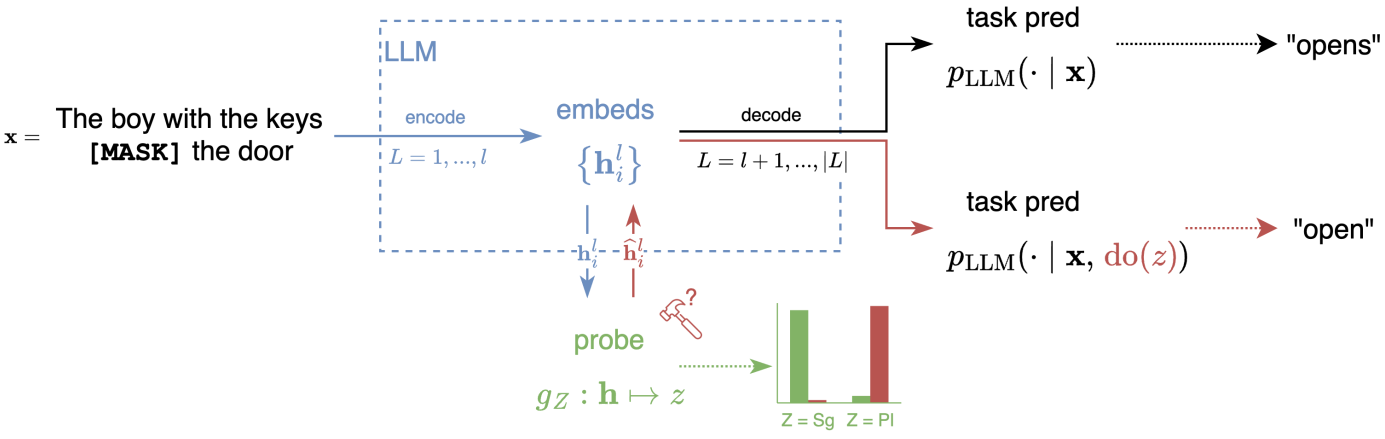
Measuring the Reliability of Causal Probing Methods: Tradeoffs, Limitations, and the Plight of Nullifying Interventions

Marc E. Canby*, Adam Davies*, Chirag Rastogi, Julia Hockenmaier
*Equal contribution University of Illinois Urbana-Champaign

How *reliable* are interp methods?


$\mathbf{x} =$ The boy with the keys **[MASK]** the door

LLM

encode $L = 1, ..., l$

embeds $\{\mathbf{h}_i^l\}$

decode $L = l+1, ..., |L|$

task pred $p_{\text{LLM}}(\cdot \mid \mathbf{x})$ "opens"

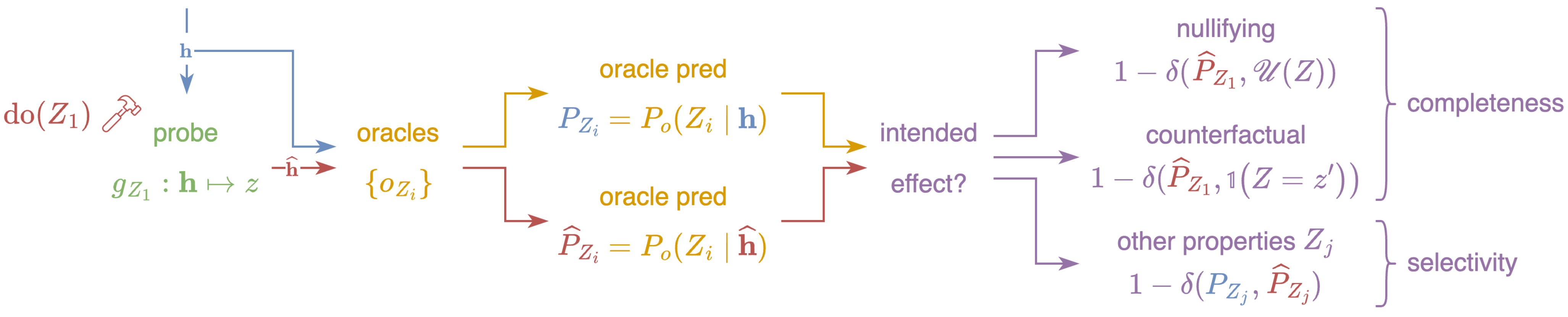
task pred $p_{\text{LLM}}(\cdot \mid \mathbf{x}, \text{do}(z))$ "open"

$\mathbf{h}_i^l$ $\widehat{\mathbf{h}}_i^l$

probe $g_Z : \mathbf{h} \mapsto z$

Completeness: How much is targeted property damaged?

Selectivity: How little collateral damage is done?


$\mathbf{h}$

do(Z_1)

probe $g_{Z_1} : \mathbf{h} \mapsto z$

$\widehat{\mathbf{h}}$

oracles $\{o_{Z_i}\}$

oracle pred $P_{Z_i} = P_o(Z_i \mid \mathbf{h})$

oracle pred $\widehat{P}_{Z_i} = P_o(Z_i \mid \widehat{\mathbf{h}})$

intended effect?

nullifying $1 - \delta(\widehat{P}_{Z_1}, \mathscr{U}(Z))$

counterfactual $1 - \delta(\widehat{P}_{Z_1}, \mathbb{1}(Z = z'))$

completeness

other properties Z_j $1 - \delta(P_{Z_j}, \widehat{P}_{Z_j})$

selectivity

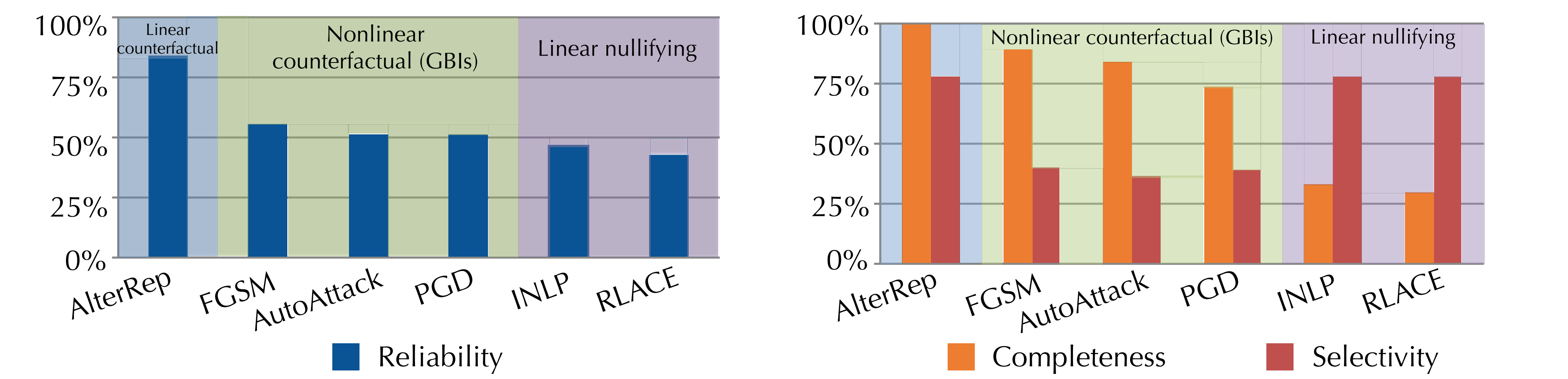
Experiments: Subject-verb agreement

`The boy with the keys [MASK] the door` → `open` vs. `opens`

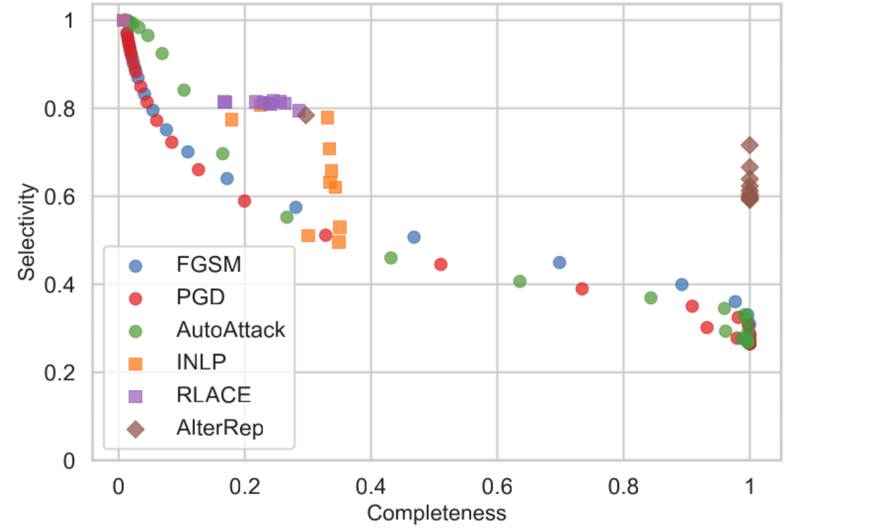
Z_1 = # subject ("`boy`") Z_j = # distractor ("`keys`")

How reliable are interventions on `[MASK]` vector in final layer of BERT?

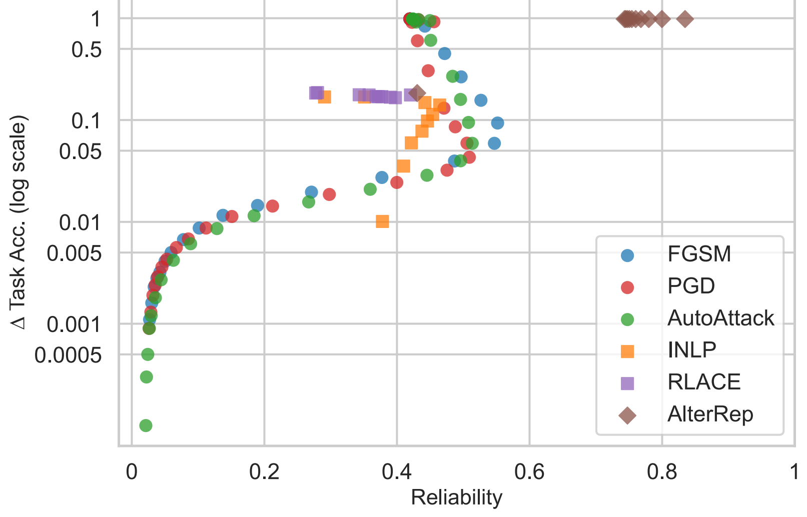
How reliable are causal probing interventions? Nullifying methods are not reliable!



Completeness & selectivity are a tradeoff!



Greater reliability → Greater Δ Task Acc.


Δ Task Acc. (log scale)
Reliability
FGSM
PGD
AutoAttack
INLP
RLACE
AlterRep