Measuring the Reliability of Causal Probing Methods: Tradeoffs, Limitations, and the Plight of Nullifying Interventions

Marc E. Canby*, Adam Davies*, Chirag Rastogi, Julia Hockenmaier




*Equal Contribution

University of Illinois Urbana-Champaign

Summary

Which features do LLMs use to perform a given task?

– **Causal interventions** on feature representations

SUMMARY

Which features do LLMs use to perform a given task?

– **Causal interventions** on feature representations

How ***reliable*** are interventions?

– **Completeness:** Is the intended intervention carried out?

– **Selectivity:** Are we damaging non-targeted features?

Summary

Which features do LLMs use to perform a given task?

– **Causal interventions** on feature representations

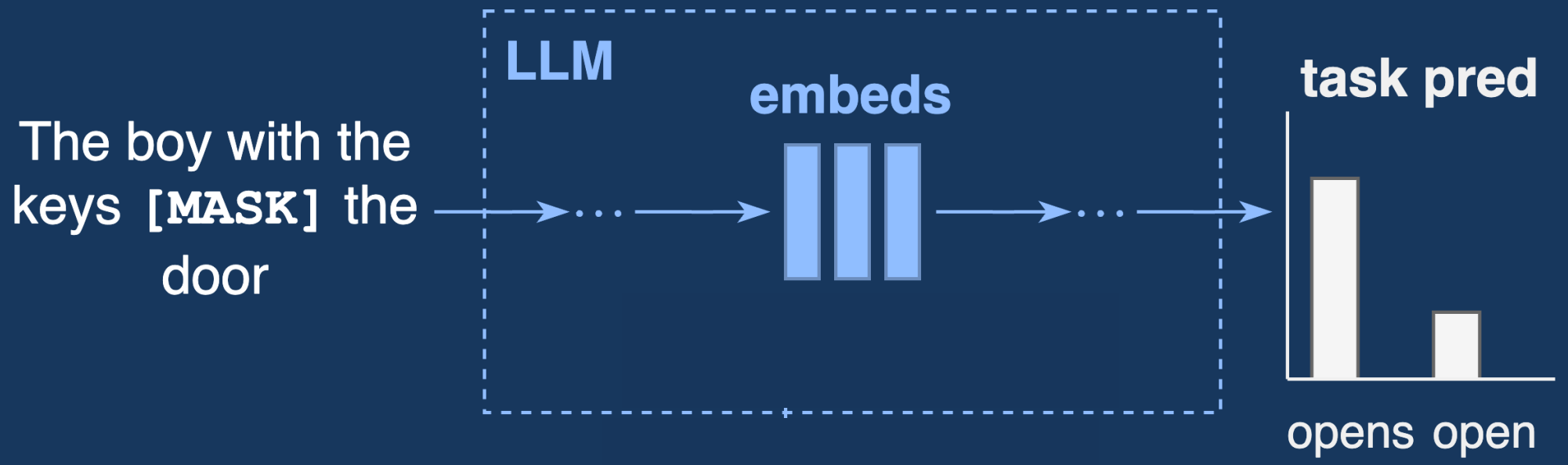
How ***reliable*** are interventions?

– **Completeness:** Is the intended intervention carried out?

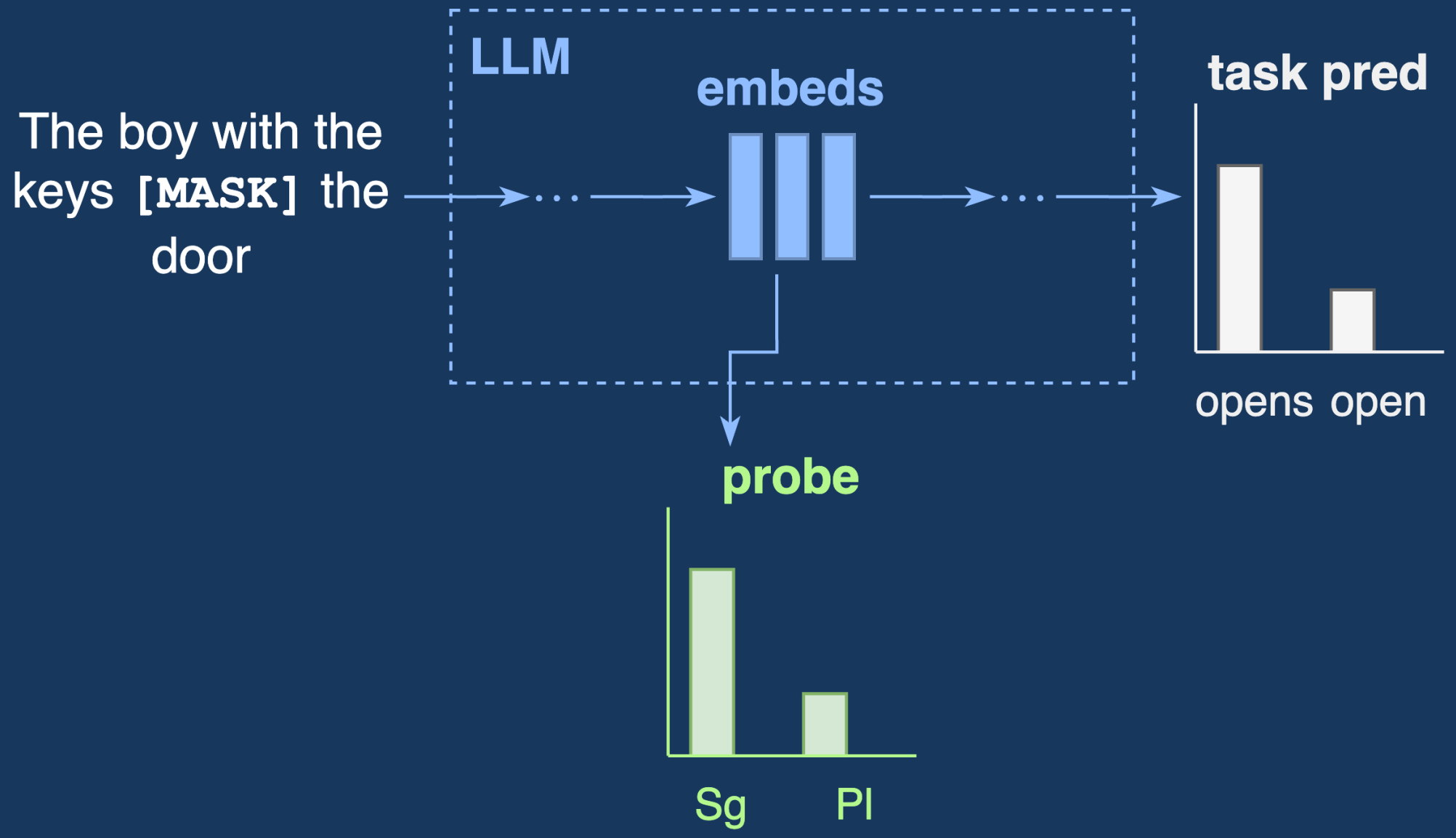
– **Selectivity:** Are we damaging non-targeted features?

We define an **evaluation framework** to compare *different classes* of interventions

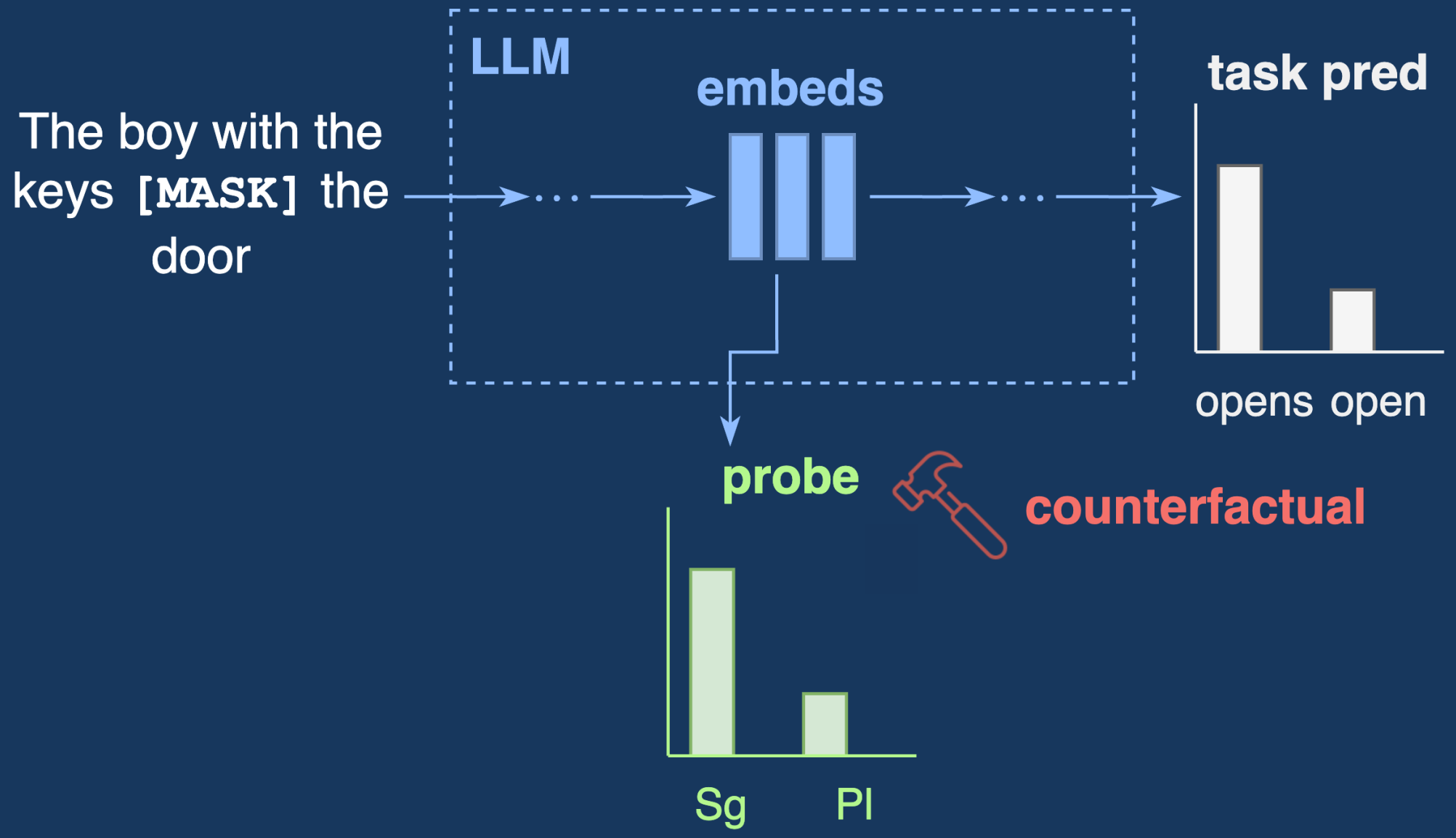
Causal Probing



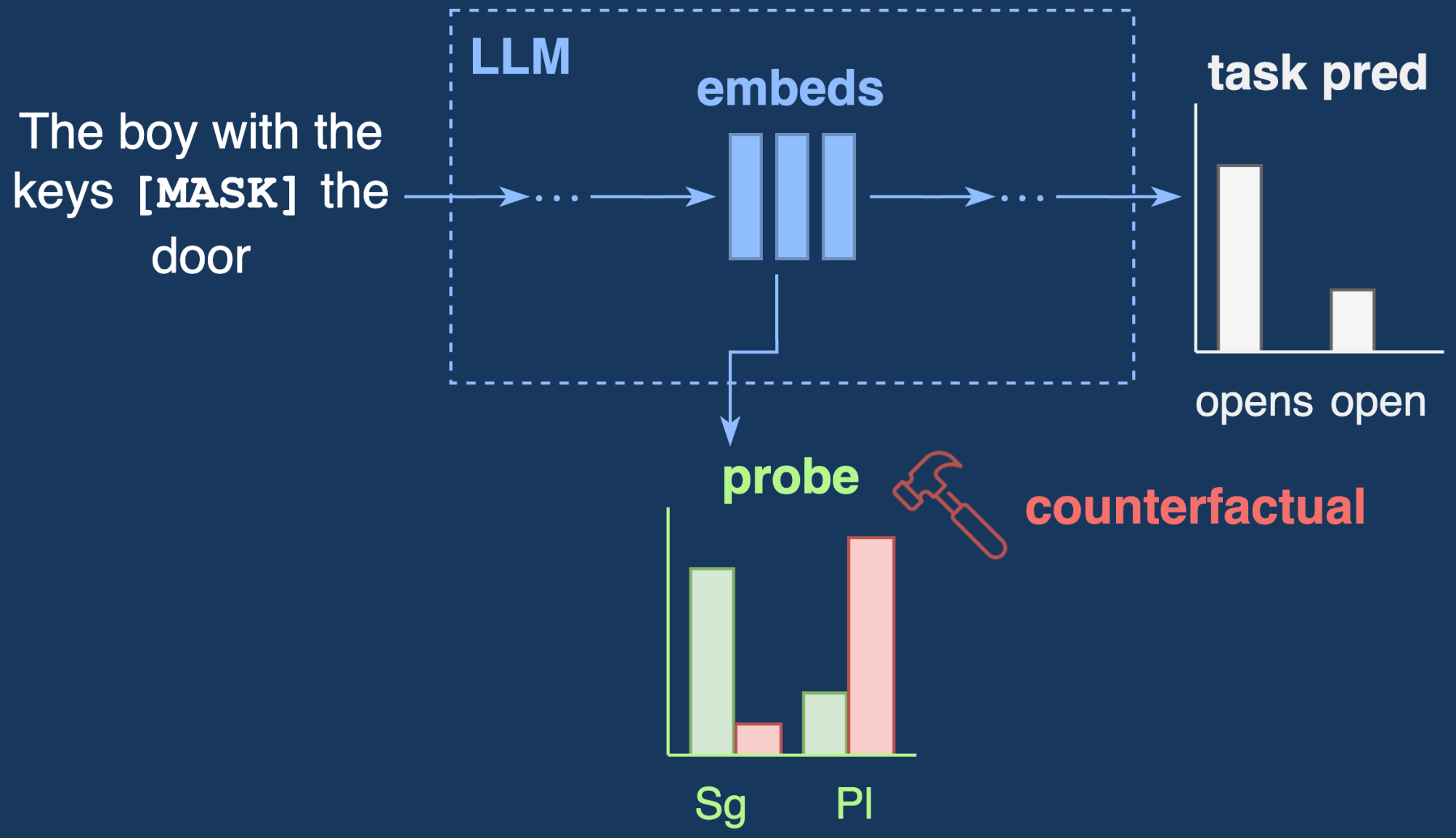
Causal Probing



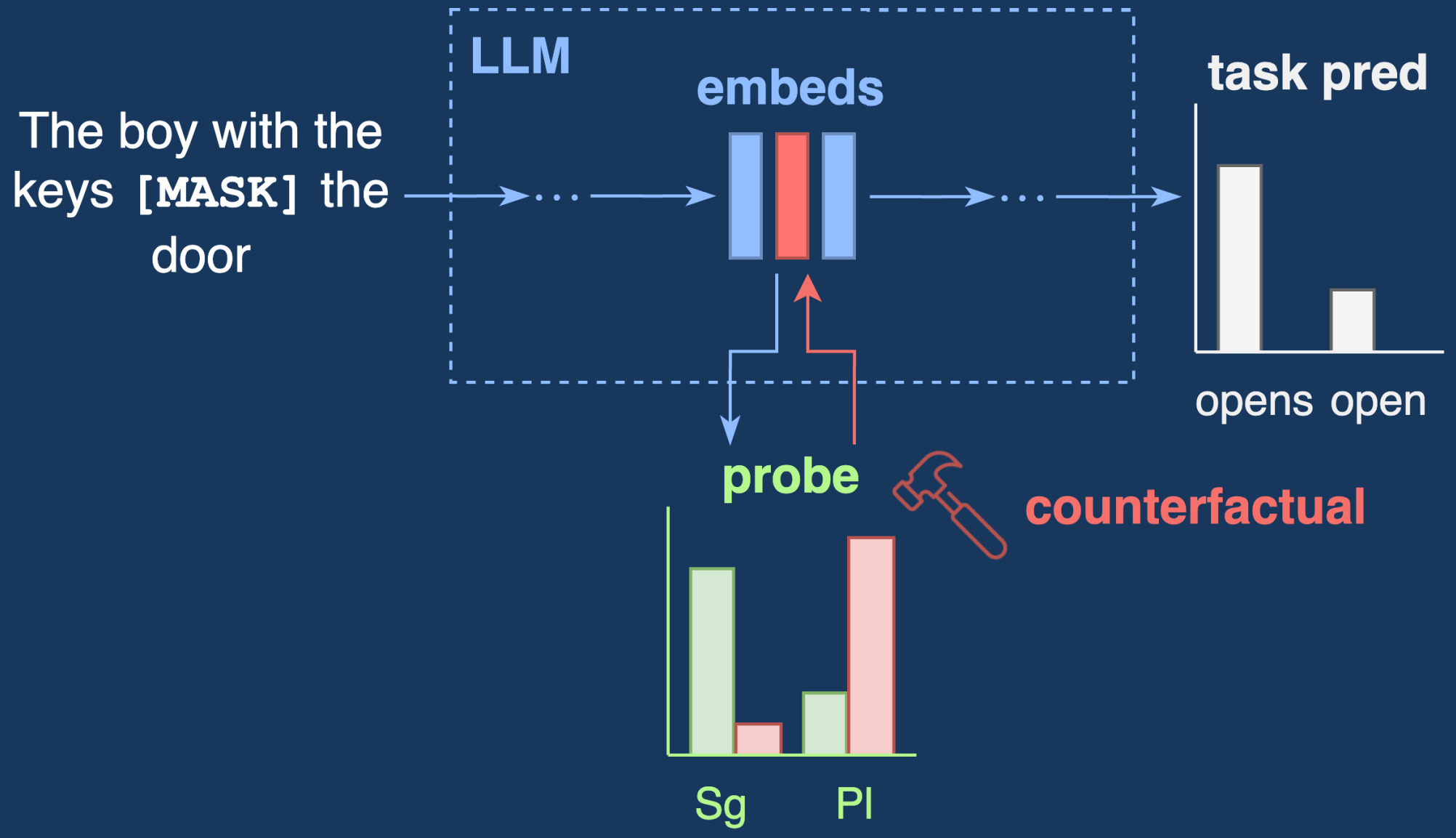
Causal Probing



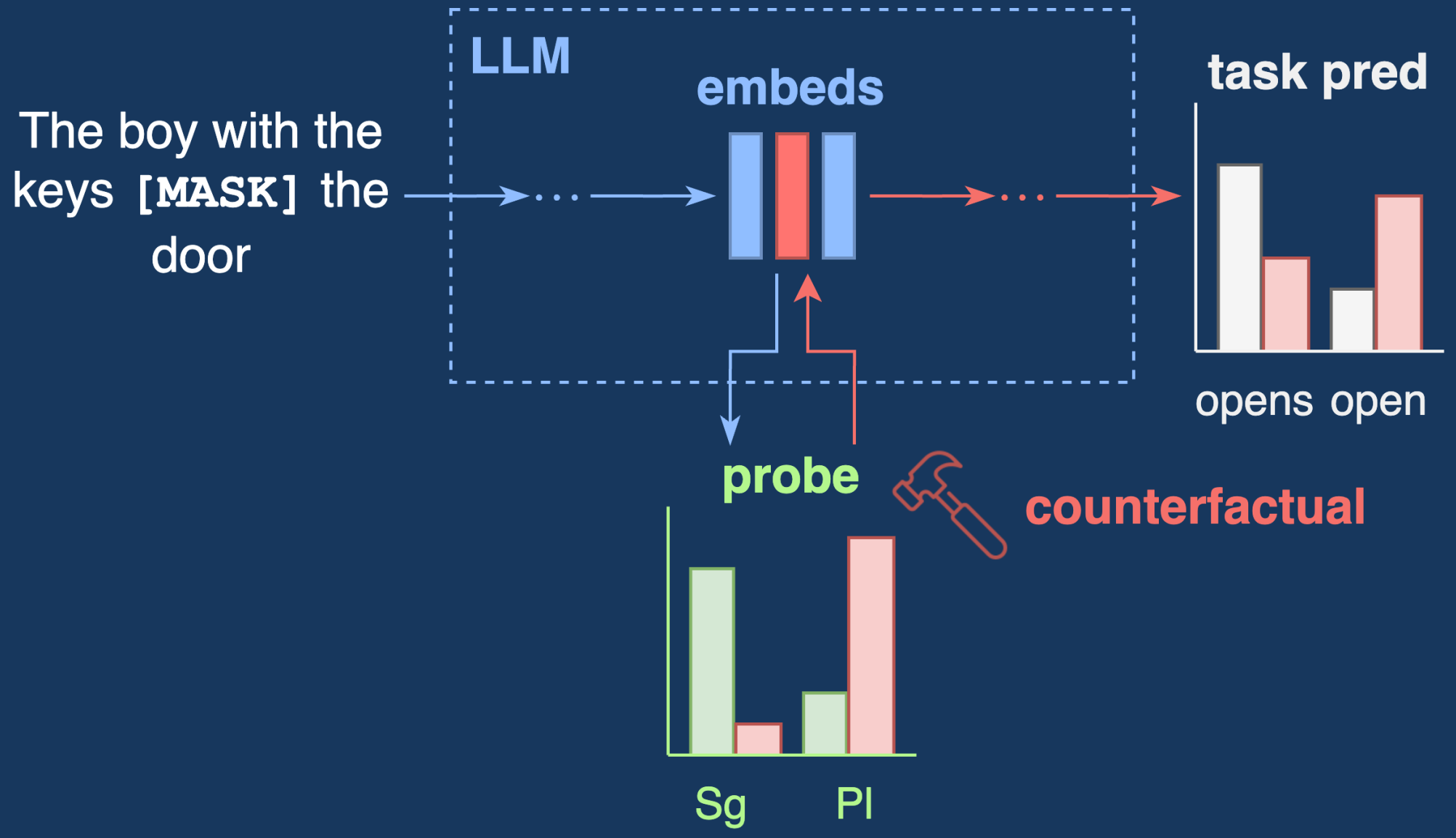
Causal Probing



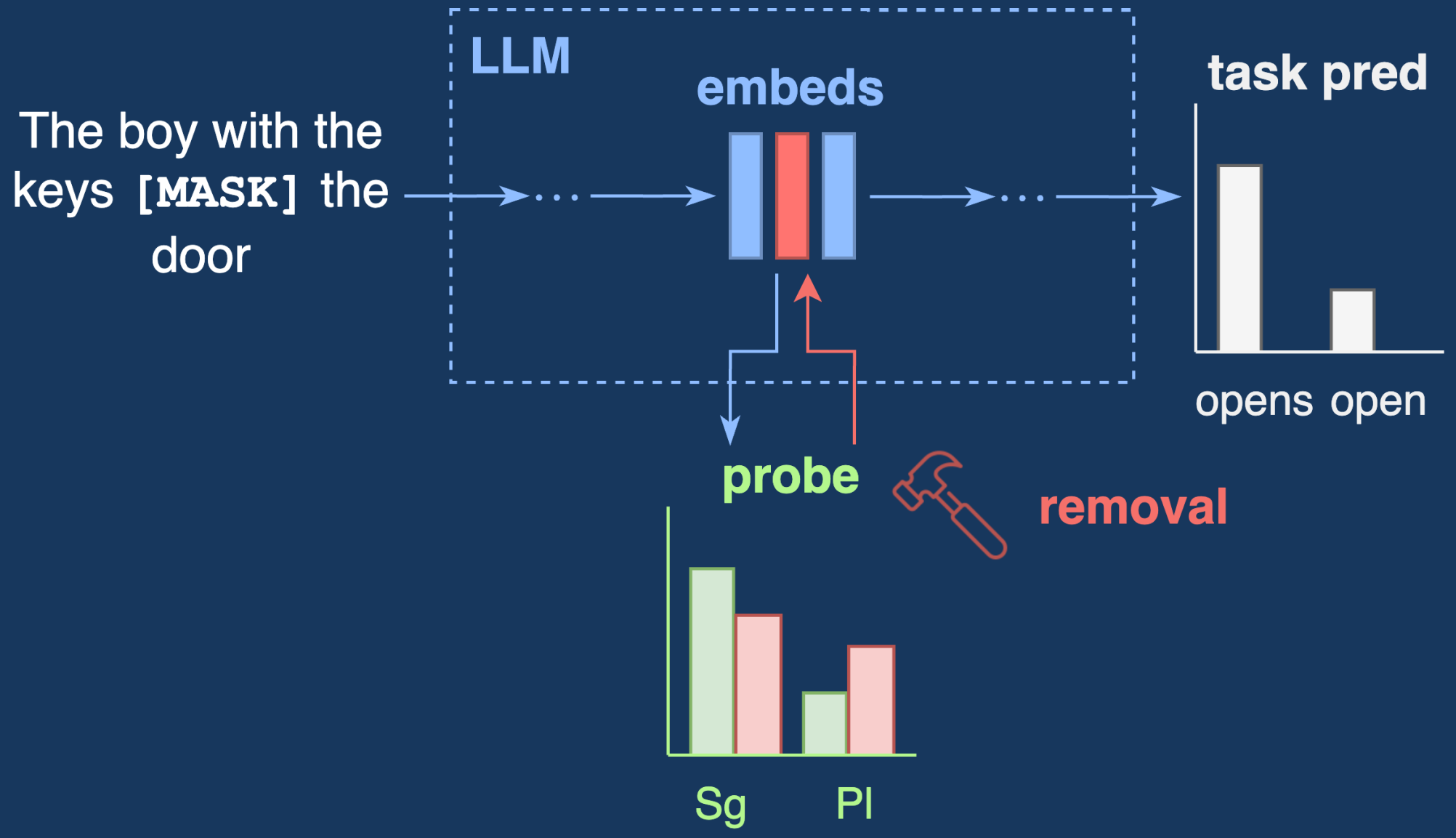
Causal Probing



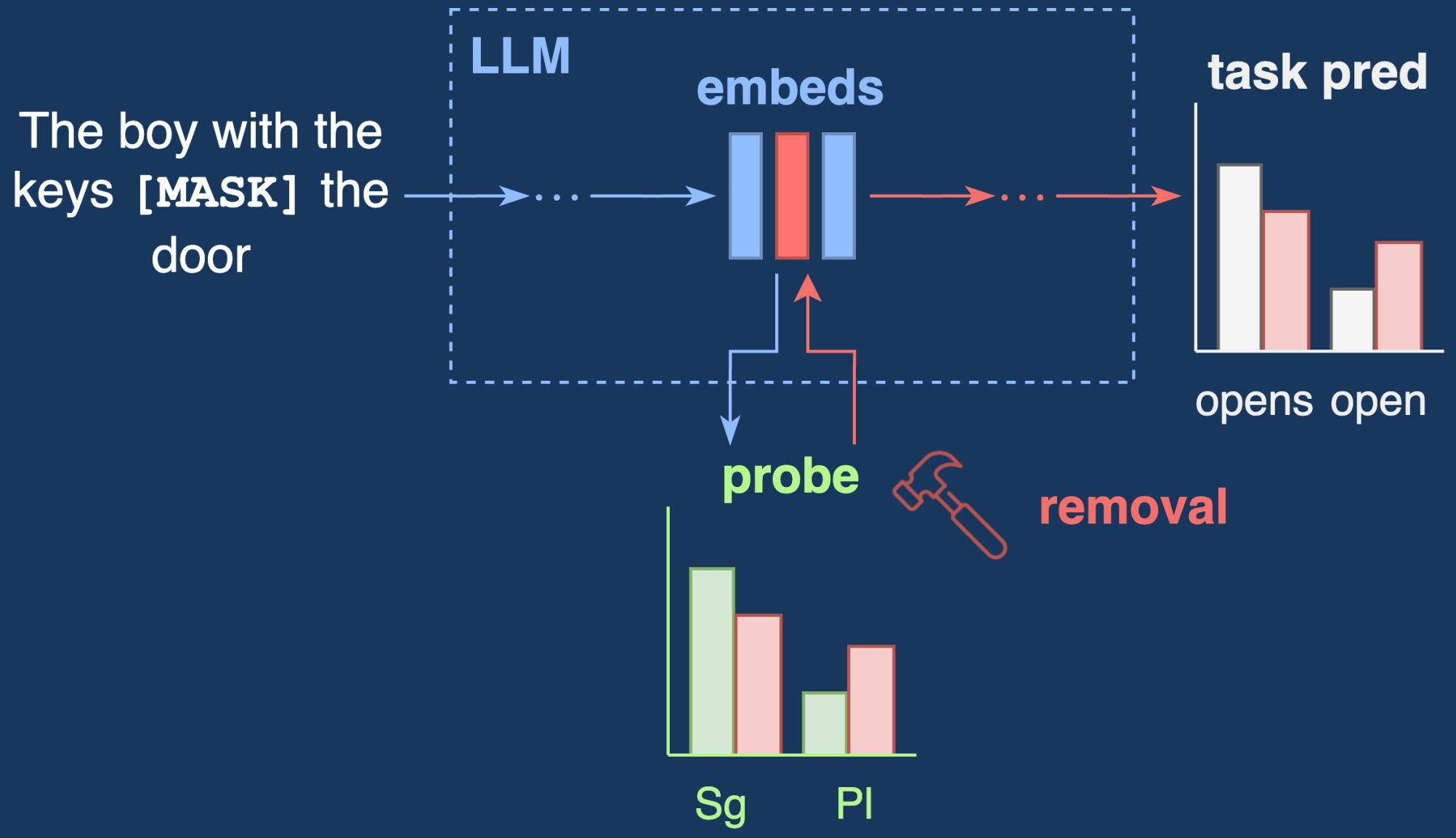
Causal Probing



Causal Probing



Causal Probing

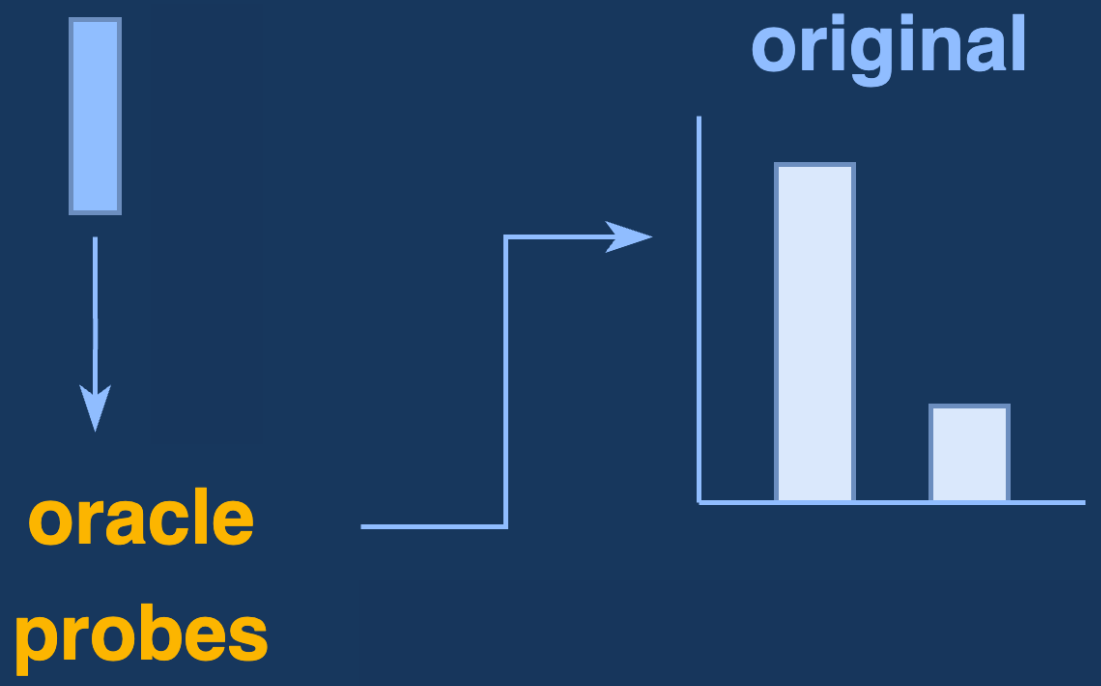


Our Evaluation Framework

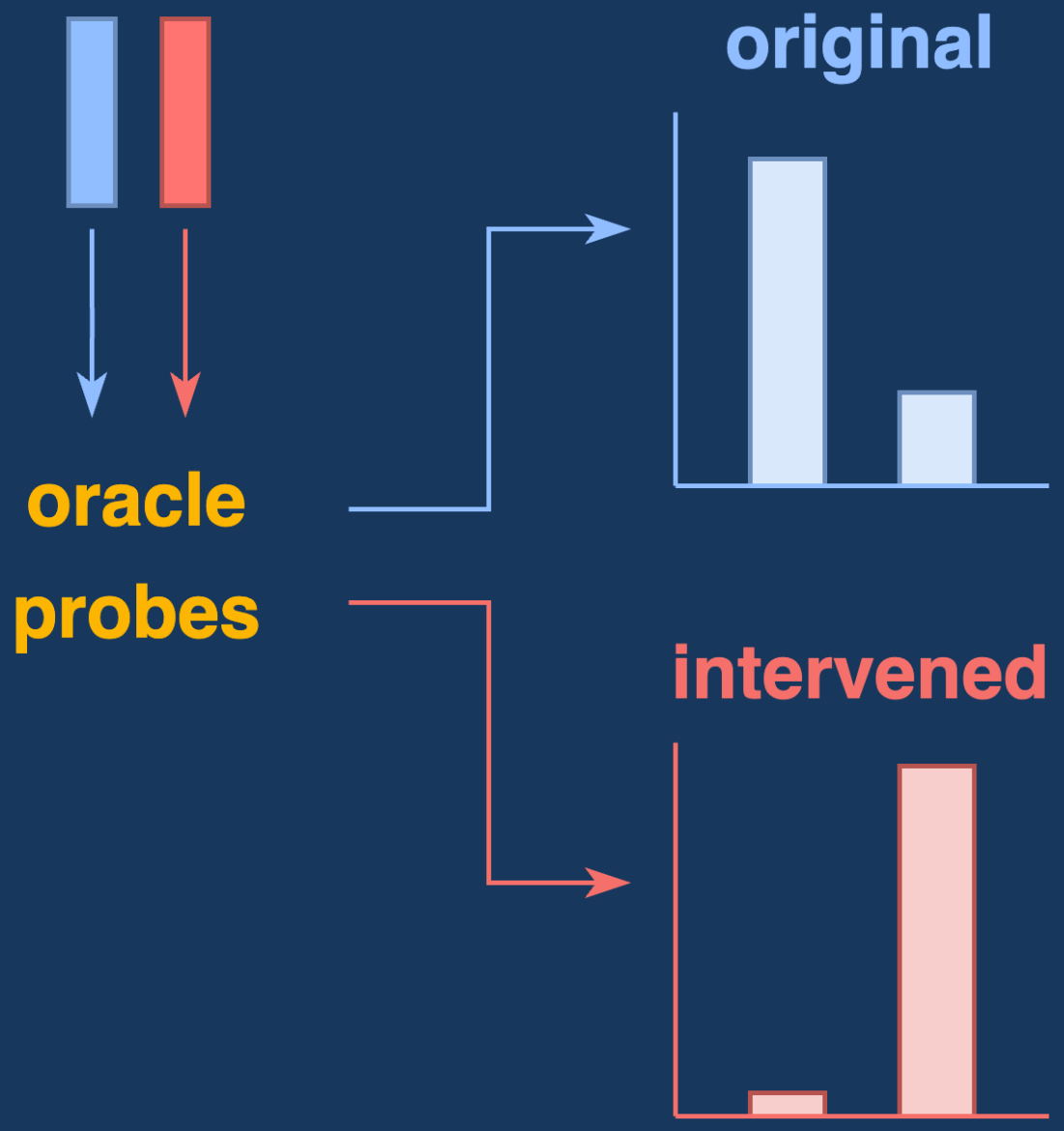
oracle

probes

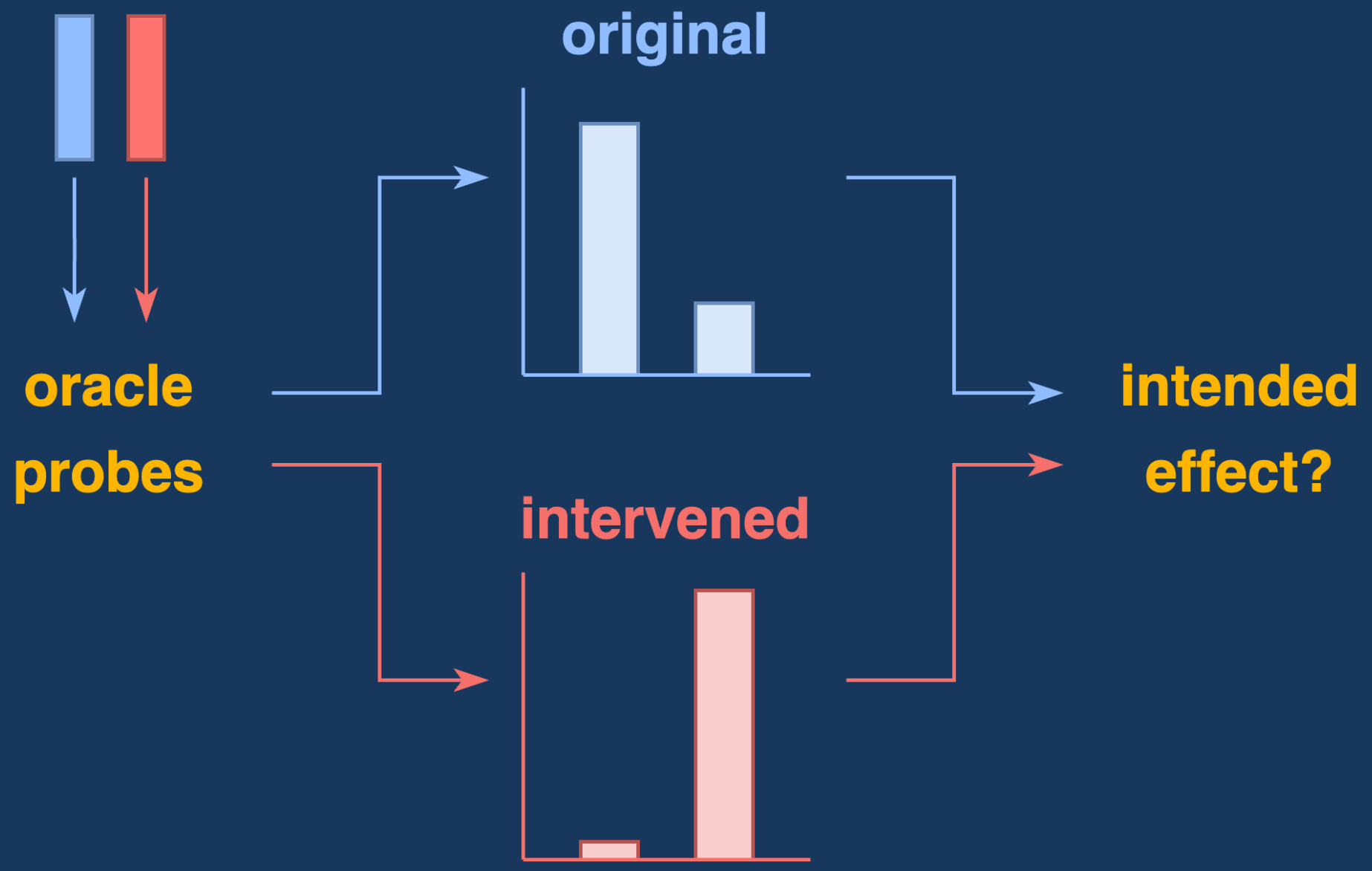
Our Evaluation Framework



Our Evaluation Framework



Our Evaluation Framework



Our Evaluation Framework

completeness

Our Evaluation Framework

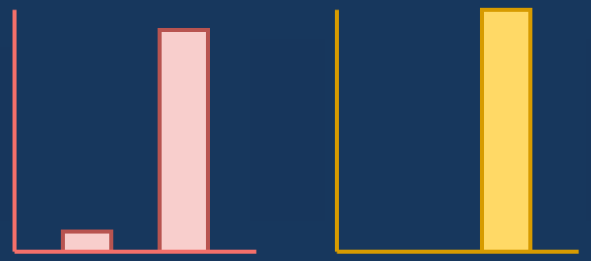
completeness

counterfactual

Our Evaluation Framework

completeness

counterfactual



Our Evaluation Framework

completeness

counterfactual

$1 - \mathrm{dist}\Big($  $,$  $\Big)$

Our Evaluation Framework

completeness

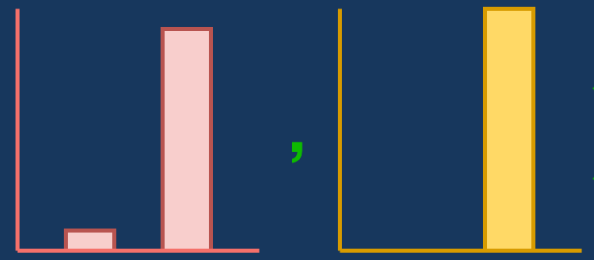
counterfactual

$1 - \mathrm{dist}\Big($  $,$  $\Big) \quad \in [0, 1]$

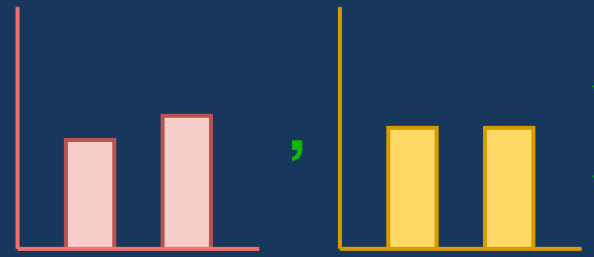
Our Evaluation Framework

completeness

counterfactual

$1 - \mathrm{dist}\Big($  $\Big) \in [0, 1]$

removal

$1 - \mathrm{dist}\Big($  $\Big) \in [0, 1]$

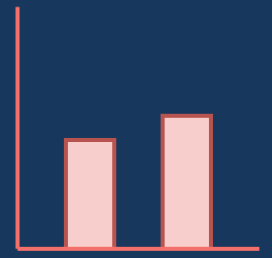
Our Evaluation Framework

completeness

counterfactual

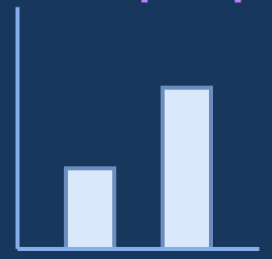
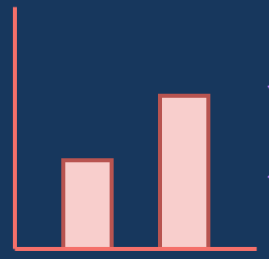
$1 - \mathrm{dist}\Big($  $,$  $\Big) \in [0, 1]$

removal

$1 - \mathrm{dist}\Big($  $,$  $\Big) \in [0, 1]$

selectivity

different properties

$1 - \mathrm{dist}\Big($  $,$  $\Big) \in [0, 1]$

Causal Probing Interventions

Concept Removal: *Remove* representation of target feature

- **Linear:** INLP, RLACE

Causal Probing Interventions

Concept Removal: *Remove* representation of target feature

- **Linear:** INLP, RLACE

Counterfactual: *Swap* representation from one value to another

- **Linear:** AlterRep
- **Nonlinear (GBIs):** FGSM, PGD, AutoAttack

Causal Probing Interventions

Concept Removal: *Remove* representation of target feature

- **Linear:** INLP, RLACE

Counterfactual: *Swap* representation from one value to another

- **Linear:** AlterRep
- **Nonlinear (GBIs):** FGSM, PGD, AutoAttack

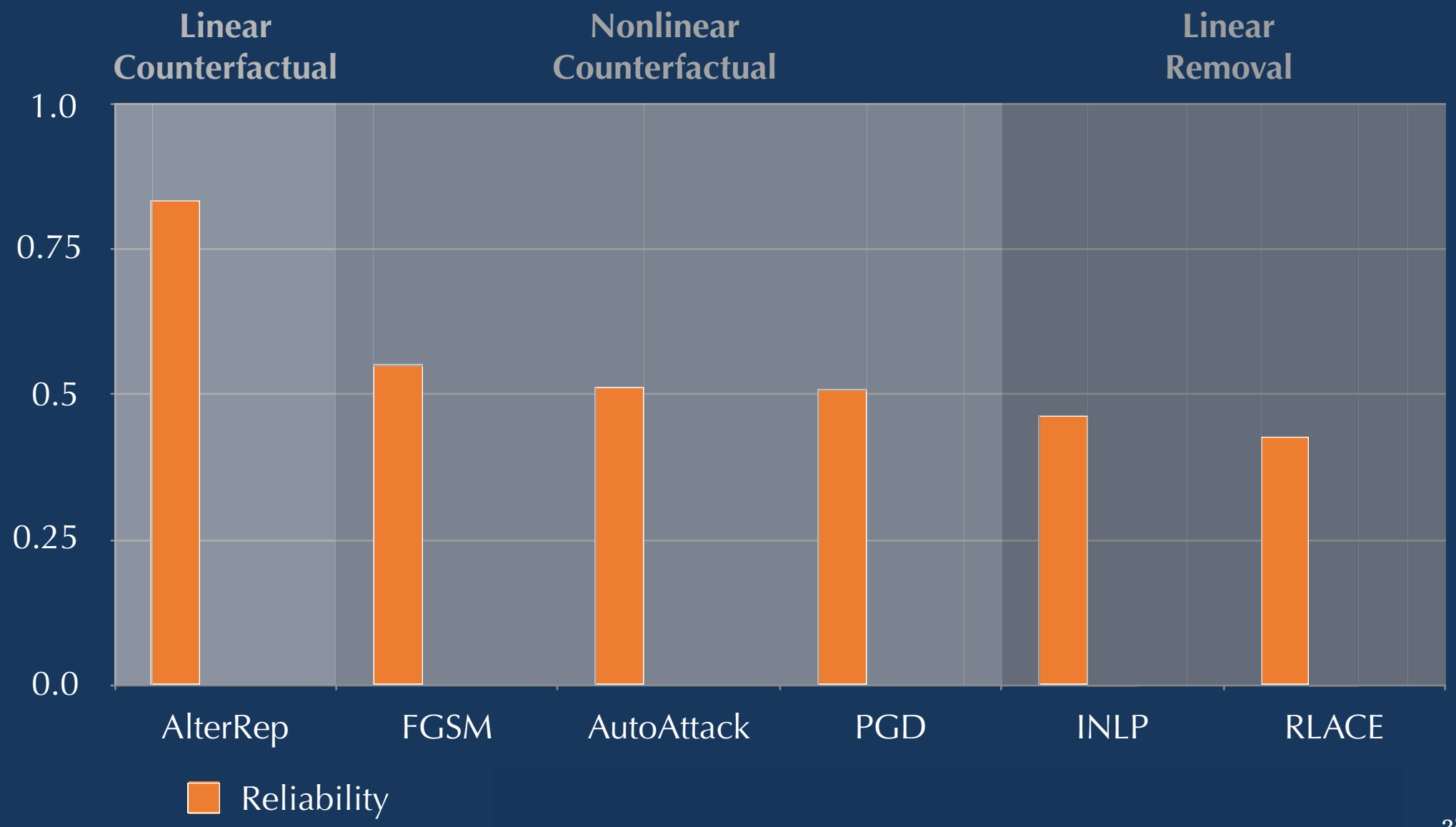
Intervene on `[MASK]` token in final layer of BERT

Our Evaluation Framework

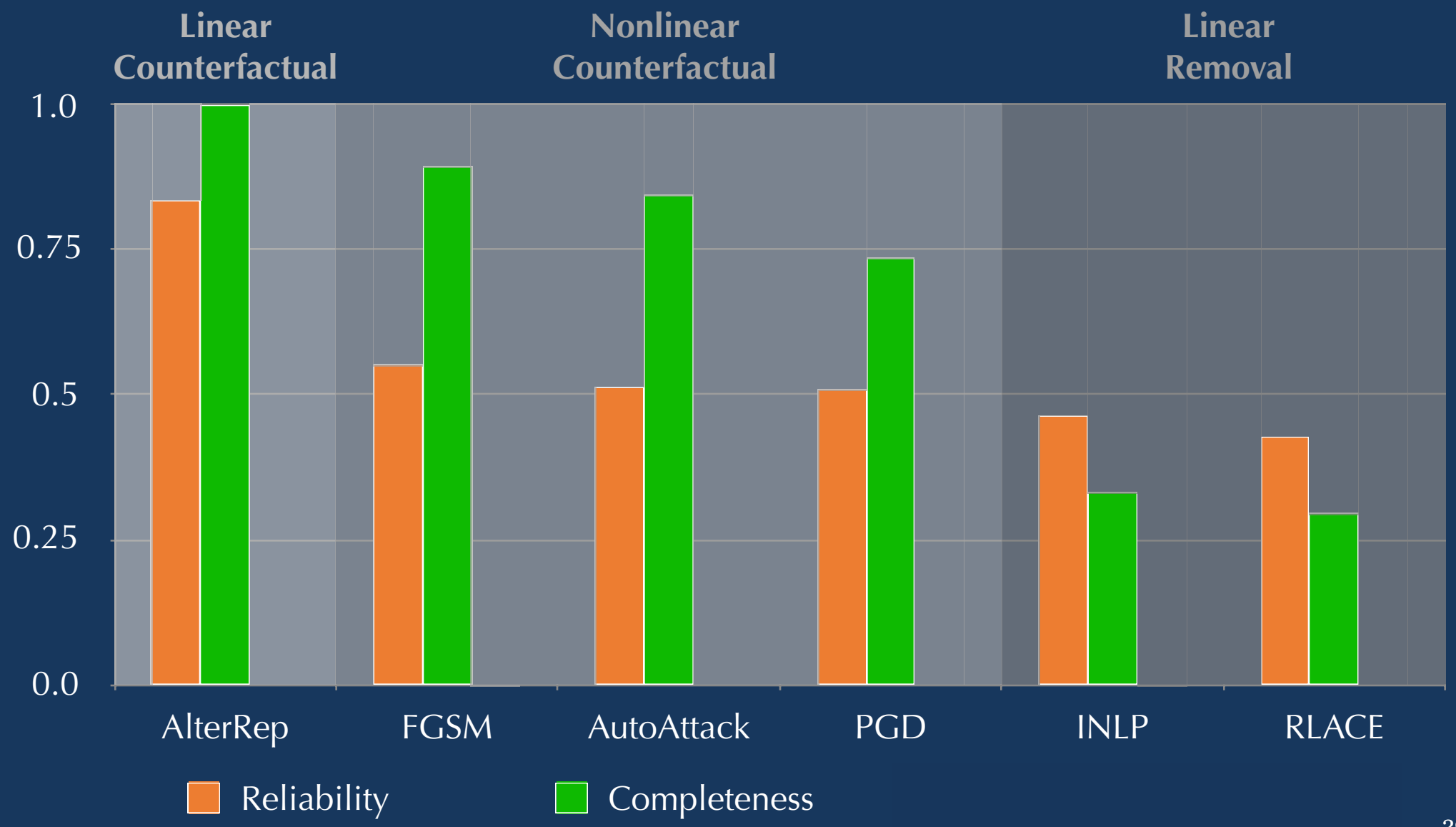
Evaluate interventions according to

- **Completeness:** Is the intended intervention carried out?
- **Selectivity:** Are we damaging non-targeted features?
- **Reliability:** Harmonic mean of completeness and selectivity

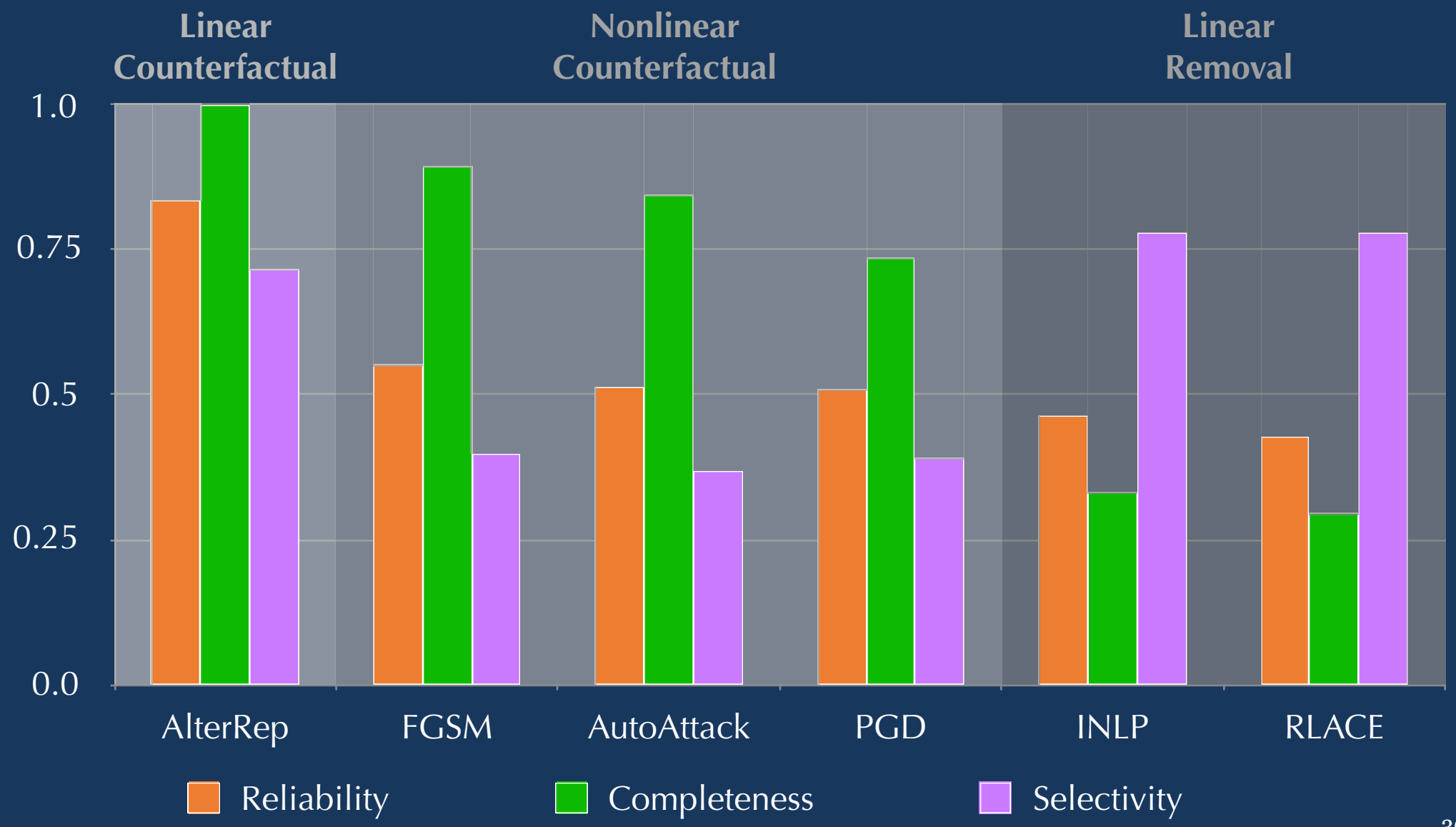
Reliability: AlterRep is most reliable!



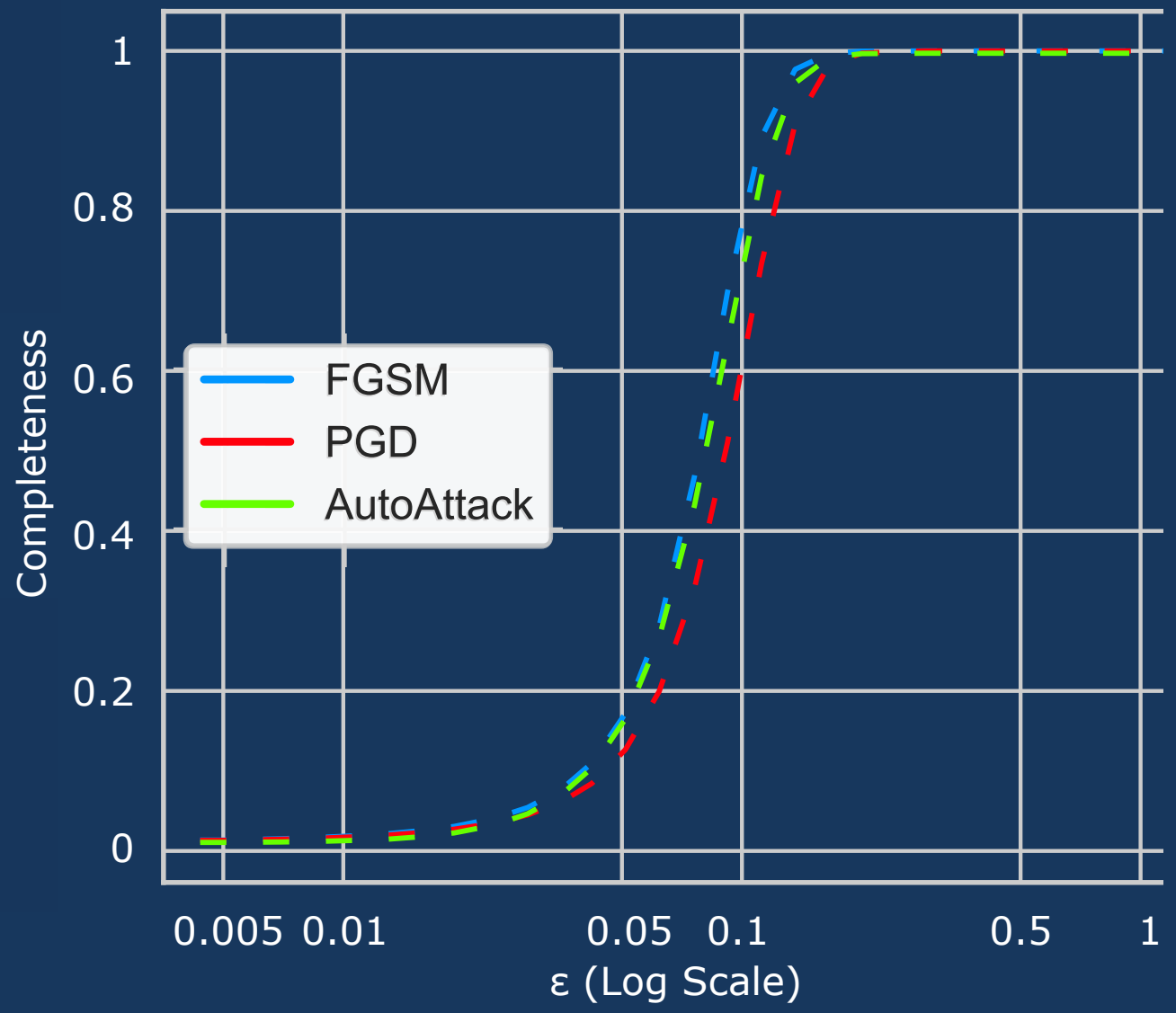
Completeness: Removal Methods are not complete!



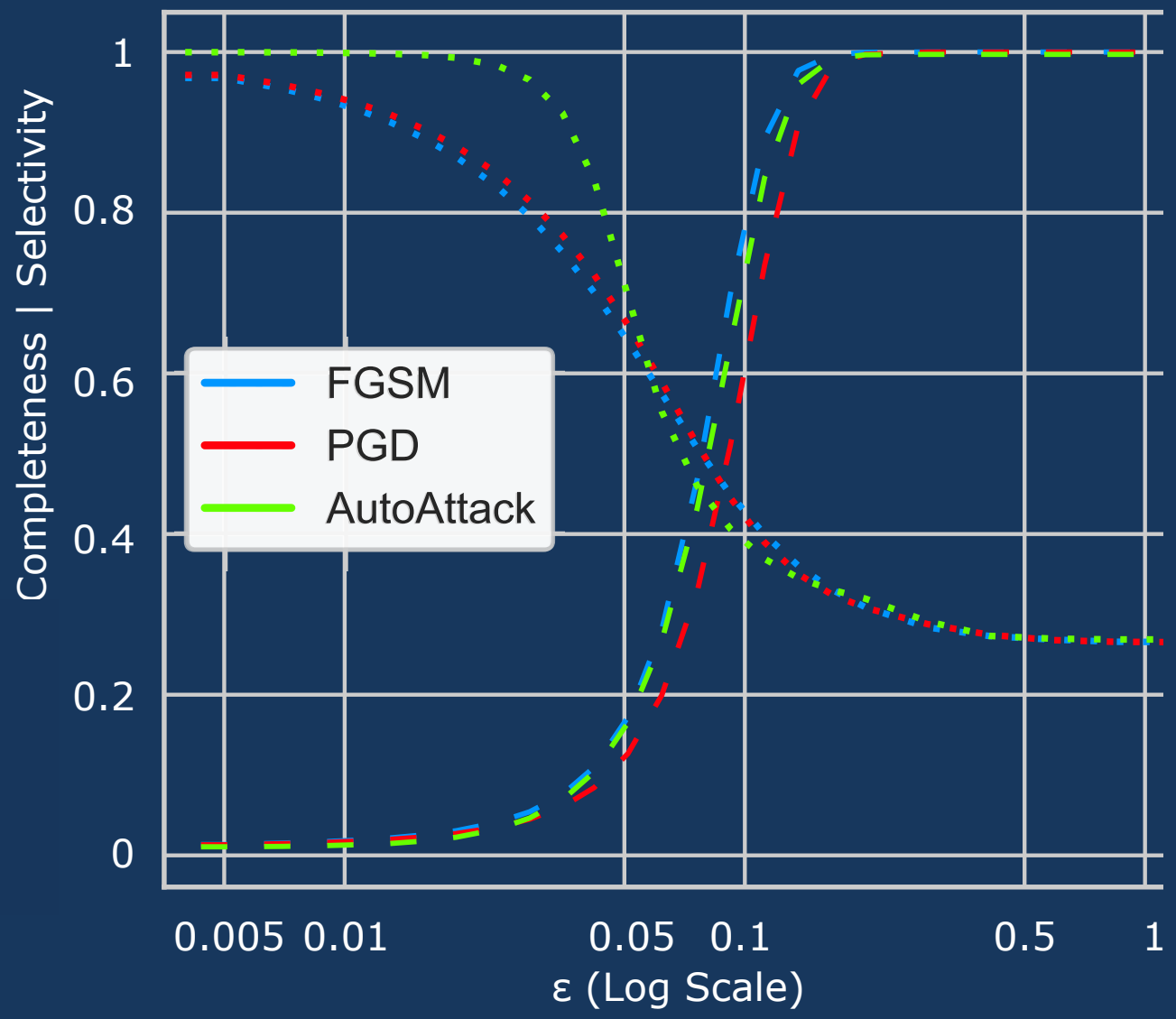
SELECTIVITY: GBIs ARE NOT SELECTIVE; LINEAR METHODS ARE



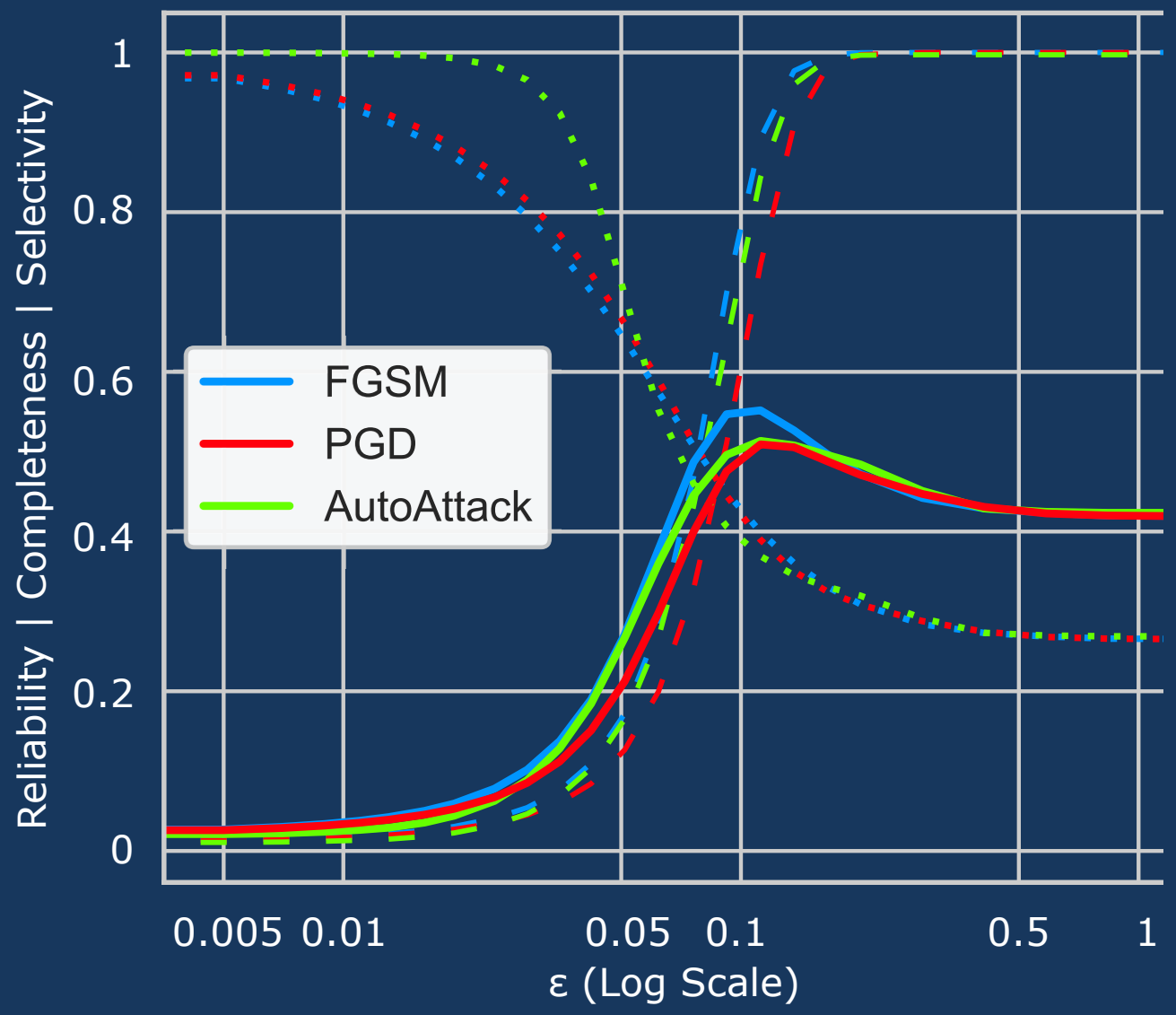
Completeness and Selectivity are a tradeoff!



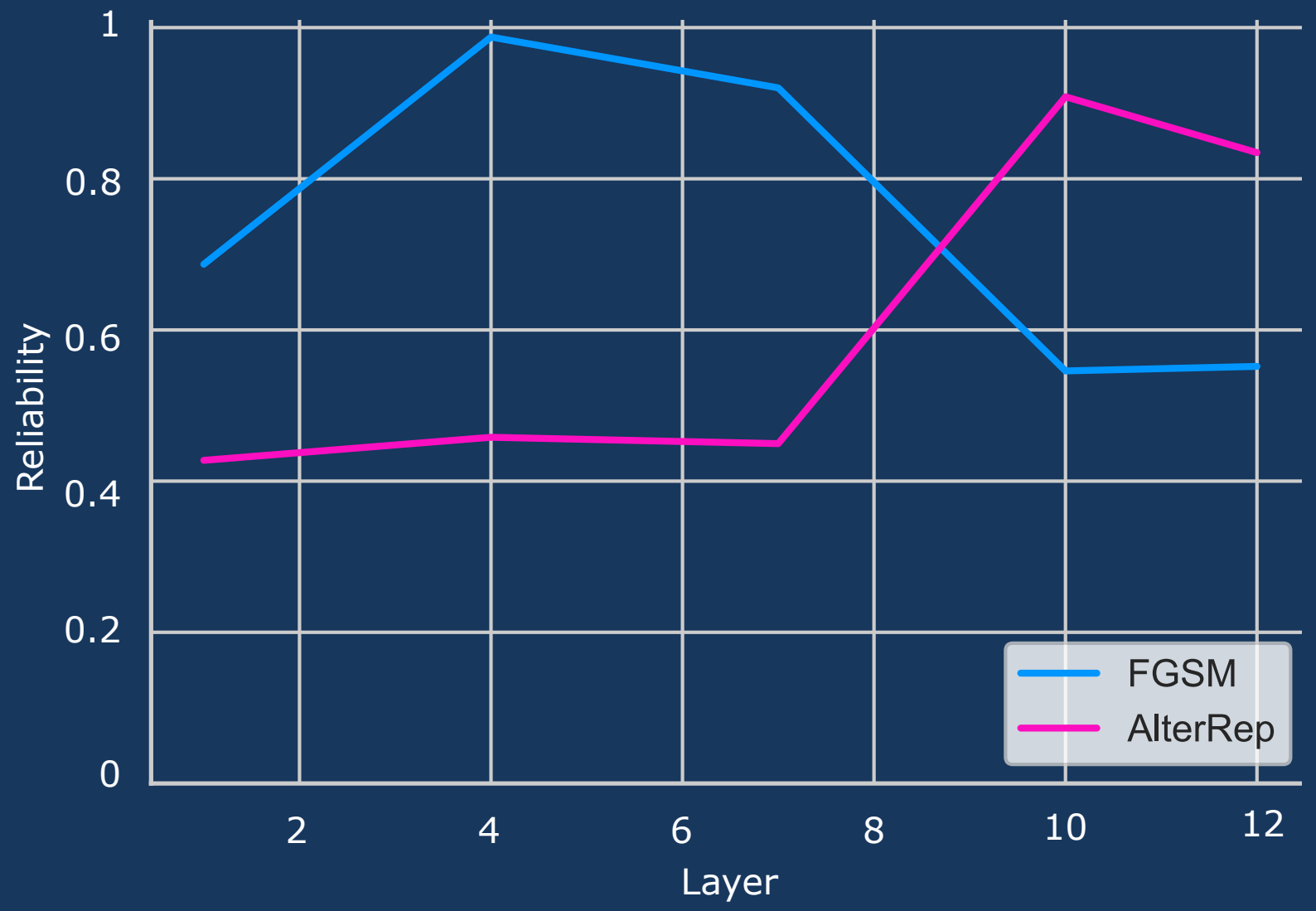
Completeness and Selectivity are a tradeoff!



Completeness and Selectivity are a tradeoff!



FGSM (NONLINEAR) MORE RELIABLE IN EARLIER LAYERS
ALTERREP (LINEAR) MORE RELIABLE IN LATER LAYERS



Takeaways

We introduce an **evaluation framework** to compare different classes of causal probing interventions

- **Tradeoff** between *completeness* and *selectivity*
- Concept removal is **not reliable** (for causal probing)
- Linear interventions better in later layers (less collateral damage)

Questions?

Supplementary Materials for QA

More Reliable methods → Greater Δ in task accuracy