Phase 2: Quantitative Evaluation 2

Overview

As described in the guidelines, the goal of Quantitative Evaluation 2 (QE2) is to evaluate, in an automated way, the capacity of submissions for collaboration and innovation.

This is done through an evolutionary optimization process that uses the submissions as initial raw material from which to discover new sets of prescriptors that make better and better tradeoffs.

Experiment Details

Evolution was implemented using the Cognizant LEAF ESP service.

To map submitted prescriptors to an evolvable format, each prescriptor was distilled into a neural network via supervised learning of the prescriptions it made in QE1. The input of each neural network was cases per 100K for the past 21 days, and the output is a setting of IPs for the next day. The data of five tasks were used for distillation (days 3-7; all-ones IP cost), resulting in ~100K samples for each prescriptor, a random 20% of which was used for validation for early stopping. Using only data from all-ones tasks keeps the distilled NNs structurally simple, and makes it easier to learn accurate distillations, since the IP costs can be ignored. The distilled models were then placed in the initial population of evolution.

The task for evolution was to prescribe for 90 days starting on 2/10/2021 for the 20 regions with the most total deaths.

Ten independent runs of evolution were run for 100 generations each.

Metric AUC of Population Ancestry %

At each generation of evolution, we can look at each prescriptor in the population and trace its ancestry to compute what percentage of its ancestry came from each submitted prescriptor. This is essentially what 23andme would do to compute ancestry percentages if it had access to complete ancestry data.

From this ancestry % of individual prescriptors, we can compute the ancestry % across the entire current population, which gives a metric of each submitted prescriptor's contribution to the current generation.

We then take the mean of this Population Ancestry % over all generations to get a metric (called Area Under the Curve (AUC) of Population Ancestry %) for how much each submitted prescriptor contributed to the entire evolutionary process.

To get a metric for a team, this AUC of Population Ancestry % is summed over all that team's prescriptors.

For fairness across teams that submitted different numbers of prescriptors, prescriptors in the initial population are selected for reproduction at a rate inversely proportional to the number of prescriptors submitted by their team.
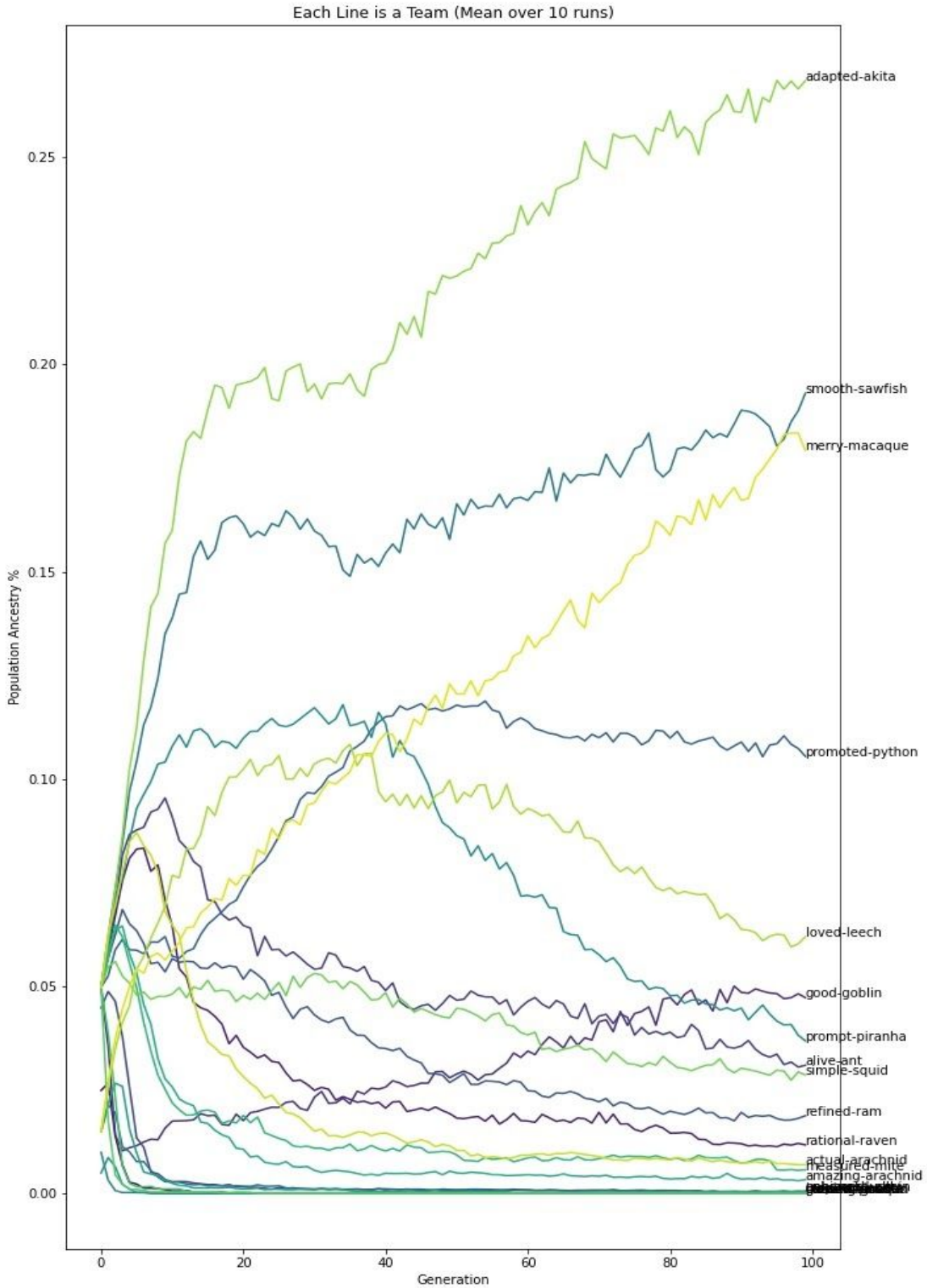
Team Results

The below table shows AUC of Population Ancestry % for each team. Mean and stdev are computed over 10 runs. Team names have been anonymized separetely for this document; these are not the team names that were used during the competition.

	Team	Mean	Stdev
0	adapted-akita	0.2141	0.0247
1	smooth-sawfish	0.1607	0.0248
2	merry-macaque	0.1170	0.0194
4	promoted-python	0.0983	0.0114
3	loved-leech	0.0834	0.0227
5	prompt-piranha	0.0809	0.0094
6	alive-ant	0.0519	0.0137
7	simple-squid	0.0413	0.0086
8	refined-ram	0.0344	0.0120
9	good-goblin	0.0312	0.0074
10	rational-raven	0.0271	0.0055
11	actual-arachnid	0.0210	0.0074
12	measured-mite	0.0143	0.0092
13	amazing-arachnid	0.0106	0.0026
14	unbiased-urchin	0.0033	0.0013
15	enhanced-elk	0.0020	0.0017
16	inspired-insect	0.0019	0.0012
17	present-poodle	0.0015	0.0007

18	glowing-goat	0.0009	0.0003
19	trusted-terrapin	0.0008	0.0002
20	causal-crawdad	0.0002	0.0001
21	current-cicada	0.0002	0.0001

The results are quite consistent across independent runs, pairwise p-values for each pair of teams is shown below. There are clear clusters of teams at different levels of performance.

p-value for each pair of teams (Wilcoxon)

The below plot gives a more detailed look at how the Population Ancestry % changes for each team throughout evolution. Each team's curve on this plot is summed across all prescriptors of the team and averaged over the ten runs.

Each Line is a Team (Mean over 10 runs)

To get even more detail, the below plot pulls apart the team results into their constituent prescriptors. Each curve corresponds to one submitted prescriptor and is averaged over the ten runs. Each curve is labeled by the team name and the prescriptor index, where a lower index indicates lower stringency.
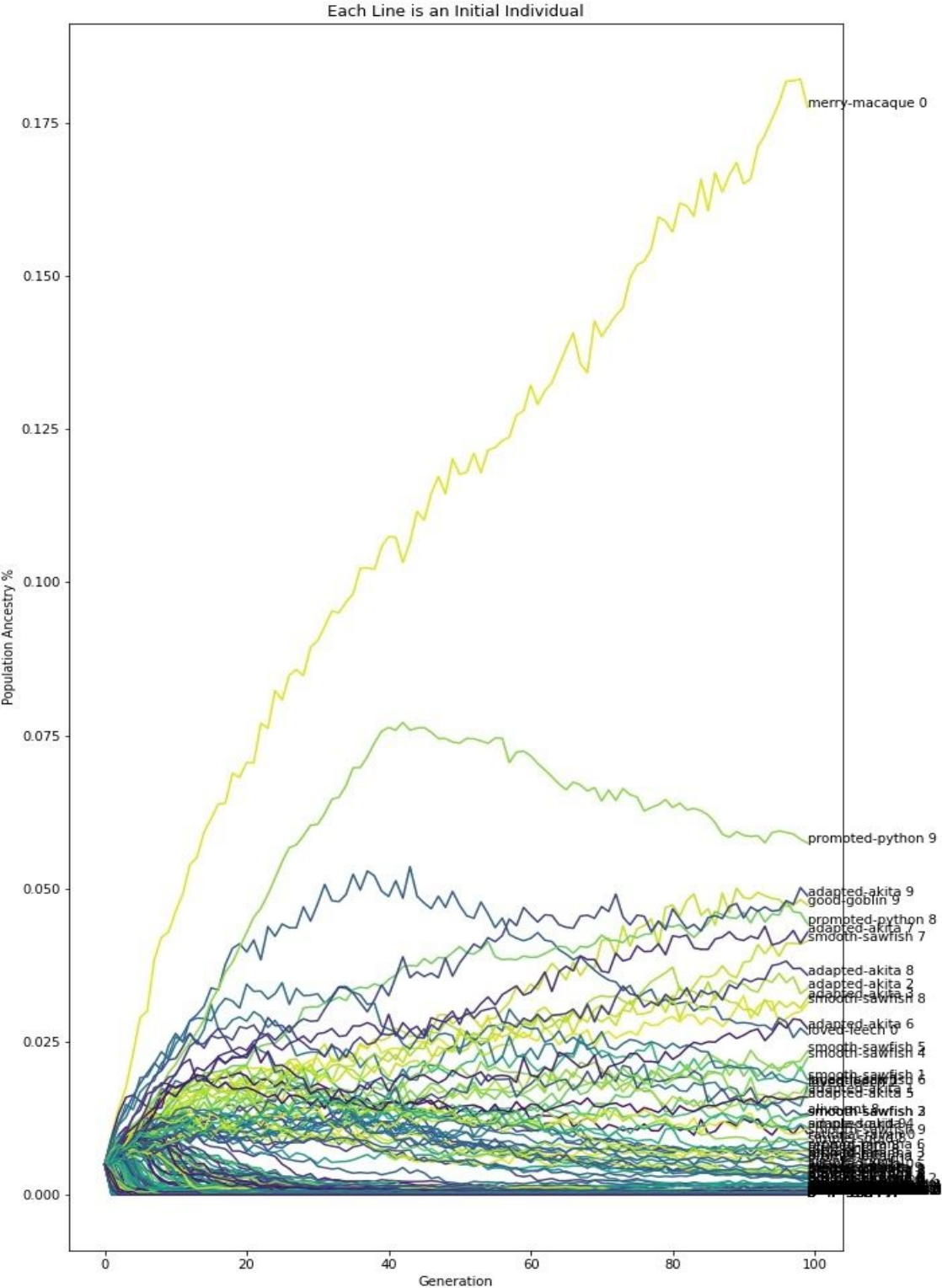
Overall Evolution Results

The previous section looked at the impact of each team on evolution, this section shows the overall benefits of using submitted prescriptors in the evolutionary process, i.e., the benefits of this collaboration across teams.

The following sequence of plots show the qualitative behavior of the evolutionary runs seeded with the teams' prescriptors. Each plot is from a single run, and the results are qualitatively consistent across the ten runs.

The below plot shows how evolution combines the initial prescriptors from teams to discover innovations that flesh out and extend the Pareto front.
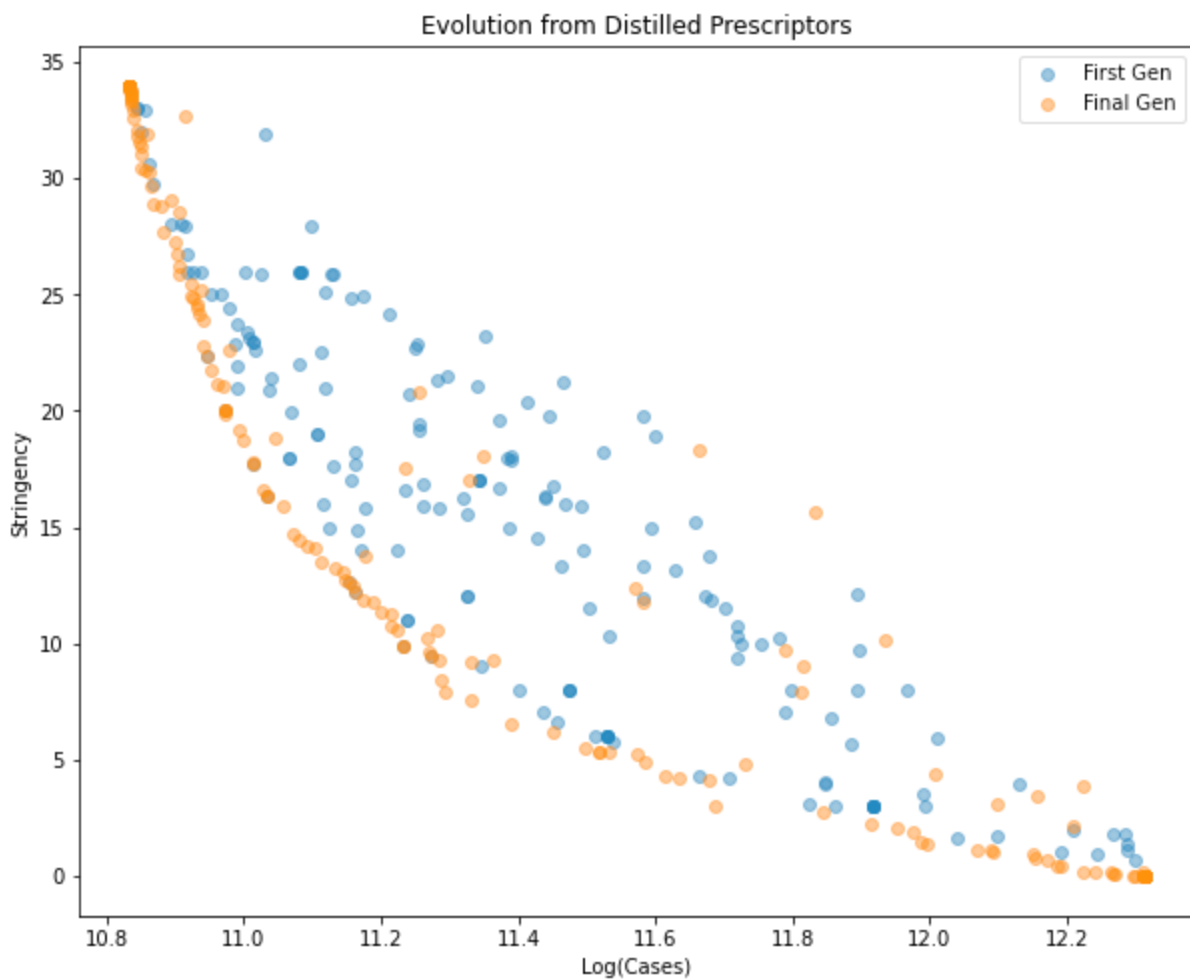
The next plot compares this result to evolving from scratch, i.e., without initializing the population from team submissions. Initialization with the teams' prescriptors clearly yields improvements across the entire Pareto front.
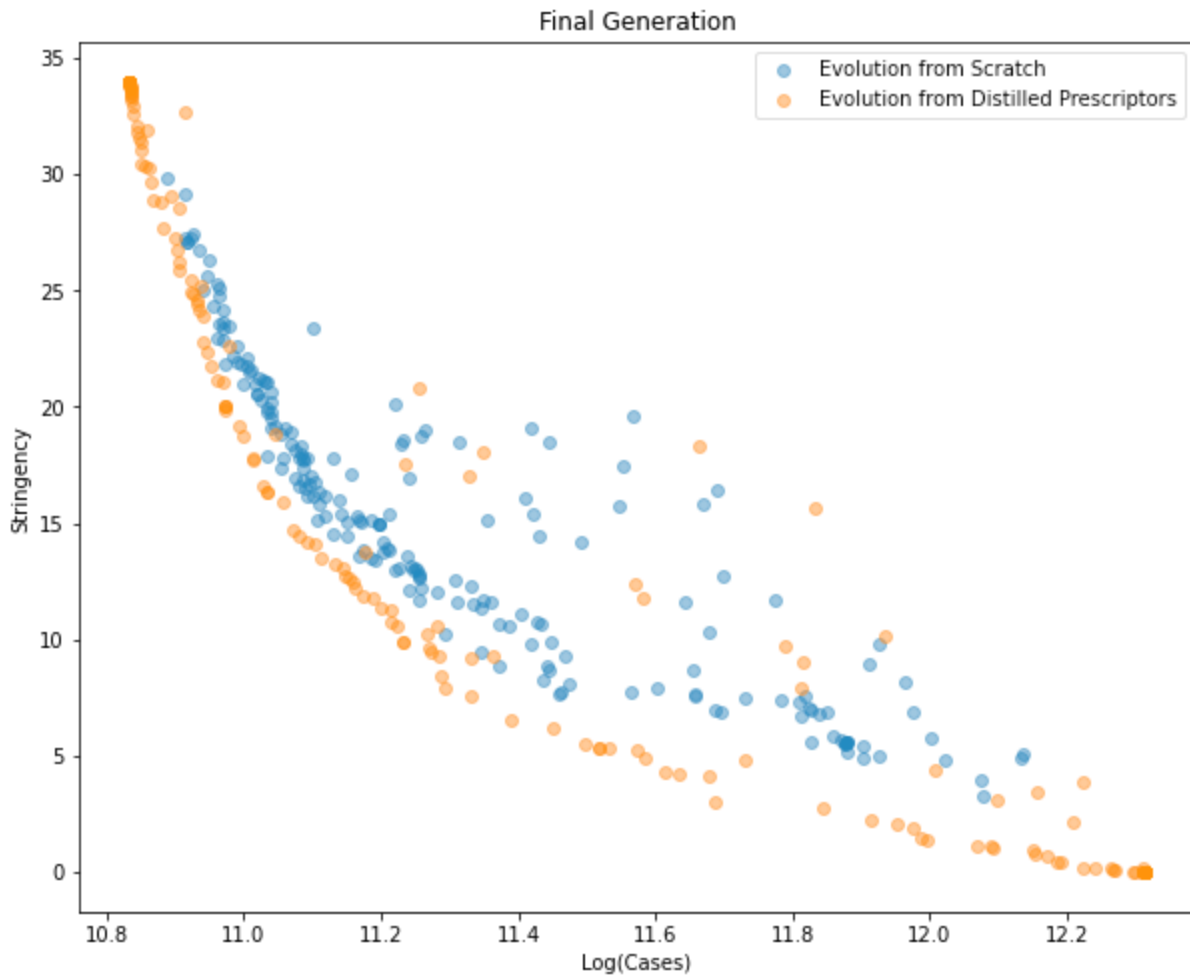
The next plot shows the difference in the first generation of runs with and without seeding. The orange dots consist of the 169 distilled prescriptors along with 32 random prescriptors; the blue dots consist of 201 random prescriptors. We can see that the teams' prescriptors provide evolution with a well-formed Pareto front along with other promising material for promising further innovations.
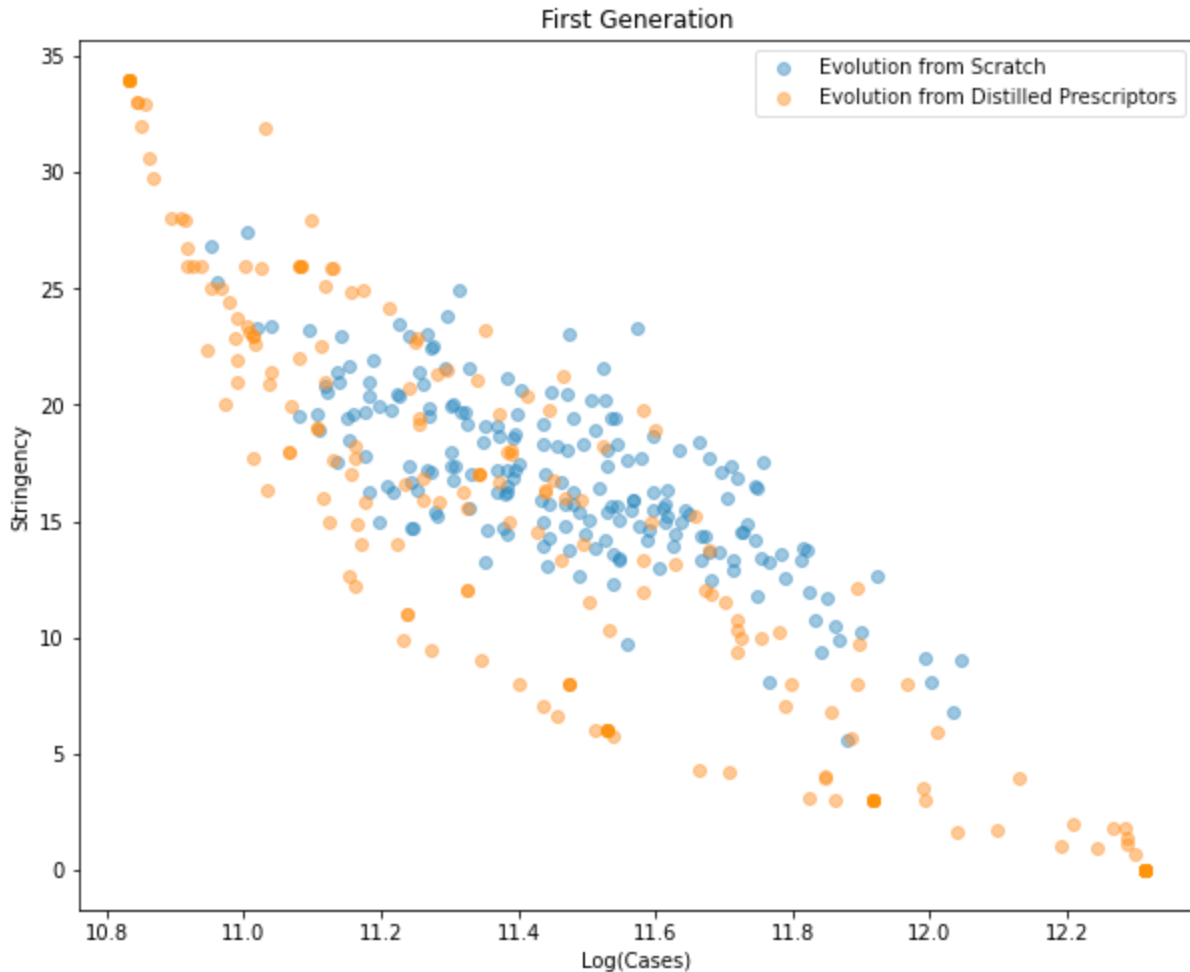
Conclusion

This document shows a way to quantitatively evaluate how useful the ideas embedded in the submissions are in conjunction with other ideas in the community. Moreover, the method makes it possible to combine them into superprescriptors that perform even better than the original prescriptors, thus forming a synergistic result of the creative efforts of the entire community.