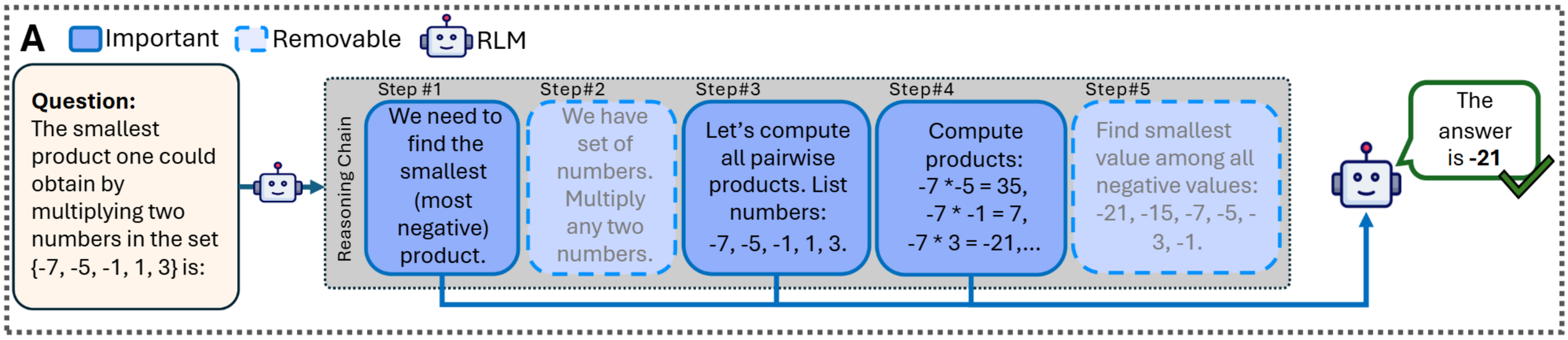
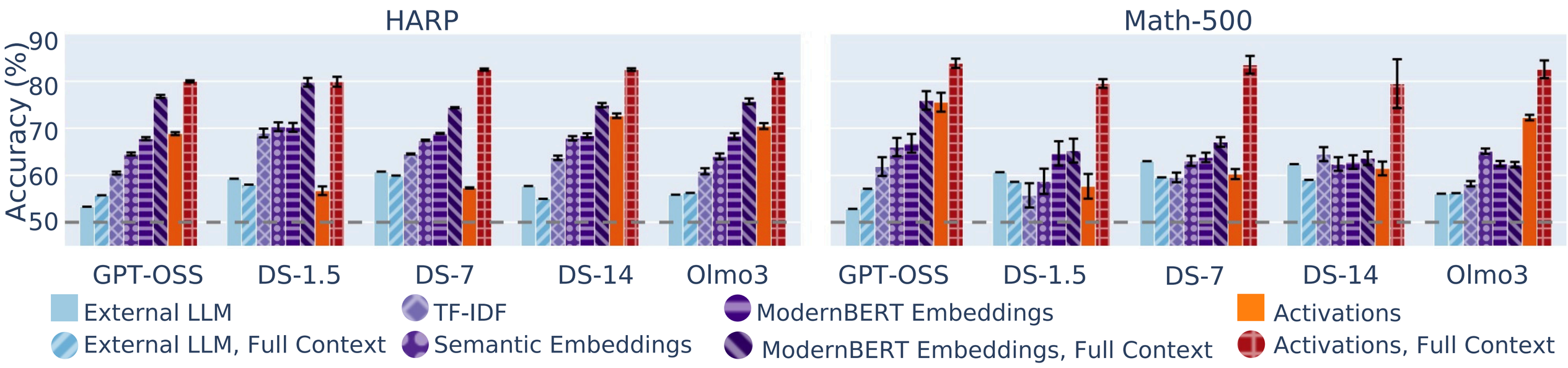
Reasoning models encode the **importance** of a reasoning step in its **activations** (**before** later steps are generated), while importance can **not** be **faithfully recovered** from **tokens** alone.

1. Identifying imporance: Not all reasoning steps are important. We use causal analysis to extract a minimal subsequence of steps that is both sufficient (high P_{answer}) and attainable (low perplexity): the core subsequence that the model actually needs.

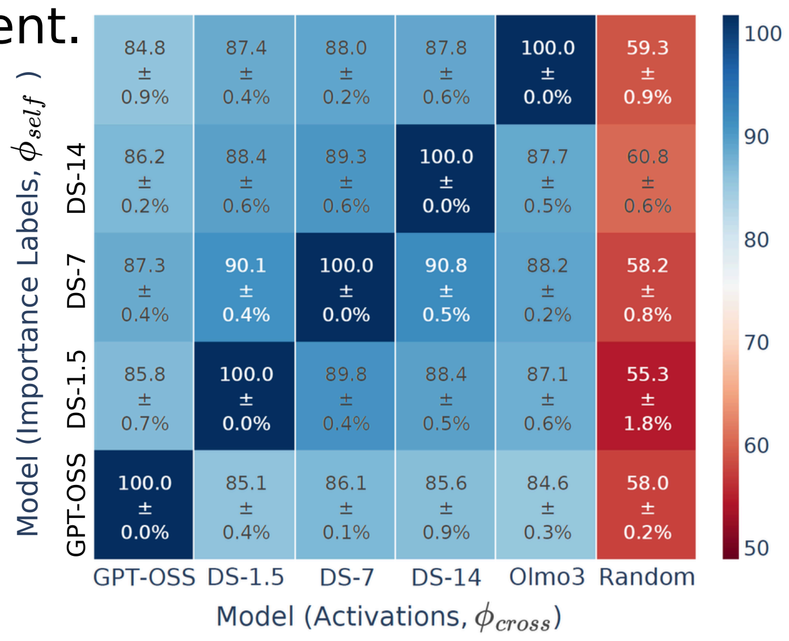


2. Importance lives in activations: Given importance labels, **probes trained on model activations** recover this concept, pre-hoc: without access to information from later steps. They beat **token-based** and **weaker-representation** baselines.

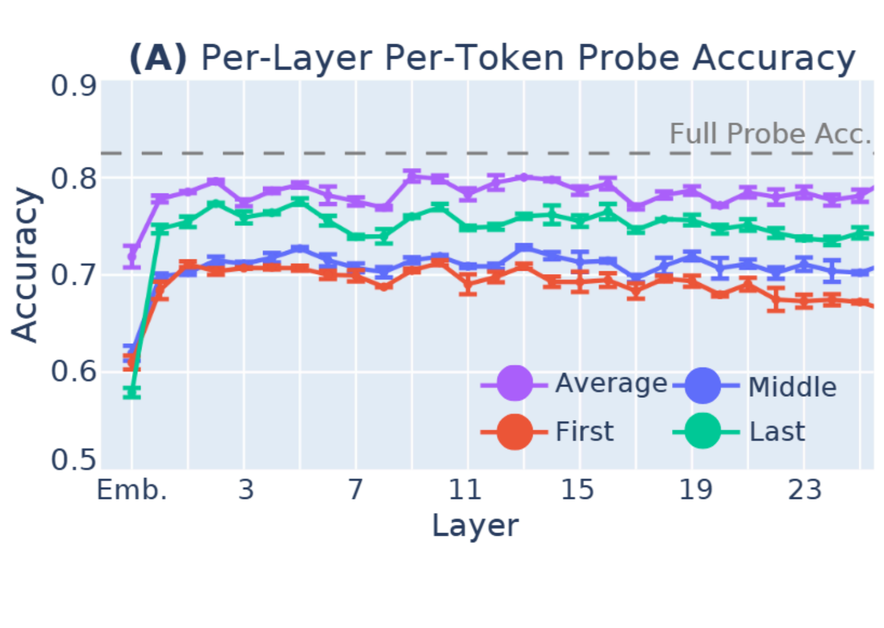


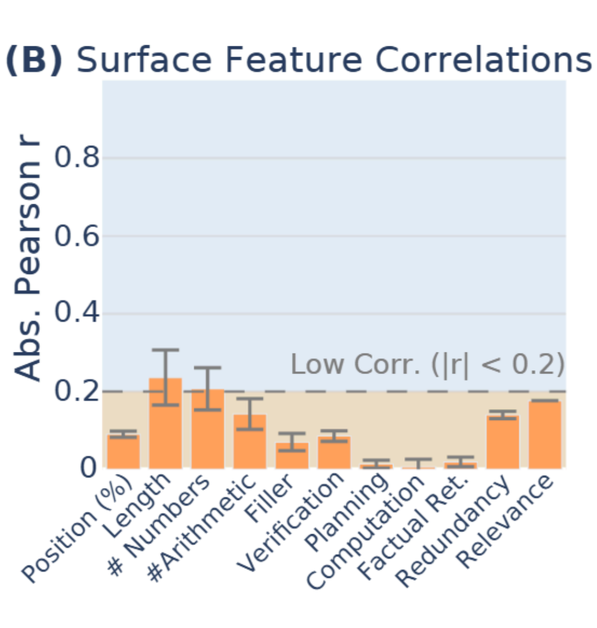
3. Importance is cross-model.
A probe trained on one model's activations predicts another model's importance with high accuracy and agreement.

Cross-model agreement ratios



4. Importance is not surface-level: Importance isn't explained by token-based surface features, nor localized to any single layer or token.





Reasoning Models Know What's Important, and Encode It in Their Activations

Yaniv Nikankin[1], Martin Tutek[2], Tomer Ashuach[1], Jonathan Rosenfeld[3], Yonatan Belinkov[1,4]

[1] TECHNION Israel Institute of Technology
[2] University of Zagreb
[3] Massachusetts Institute of Technology
[4] Kempner INSTITUTE