LaSR: Symbolic Regression with a Learned Concept Library

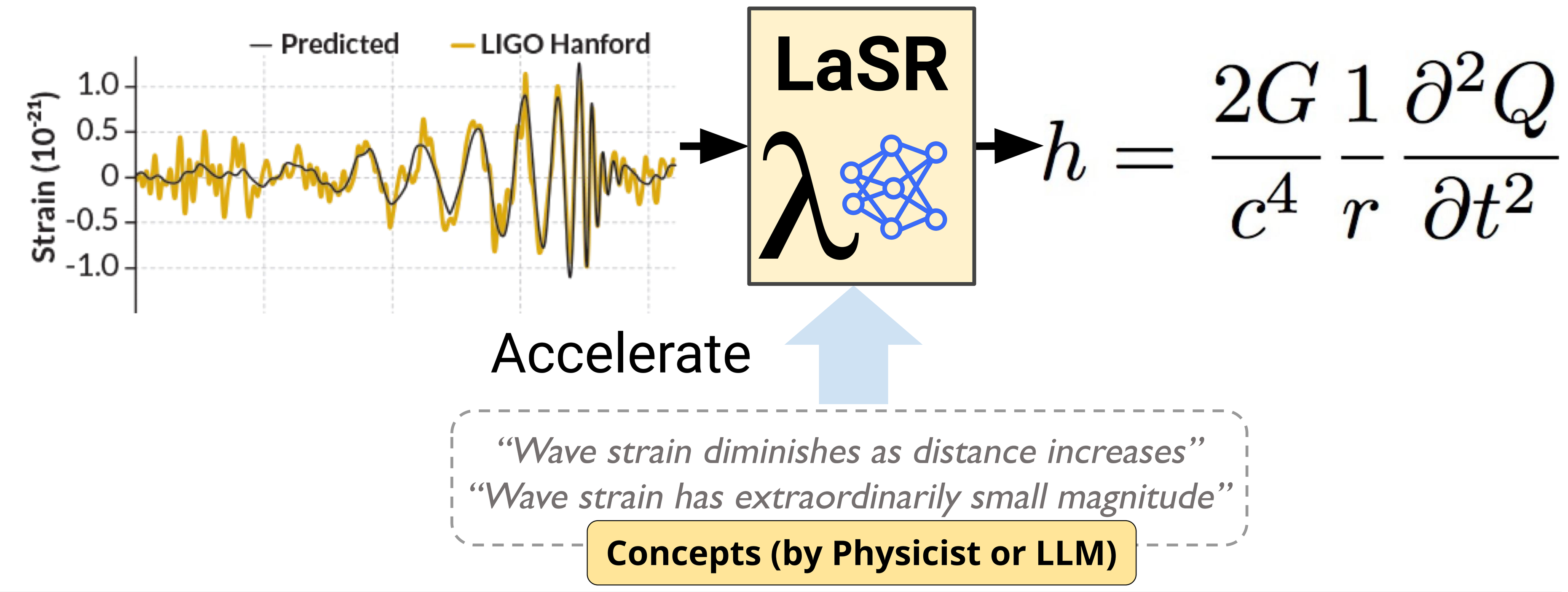
Arya Grayeli[1,4*], Atharva Sehgal[1*], Omar Costilla-Reyes[3], Miles Cranmer[2], Swarat Chaudhuri[1]

[1]UT Austin [2]Cambridge University [3]MIT CSAIL, [4]Foundry Technologies *Equal Contribution; (contact: atharvas@utexas.edu)

Problem

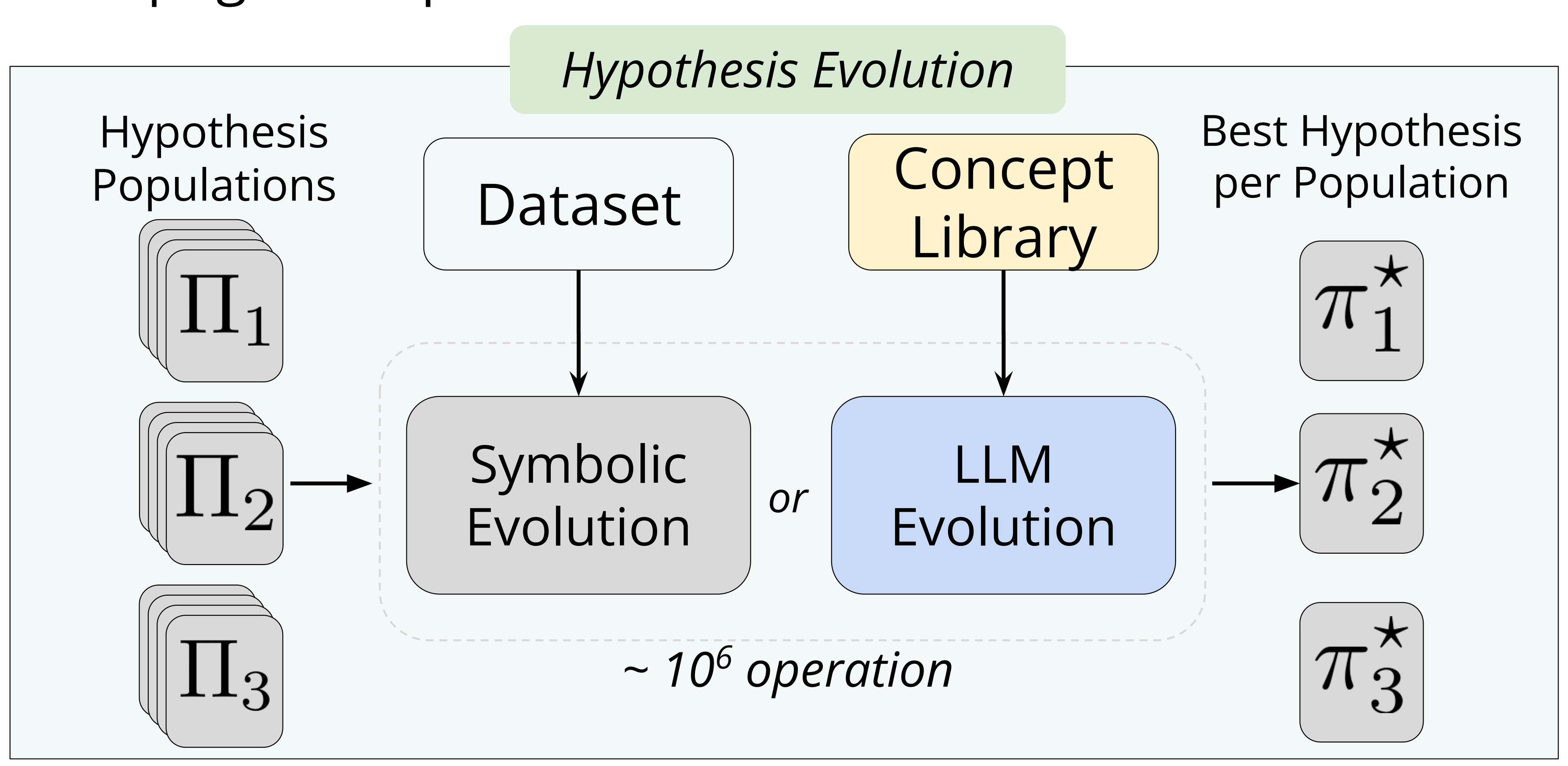
Goal: Discover empirical laws from raw experimental data.

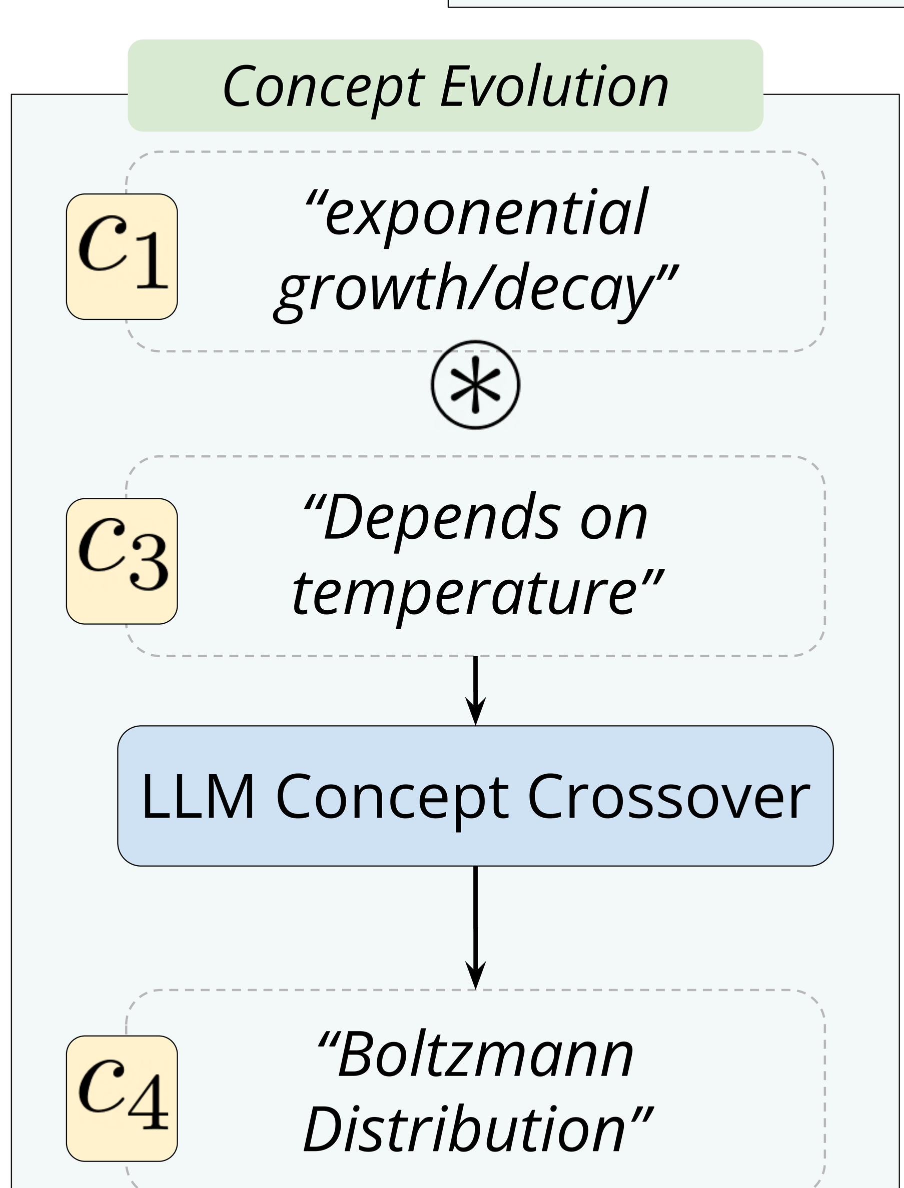
Overview

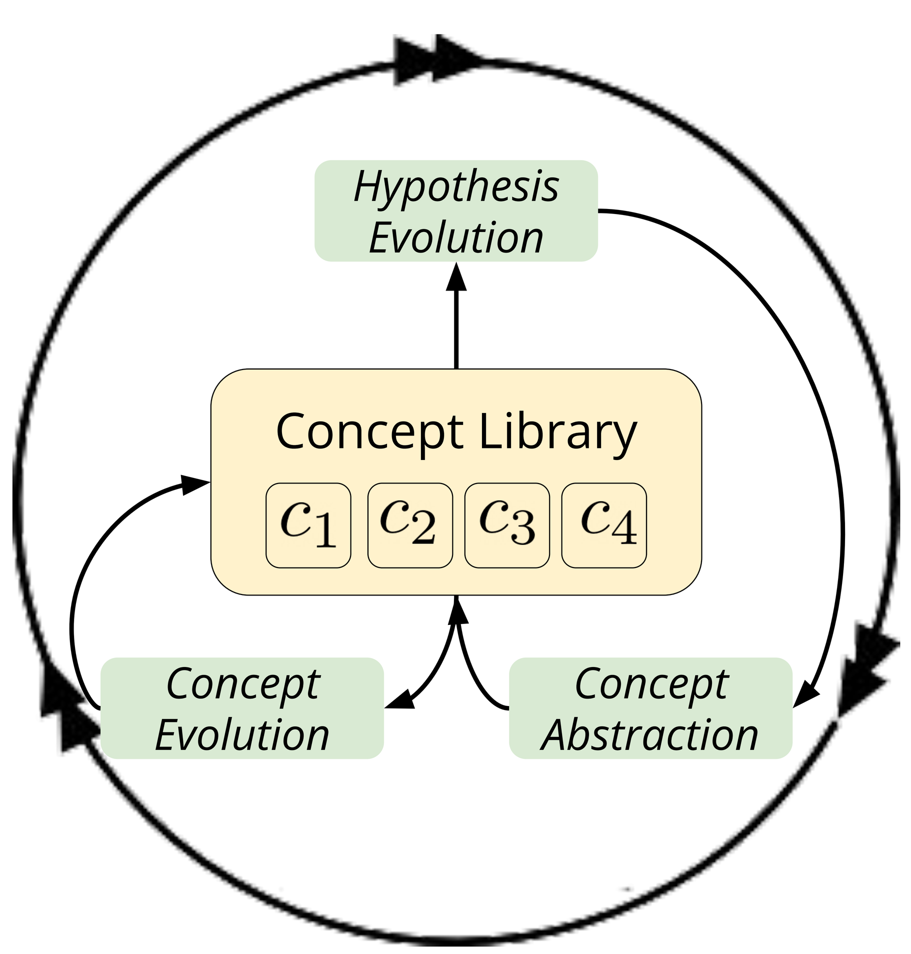


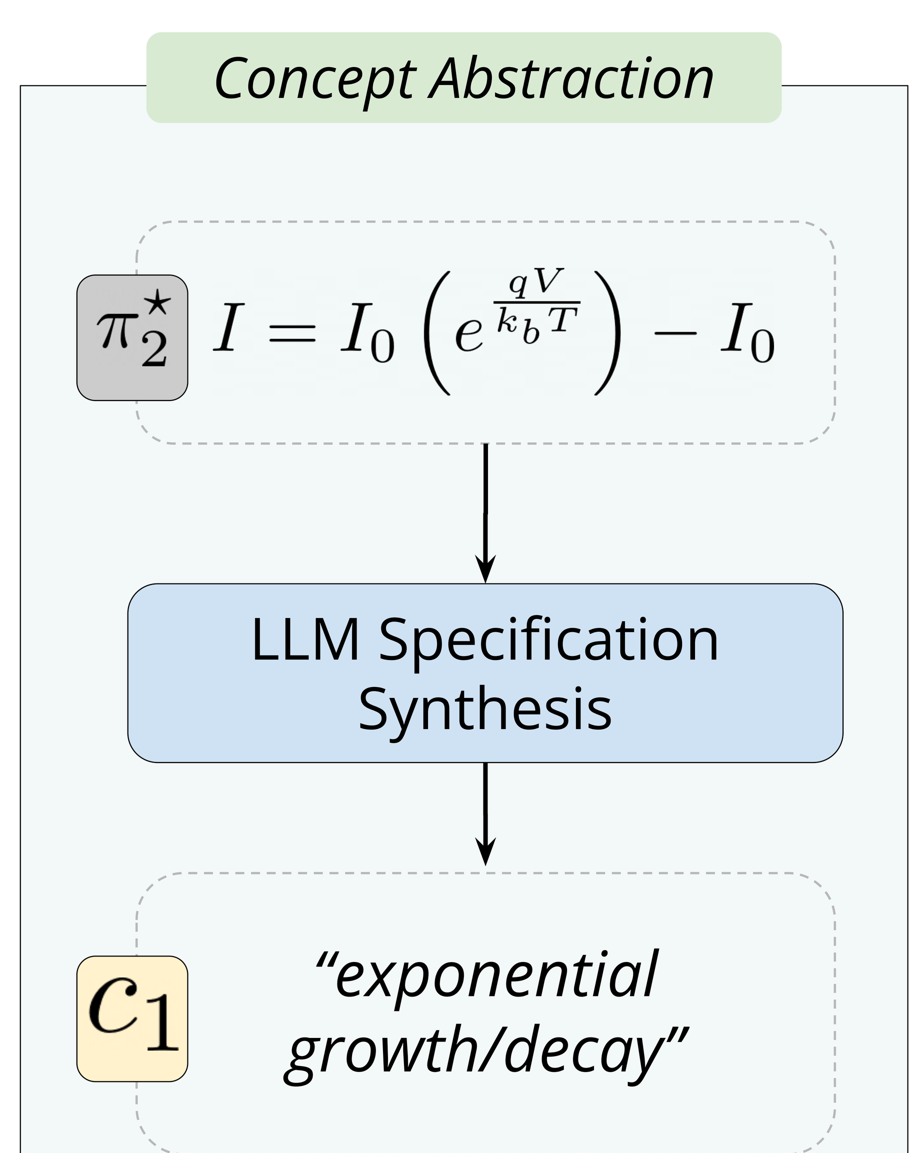
Algorithm

Key Ideas:
I. Use LLMs to generate abstract concepts that summarize high-performing equations and to produce equations aligned with those concepts.
II. Alternate between finding the best equation given concepts, and the best concept given equations.









Can LaSR rediscover known scientific equations?

Observation 1: Concept guidance accelerates scientific discovery.

GPlearn	AFP	AFP-FE	DSR	uDSR	AIFeynman	PySR	LaSR
20/100	24/100	26/100	23/100	40/100	38/100	59/100	**72/100**

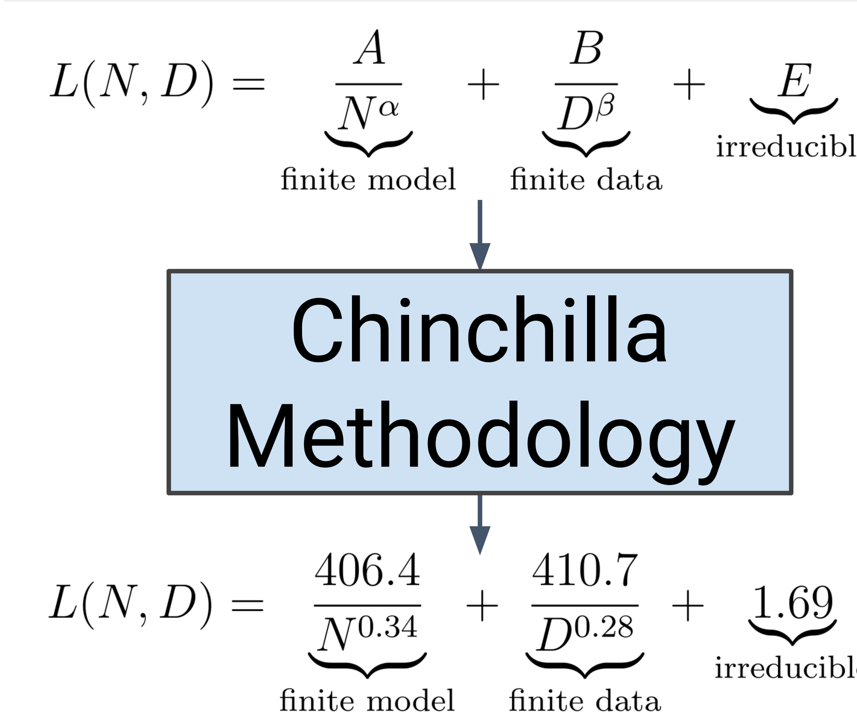
Table 1: Results on 100 Feynman equations from [49]. We report exact match solve rate for all models. LaSR achieves the best exact match solve rate using the same hyperparameters as PySR.

Observation 2: LaSR outperforms PySR even with local language models (`llama3-7b, 1%`)

		LaSR (Llama3-8B)			LaSR (GPT-3.5)
Type of Solve	**PySR**	$p = 1\%$	$p = 5\%$	$p = 10\%$	$p = 1\%$
Exact Solve	59/100	67/100	69/100	71/100	72/100
Almost Solve	7/100	5/100	6/100	2/100	3/100
Close	16/100	9/100	12/100	12/100	10/100
Not Close	18/100	19/100	13/100	16/100	15/100

Table 2: Evaluation results on Feynman dataset by cascading LaSR's LLM backbone (llama3-8b, gpt-3.5-turbo) and changing the probability of calling the model (p = [0.01, 0.05, 0.10]) in the order of increasing concept guidance. LaSR outperforms PySR even with minimal concept guidance using an open-source LLM.

Can LaSR discover new equations?

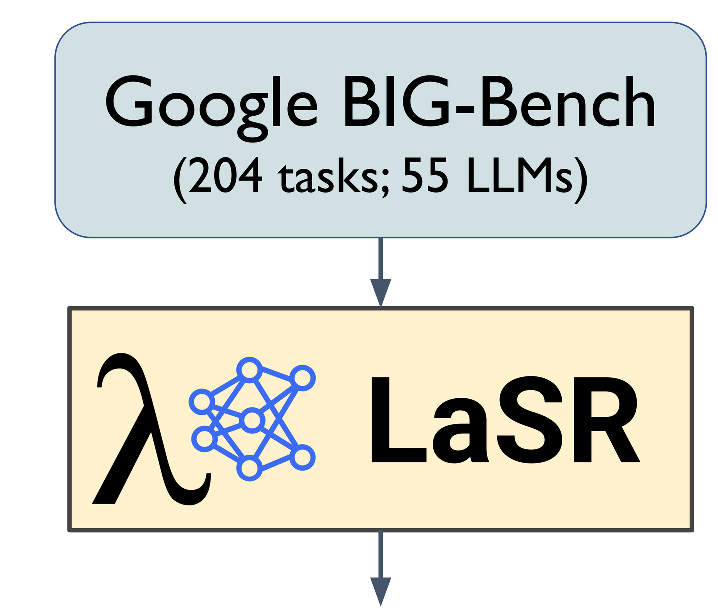


Previous work necessitates manually postulating a scaling law and fitting free parameters to a dataset.

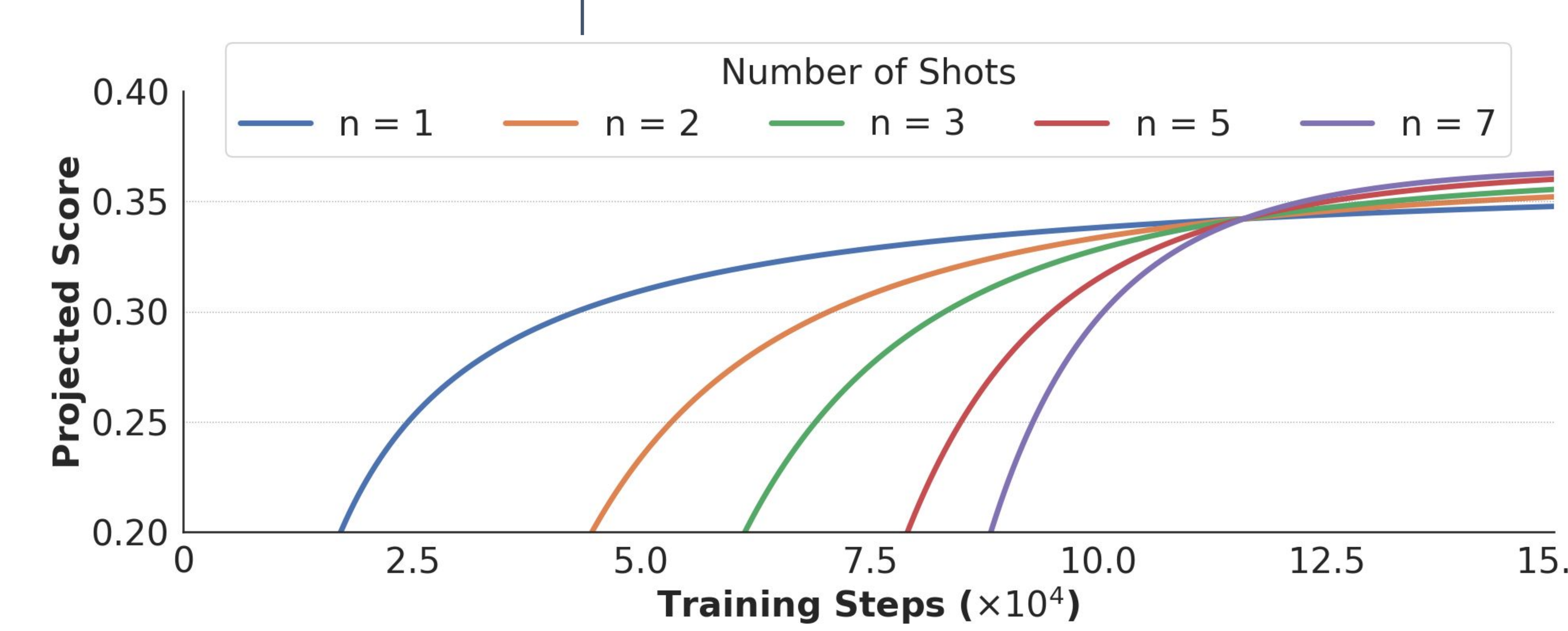
Step 1: Catalogue model performance w.r.t hyper parameters

Step 2: Use symbolic regression to postulate and fit scaling laws.

Step 3: Choose the scaling law that fits the data the best while using the least free parameters.



$$\hat{y} = \frac{-0.0248235}{\left(\frac{\texttt{train\_steps}}{116051}\right)^{\texttt{\#shots}}} + 0.367$$



Visualization of the projected values of LaSR's scaling law for various inputs.

Takeaways

- **LaSR generalizes beyond SR:** Concept guidance may be useful in domains other than scientific discovery.
- **Scientific Knowledge is Code.** Many scientific theories are often represented as code, and discovering non-trivial codes enables new scientific discoveries.
- **Local Language Models** are capable of making non-trivial discoveries!