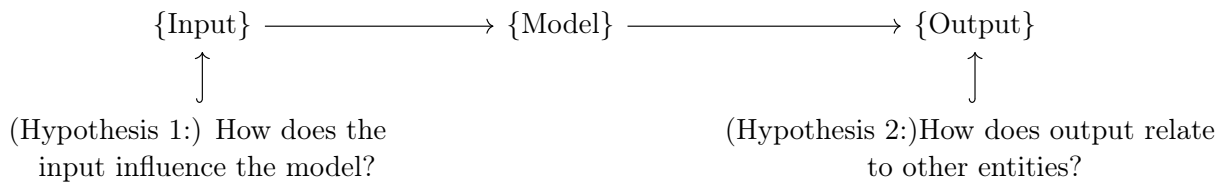
SeriMATs: Response to Jessica Rumbelow's Question 1. In this exploration, we adopt a category theory perspective[1] - a key area of my research - to understand machine learning models. Our approach is predicated on the notion that a model's *properties* is revealed through its interactions with outer world.



General Model Properties:

- *Hypothesis 1: Task-specific Design.* The model is likely tailored for specific tasks. We can ascertain this by inputting a variety of machine learning tasks to observe its performance across domains like image classification and sentiment analysis. This approach acknowledges that our understanding of the model's capabilities might be incomplete[2].
- *Hypothesis 1b: Access to Partial Outputs.* If only partial outputs are accessible, they warrant a detailed analysis to decipher their implications, leading back to Hypothesis 1.
- *Hypothesis 2: Output Utility.* Outputs might be instrumental in representation learning, enabling further insights into the model's properties.

Now that we have understood general properties: we may have a good idea have the type of inputs that really affects the outputs. In this case:
Domain-specific and Non Domain-specific Analysis:

- *Domain-specific Challenges.* Designing domain-specific problems (e.g., in sentiment analysis) helps identify the model's linguistic understanding, such as semantic and syntactic comprehension.
- *Task Perturbation.* Investigating the model's responses to input perturbations reveals its robustness

Lastly, we point on analysis with gradient access: Access to gradient information enriches the analysis, particularly by quantifying the 'magnitude of effect.' This could involve counterfactual reasoning or robustness tests, providing deeper insights into how inputs influence outputs and the model's internal mechanics.

[1] Under category theory, focus is not on the substance of objects but on their relations with other objects.

[2] Complete knowledge of all the tasks a model can perform may be unattainable.