Reliability in Systems of Agents

Nathan Rigoni
AI/ML Engineer Staff
Generative AI Team Lead
Lockheed Martin

Objective

Examine the metrics we use to determine the expected failure rate of LLM agents in performing tasks.

- Definitions
- Context
- Failure Modes
- Benchmarks

The real objective is to start a conversation

What is an agent?

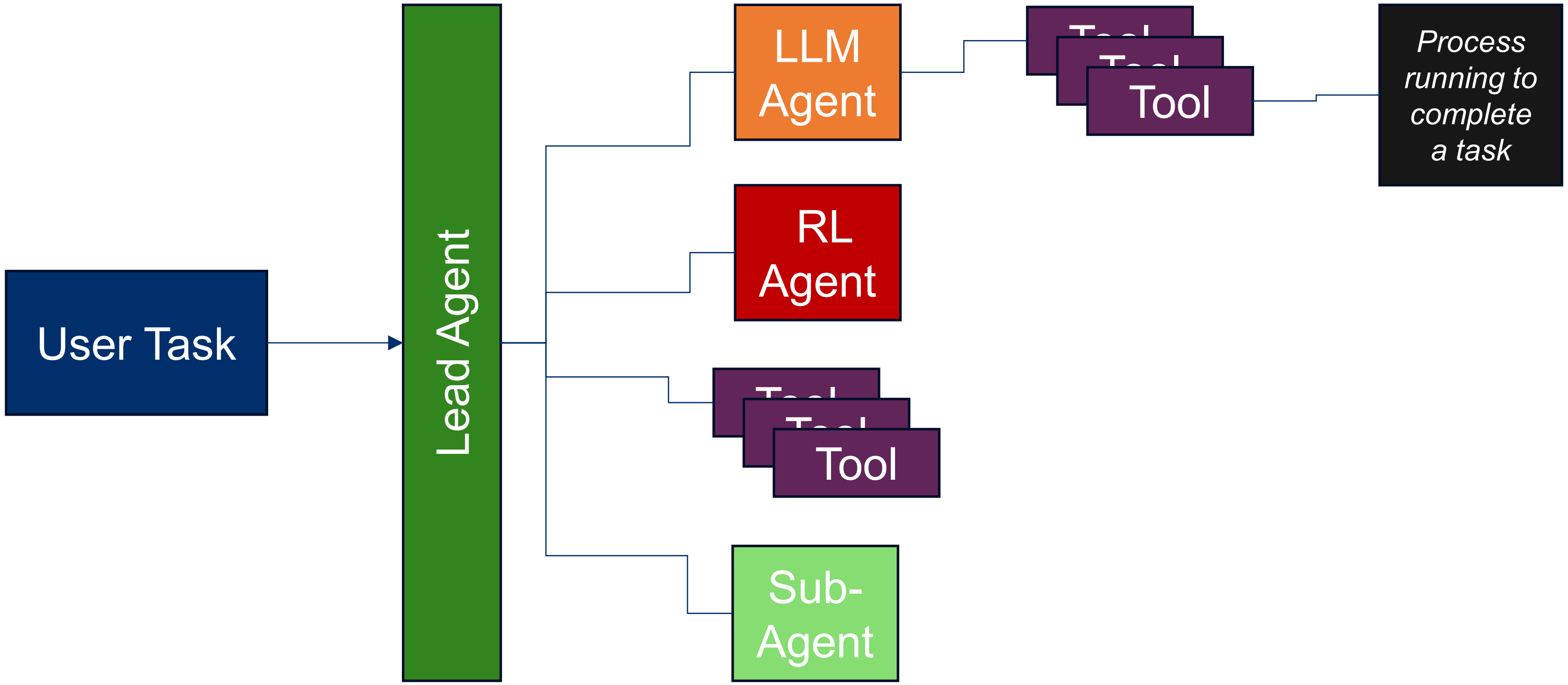
An artificial intelligence (AI) agent is a software program that can interact with its environment, collect data, and use the data to perform self-determined tasks to meet predetermined goals.

Types of Agents:

- Large Language Model Agents (LLM + tools)
- Reinforcement Learning (RL) Agents
- Chat-bots (LLM no tools)

Focusing on LLM agents in this presentation

System of Agents (example)



DO NOT EXTRAPOLATE THESE OBSERVATIONS TO ALL OF AI!

We are only examining LLM Agents

What type of system is language?

Nondeterministic

(philosophy) *The opposite of determinism: the doctrine that there are factors other than the state and immutable laws of the universe involved in the unfolding of events, such as free will.*

(computing) *Dependence on factors other than initial state and input.*

(computer science) *The property of being nondeterministic, involving arbitrary choices; necessitating the choice between various indistinguishable possibilities.*

Language as a component of communication relies on interpretation, which is informed by culture, experience, and other details and information not present in the initial state or the input. This makes language a nondeterministic system.

"You keep using that word…
I do not think it means what you think it means"
- Inigo Montoya

What can we know?

What is a failure?
A binary bit labeling a contextually undesired state or action.

What is Information?

Context

“The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. **These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages.** The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.” - C.E Shannon *A Mathematical Theory of Communication*

Bit

A bit is the most basic unit of information in computing and digital communication. **Bi**nary Digi**t**. The bit represents a logical state between two possible values.

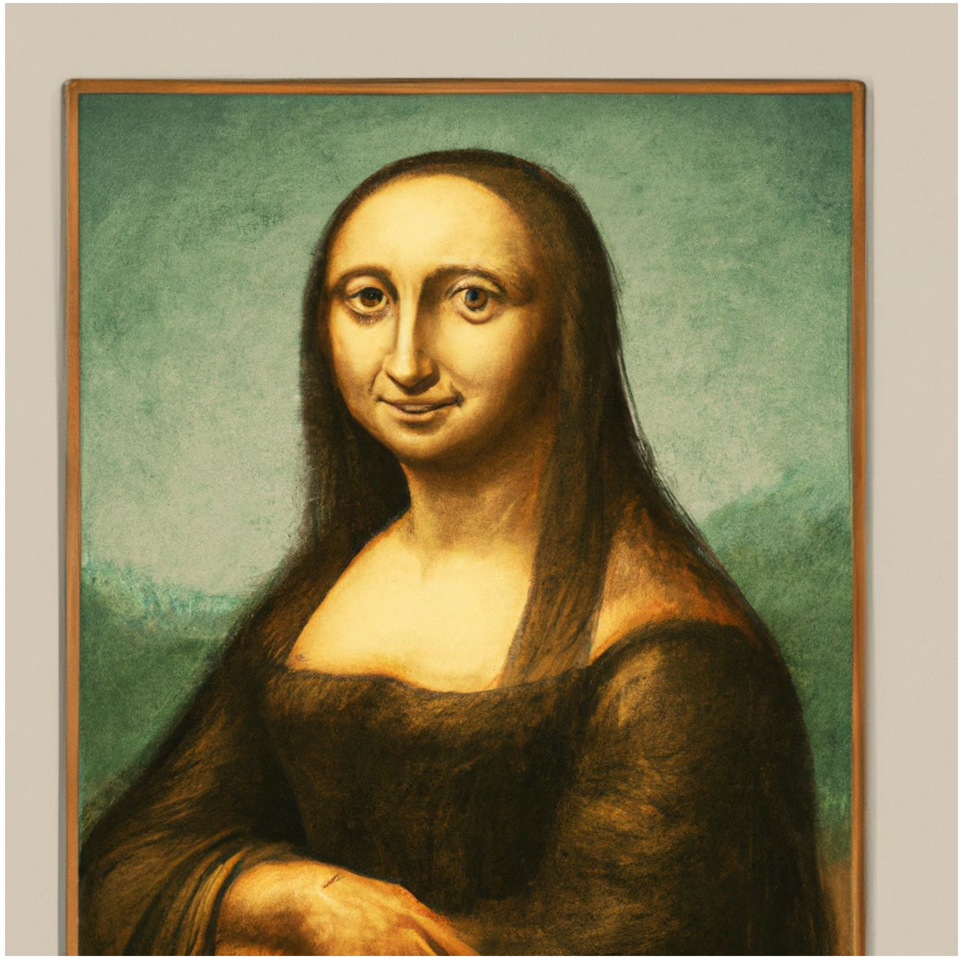
Information = Bits + Context

A Complicated Problem

When using attention, the model weights are dependent upon input. Because of this it means the for every input there is a different function used to predict next token and therefore final answer.

- Any metric of failure is only and independently descriptive of the prompt used to produce that failure.
- The relationship of variance between different prompts will not always be the same and is not measurable in a generalizable way (i.e. variance has to be measured between 2 prompts and is not an indicator of variance between any other 2 prompts)

Variance Contributors:
Prompts
Temperature (Sampling)



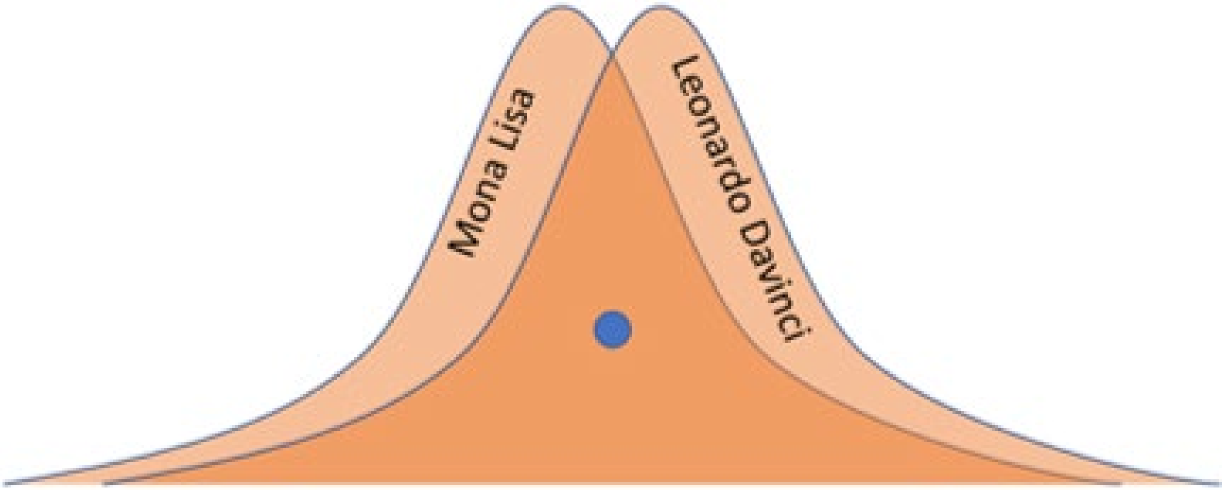
So, what do we do?

Rely on proven statistical methods for measuring reliability in non-deterministic systems
(i.e. Measure it like we do with Humans. Sampling)

Gather statistical data of success rates within context:

Given a specific prompt:

- How often is the answer correct?
- What modes of failure are present in this test?
- What are the statistics of failure for this set of tests?
- How often should we expect success?



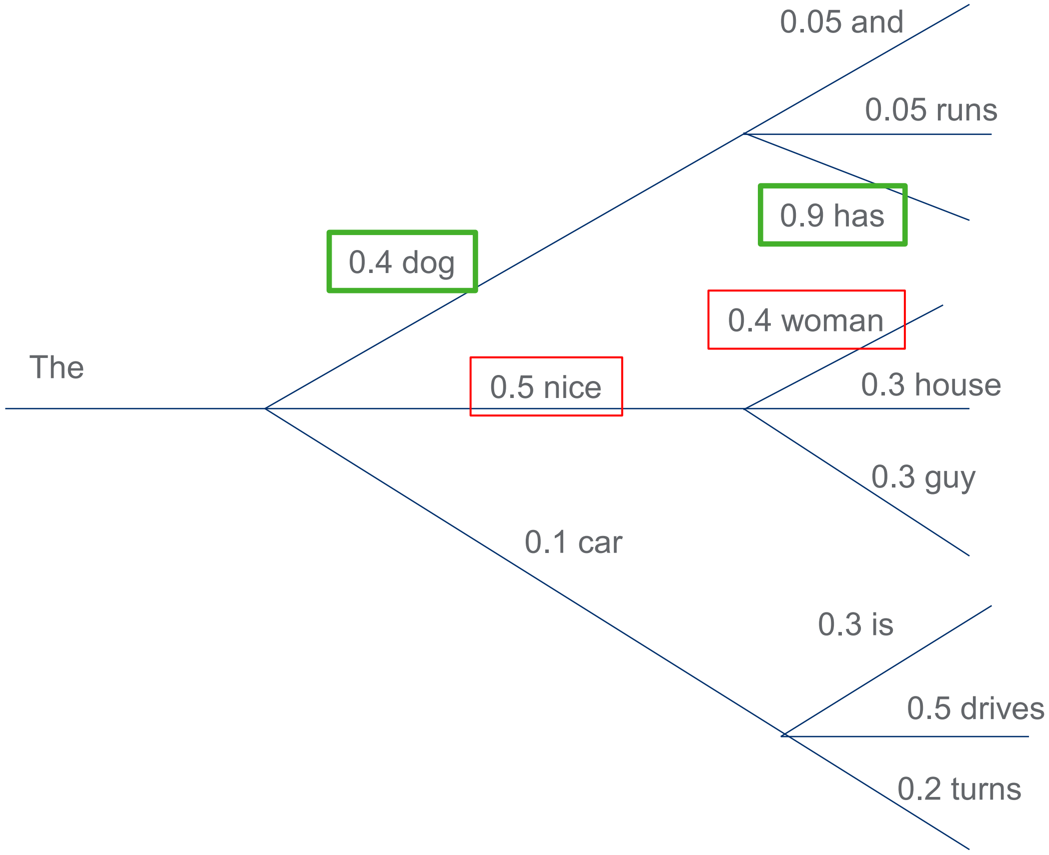
Agent Failure Modes

Temperature Decision Tree

- Sometimes the most likely next token doesn’t lead to the best answer https://huggingface.co/blog/how-to-generate

More Agents is All You Need:

- Sampling the mode of multiple asynchronous generation passes for the entire sequence of tokens, improves performance on benchmarks. https://arxiv.org/abs/2402.05120



Agent Failure Modes

Undefined/Underdefined Context

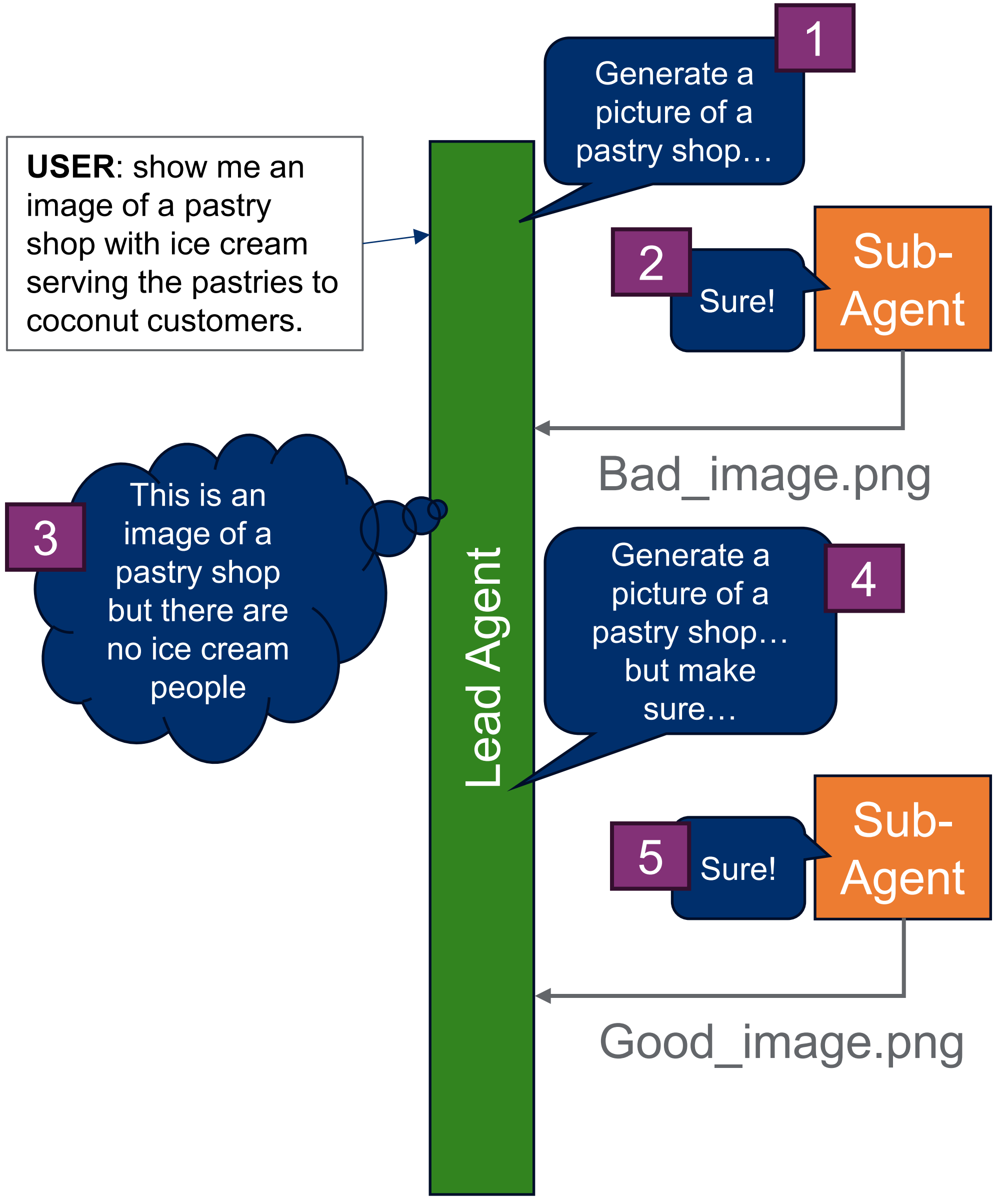
Prompting an LLM with an under defined context creates a wider distribution of probabilities of the next token and therefore a wider distribution of possible answers.

LLMs and agents perform better when their role/context in a system is well and narrowly defined as well as when the prompts or question is well or narrowly defined.

Systems of Agents architectures help with this by narrowly defining the job of an agent.

Repeated attempts = redundancy

In agent systems we can define a role for an agent to check the answers of other agents to make sure that it conforms to the prompt of question being asked. Failing to conform results in a repeating of the task.



USER: show me an image of a pastry shop with ice cream serving the pastries to coconut customers.

Lead Agent
1
Generate a picture of a pastry shop…
2
Sure!
Sub-Agent
Bad_image.png
3
This is an image of a pastry shop but there are no ice cream people
4
Generate a picture of a pastry shop… but make sure…
5
Sure!
Sub-Agent
Good_image.png

Agent Failure Modes

Training data provides some context to LLMs:

The use of a phrase or word with a narrow context helps to narrow the distribution of expected next tokens

This is in essence how we and LLMs understand questions and answer them correctly

Using a phrase or a word whose context in not included in the training data will results in wrong answers or responses.

What does skibidi mean? What does rizz mean?

For many of us these words are out of sample.

Agent Failure Modes

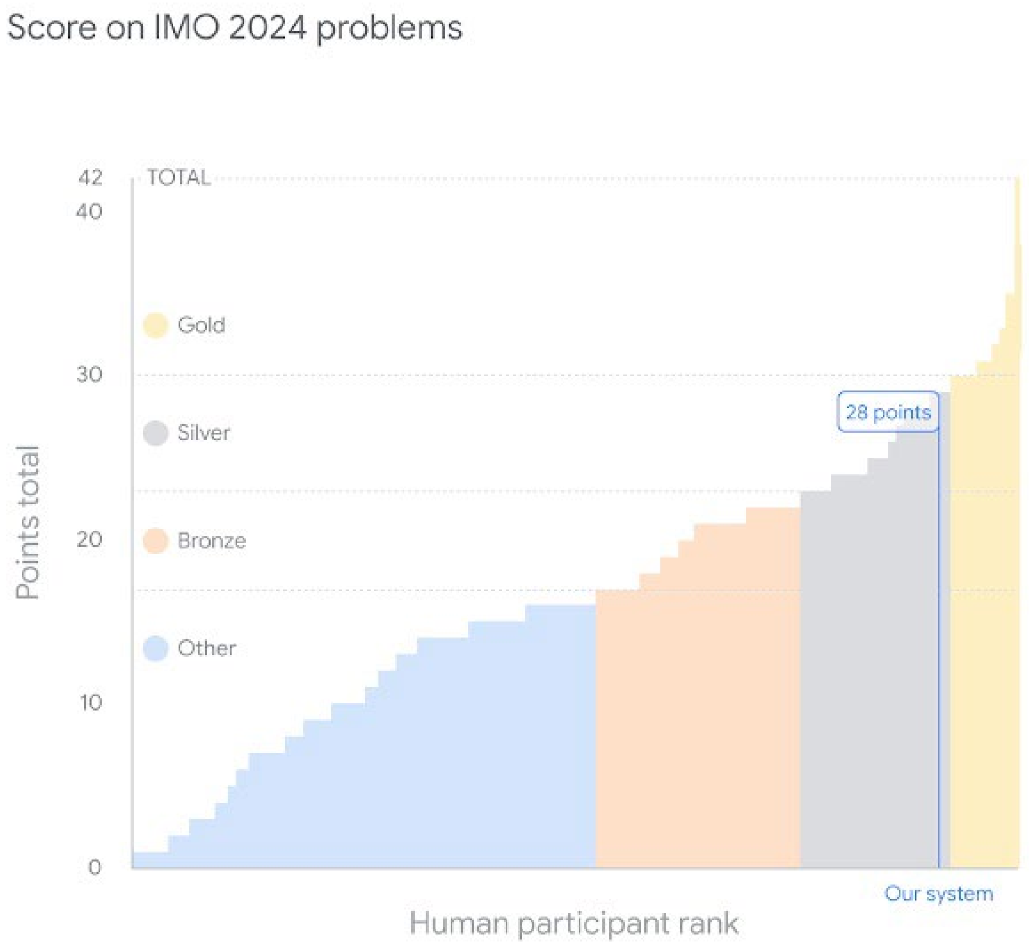
Oddly One of the most challenging problems

Mathematics is not a language. Mathematics are by rule **deterministic**.

Understanding how to do math cannot really be learned by modeling language

Numbers are:

- Categorical
- Ordinal
- Numerical
- Proper Nouns
- Pronouns
- Verbs



https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/

Benchmarks

Open-source benchmarks are an aggregation metric on a particular subject:

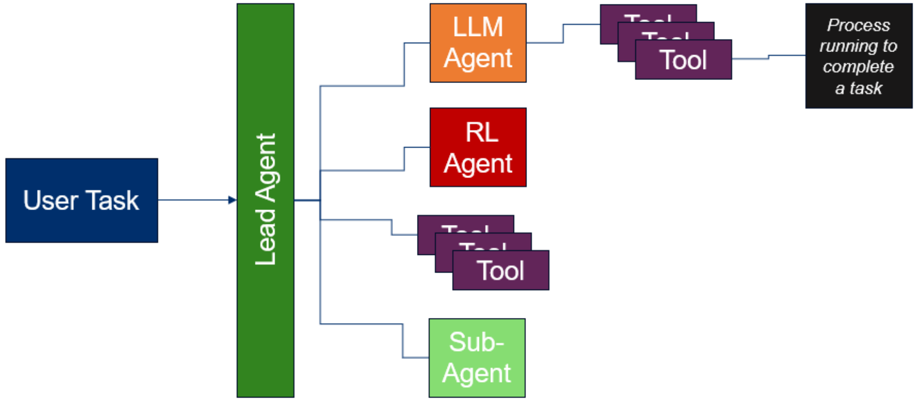
- Thousands of prompts
- Sometimes only given 1 shot (temperature=0)
- Metrics differ at different temperature settings

These measurements are not a good indicator of model performance in a narrow context. (i.e. they don’t tell you how well it can work in your specific use case)

System of Agents Reliability Calculation

The reliability of a System of Agents depends on what and how you ask it.

- What is Lockheed Martin doing to tackle this problem now?
  - Context defined benchmarking
  - Modeling Systems of Agents in classic block diagram framework
  - Measuring reliability growth over experimentation (i.e. test, fix)


User Task
Lead Agent
LLM Agent
Tool
Process running to complete a task
RL Agent
Tool
Sub-Agent