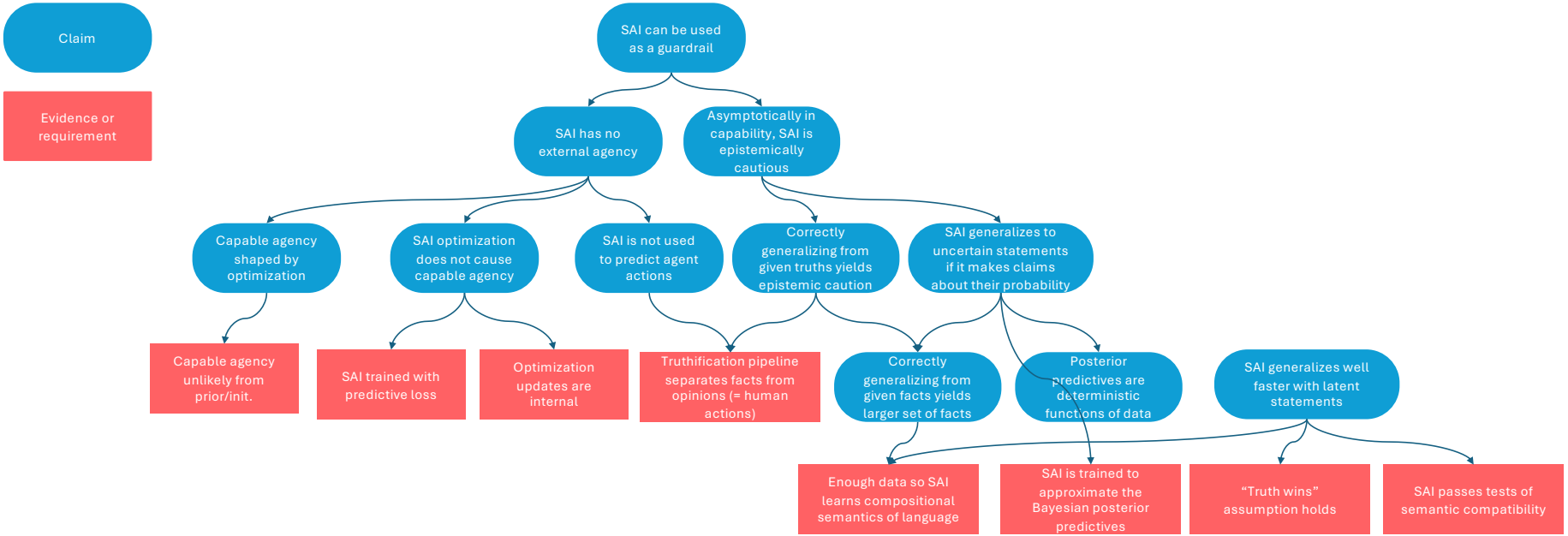
Claim
Evidence or requirement
SAI can be used as a guardrail
SAI has no external agency
Asymptotically in capability, SAI is epistemically cautious
Capable agency shaped by optimization
SAI optimization does not cause capable agency
SAI is not used to predict agent actions
Correctly generalizing from given truths yields epistemic caution
SAI generalizes to uncertain statements if it makes claims about their probability
Capable agency unlikely from prior/init.
SAI trained with predictive loss
Optimization updates are internal
Truthification pipeline separates facts from opinions (= human actions)
Correctly generalizing from given facts yields larger set of facts
Posterior predictives are deterministic functions of data
SAI generalizes well faster with latent statements
Enough data so SAI learns compositional semantics of language
SAI is trained to approximate the Bayesian posterior predictives
"Truth wins" assumption holds
SAI passes tests of semantic compatibility