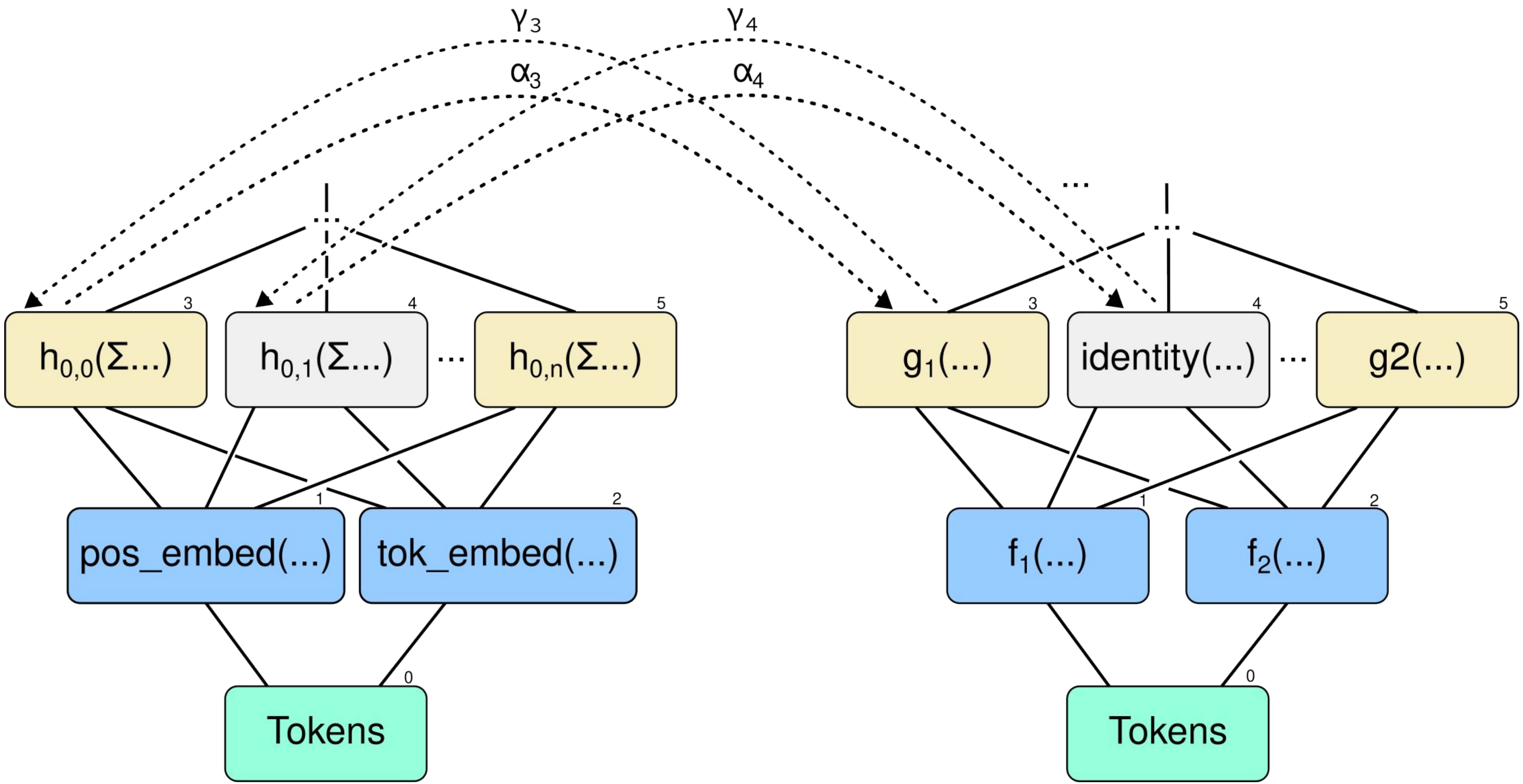
Mechanistic Interpretations: Candidate Interpretation



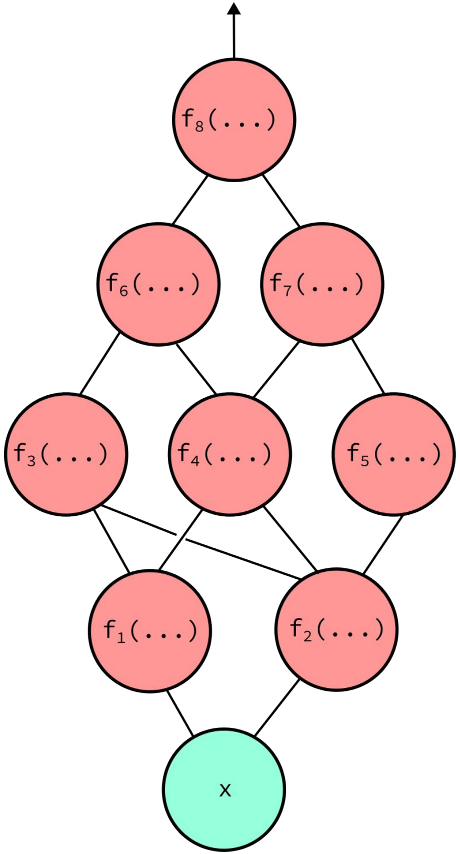
Concrete and Abstract Models



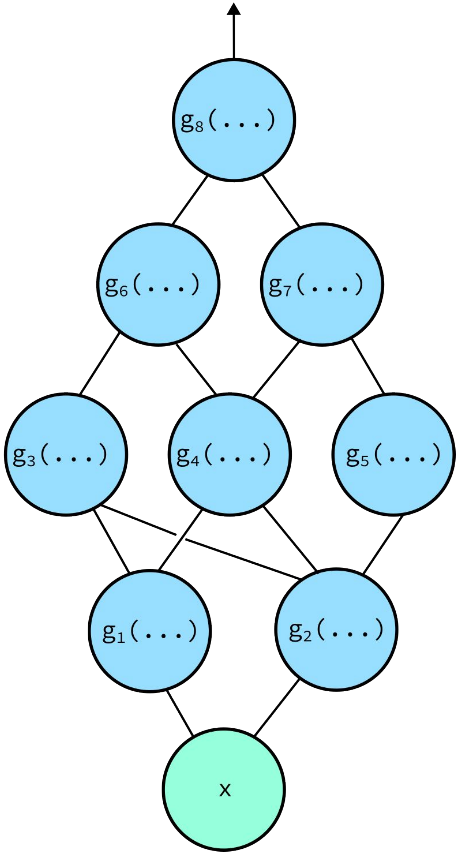
Abstraction and Concretization



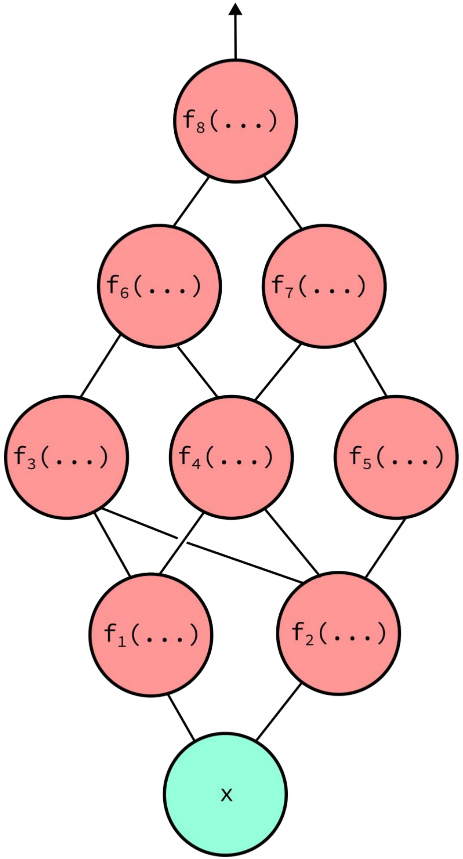
Validating Equivalence of Models



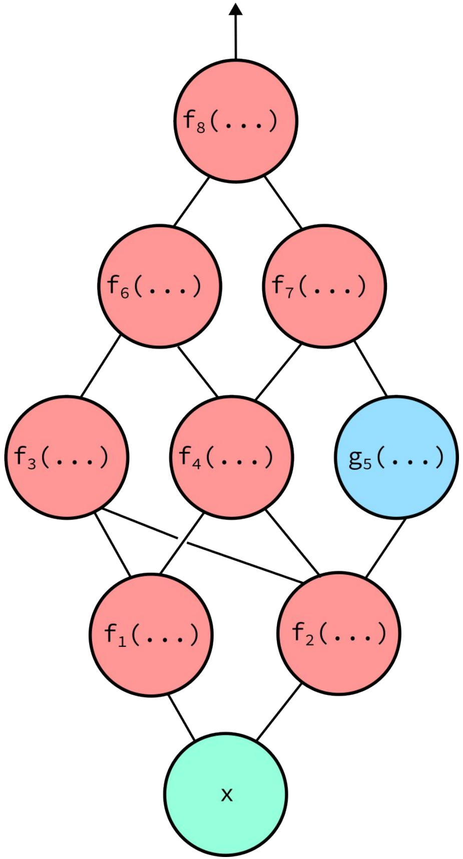
$\overset{?}{\equiv}$



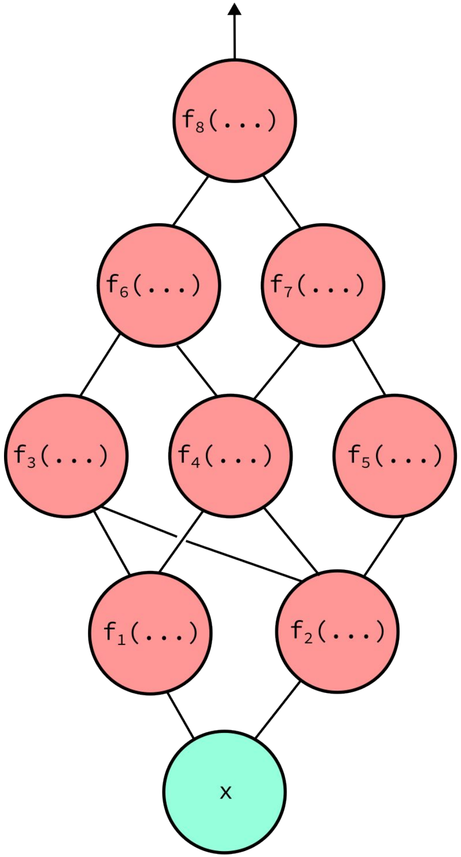
Equivalence up to Interleaving



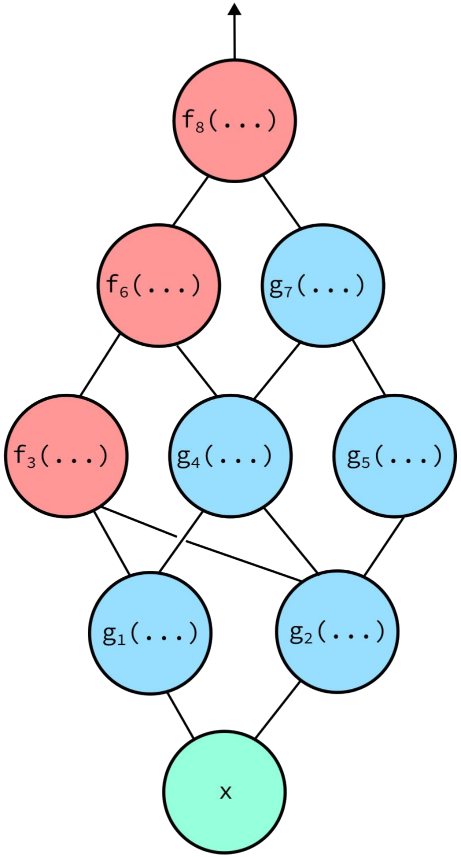
$\overset{?}{\equiv}_{\alpha}$



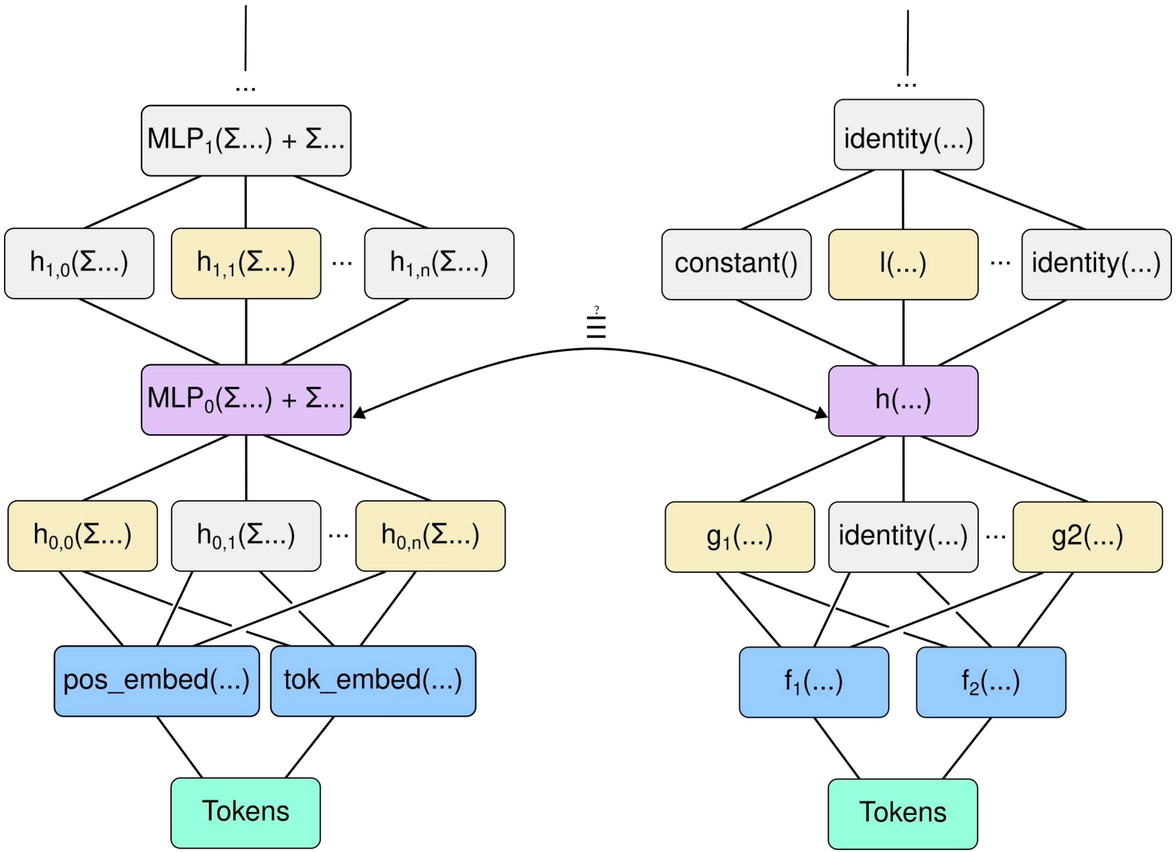
Equivalence up to Interleaving



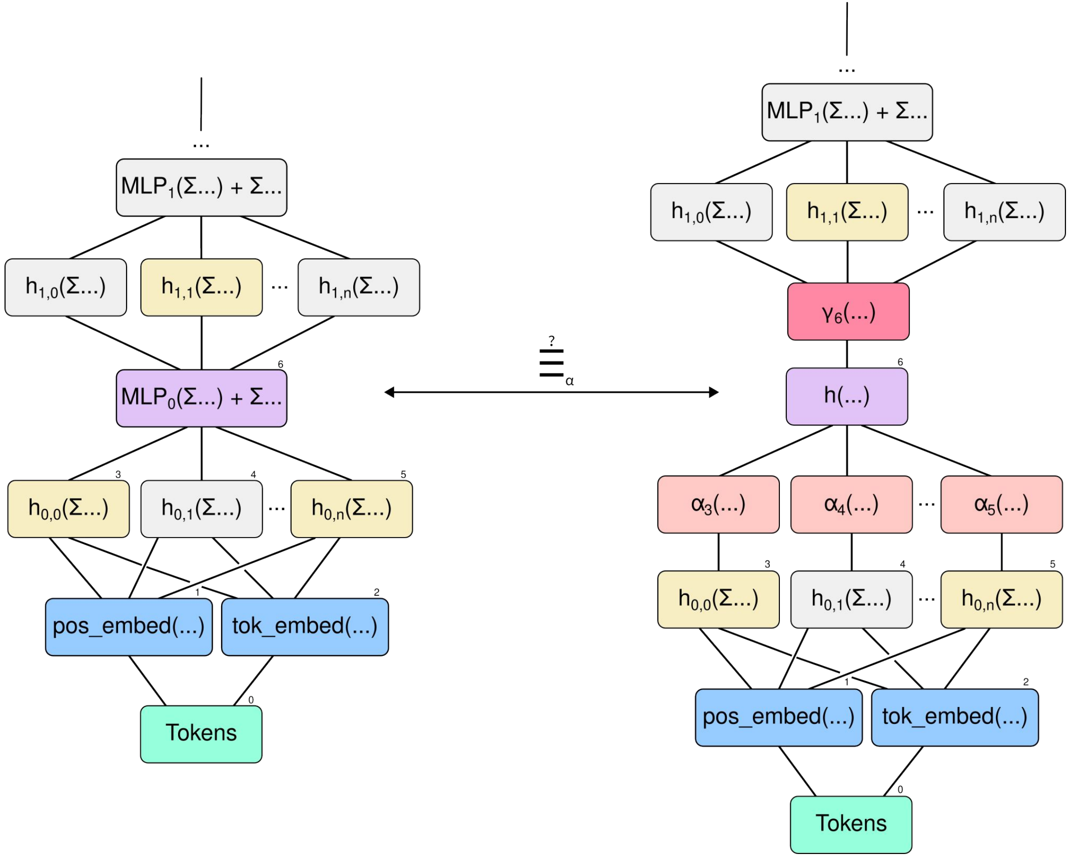
$\overset{?}{\equiv}_\alpha$



Interleaving the Concrete and Abstract Models



Handling Type Mismatch with Abstraction and Concretization



See More in the Paper!

- An axiomatic definition of a valid mechanistic interpretation
  - Characterized by invariance to interleaving
- We validate our approach with two case studies using our evaluation framework
  - A detailed original analysis of a model trained to solve the 2-SAT problem:
    - The model implements a simple parser followed by an *approximate brute-force evaluation*
  - Evaluating a well-known mechanistic interpretation:
    - A model trained to perform modular addition (Nanda et al., 2023)