caffeinated.bio: what it puts in front of scientists each week, and what a model of it shows so far

Draft r6, 4 November 2026. Raminderpal Singh's caffeinated.bio project.

This document describes a design under test. Everything it reports comes from a behavioral model of the system, run on made-up worlds with no real compounds or scientists, and nothing is deployed.

1. What caffeinated is

caffeinated.bio is a computational workflow for early small-molecule drug discovery. Alzheimer's disease is its first indication. It is designed to run in the background, week after week, on one desktop computer the team controls. It runs beside the team's own discovery work and replaces no part of it. A scientist looks at what it has produced when they choose to.

Most discovery teams work from a ranking of their compounds. A ranking shows what is at the top. It does not show where its order is uncertain or where it may be wrong. caffeinated is meant to show those places. Its weekly outputs, above all its maps, are new views of the team's data and the public record. Each is designed to point an experienced scientist's attention somewhere the ranking does not. The hope is that the scientist then sees a target, a test or a series worth a closer look.

This follows an argument by Alexander Krauss in *The Engine of Scientific Discovery* (Oxford University Press, 2026). Krauss studied science's major discoveries, including every Nobel Prize discovery. He found that all of the more than 750 he examined were triggered by first developing a new method or instrument that made them possible. caffeinated is meant to be an instrument of that kind for a discovery team. Whether it works as one can only be shown by scientists using it. This document reports what a model can show before then.

The system is designed to run itself, with software agents driving its learning. Each week the agents read the outputs, write to the scientists about what they find, and propose changes for the next week. A learning step scores how the agents' earlier predictions turned out and decides which changes go in. Every change stays within bounds the scientists set, is logged and can be undone. The scientists monitor the system, adjust its bounds, weights and settings, and extract the results and insights they need. By default the model of the system runs a narrower version of this loop. A version in which the loop runs itself has now been built in the model and tested on made-up worlds, as sections 2 and 4.9 explain.

In the design, the team's in-house data stays on the machine. A reasoning engine called fred, which runs a language model locally, reads public evidence only. This is the design aim. Today a model of the system runs on that desktop. In the model, fred, the stress test and the agents' language model are scripted stand-ins, and fred itself runs on a separate laptop.

One limit applies to every map and message in this document. In-house records, including the team's lab results, reach only a private version of the ranking and the list of tests. The maps, the agents and their track records see public data only. A fix that would let them see the team's own results has been chosen and is not yet built.

2. One week of caffeinated

Each week runs in five steps (Figure A). First, caffeinated gathers new records on compounds and targets and measures every compound. Each estimate, such as potency or predicted toxicity, carries a

range. Second, it ranks the compounds many times. The ranking formula's weights are the scientists' own and never vary. What varies are the settings the system is unsure of, such as how noisy the assays are. Each combination of settings is called a configuration, and each configuration is run several times. Configuration 0 uses the base settings, and its ranking is called the base run. The whole set of runs is the batch.

Third, caffeinated looks past the top. fred is given public evidence about leading compounds chosen by fixed rules. caffeinated then puts forward a few claims a scientist would not get from the top of the ranking, called caffeinated hypotheses. A few of them each week go to a stress test, which argues the case for and against each one. Fourth, it draws four maps and writes a list of tests worth running and a ledger of dated predictions. Fifth, the agents read these outputs, write messages to the scientists and propose changes, and a learning step decides which of them go into the next week.

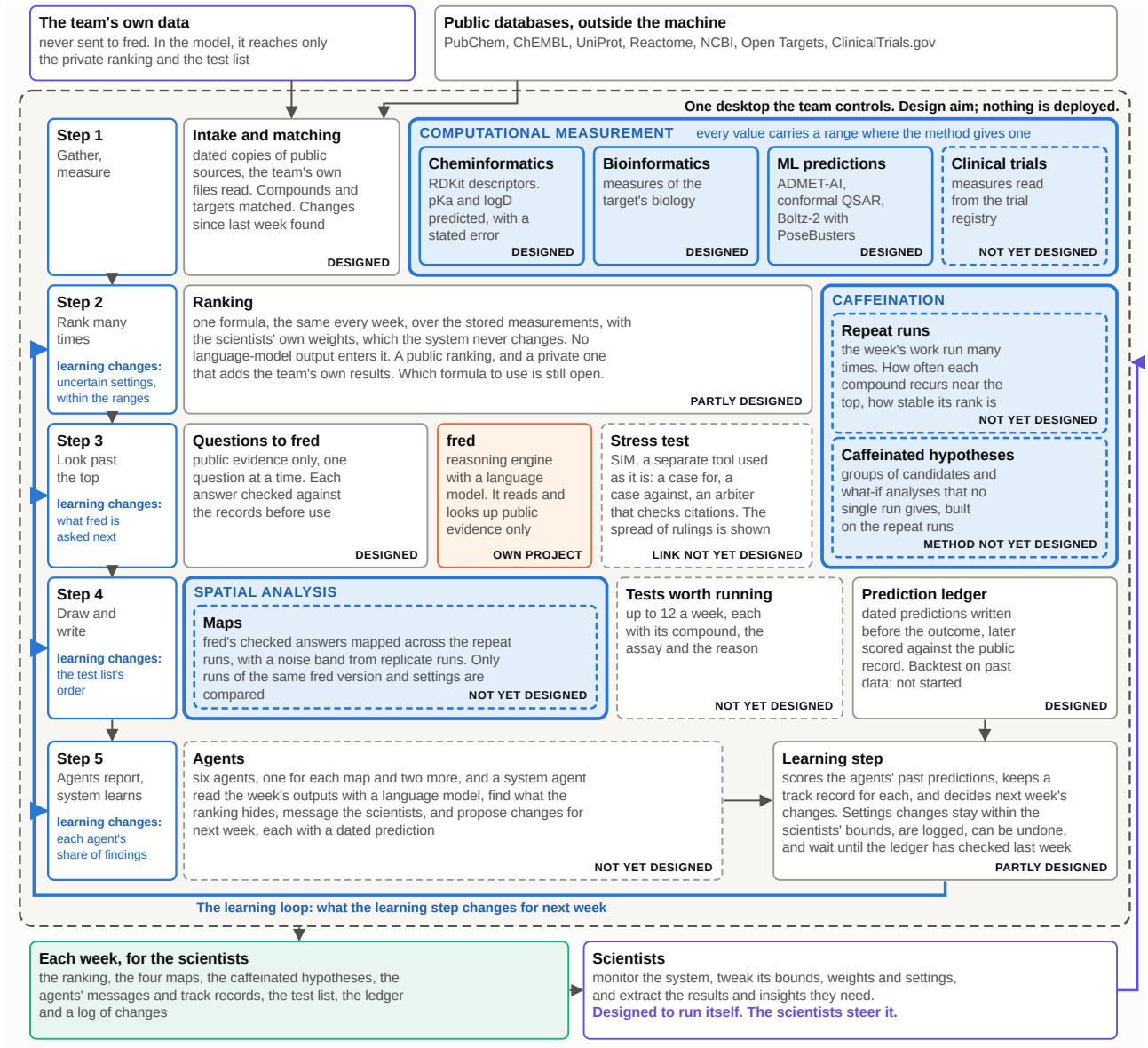

Figure A. One week of caffeinated, as designed. Nothing is built. Each box says what one part does and names the methods the design gives it. Blue shading marks the computational measurements, caffeination and spatial analysis. Capitals give design status, and a dashed box is not yet designed. The system is designed to run itself: the agents propose changes, and the learning step decides which go into next week, along the blue line. The scientists tweak its bounds, weights and settings along the purple line. Caffeination spans steps 2 and 3. Only the record format of its repeat runs is set, and its own method is not yet designed. In the model, simpler stand-ins take the place of the shaded parts. Potency is the mean of a compound's own reports, or its series' mean when it has none. The target score is the mean of the target's reports, toxicity comes from a linear predictor, and there are no trial measures. Caffeination asks, on the base run alone, whether removing one of a compound's own reports would move it across the top-10 line. Spatial analysis is the four maps of section 4. fred, the stress test and the agents' language model are scripted stand-ins.

In the design, the learning step can change the uncertain settings within the scientists' ranges, what fred is asked next and the order of the test list. These three come from the design files. It also sets each agent's share of the findings, which comes from the model. Settings changes are logged, can be undone, and wait until the ledger has checked the previous week's predictions, though most of those cannot be scored for months. Whether the same checks cover the other changes is not yet designed. In the design files the changes are proposed by fixed rules. Whether agents that use a language model may propose them, and how that fits with keeping language-model output out of the ranking, is not yet designed. The ranking's weights never change on their own. No result yet shows that the learning loop improves what the system shows.

By default the model runs a narrower loop. In it the agents only message the scientists, and the one change made on its own is how many findings each agent may write the next week, set from its track record. Sections 4.1 to 4.8 report results from that narrower loop or from earlier runs, as section 4 says. The model now also has a version in which the loop runs itself. There the learning step changes the assumed assay noise and the shape of the batch, and the scientists' top 10 comes from each compound's score averaged over the batch. A separate test tried two orders of the test list that use what the system has recorded. Section 4.9 reports both. On made-up worlds the loop passed every declared check of how it works but one, and it has not yet improved what the scientists see.

In the model's earlier runs, the track records ran against the agents' real accuracy. One agent was right on 82% of its findings and was credited with 28% of them. Another was right on 24% and was credited with 67%. Section 4 names both. In the self-running version the track records score claims about a compound's next public report, not moves in the ranking. In earlier runs of that version they ranged on average from slightly against the agents' real accuracy to weakly in line with it, and single runs varied widely. Until they follow accuracy closely, a learning step that steers by them may give room to the wrong agents. That is an inference and has not been measured.

3. The ranking, and what it leaves in the dark

The ranking is the team's familiar view. Throughout this document, "the top 10" means the top 10 of the base run, which uses public data and the base settings. The one exception is section 4.9, where it means the top 10 when each compound's score is averaged over the batch's configurations. The ranking can hide several things without anyone noticing. A compound can rank low only because nobody has measured it, since an unmeasured compound is scored at its series average with a wide range. A compound near the top-10 line can enter it under some settings and not others. A position can depend on an assumption the system is not sure of, such as the assay noise. A position can rest on one record, and one record can hold up several compounds at once. Public reasoning about a compound can disagree with its rank. A compound inside the top 10 can also be there by mistake.

Each weekly output is meant to show one or more of these. Section 5 sets out what each output is compared with.

4. The weekly outputs

Figure B sets out each output, what it is meant to show, a question it might raise for a scientist, and what the model shows so far. The subsections that follow take each output in turn. Each says what the output is, what it is meant to show, what the model's runs show and what they cannot show.

Artifact	Dark area it lights	A question it may raise	What the model shows so far
Map 1 public reasoning	Where fred's reading of public evidence disagrees with the rank	Is this rank right? Is there evidence we have missed?	Its agent's messages were no better than chance
Map 2 assumptions	Positions that depend on a setting the system is unsure of	Which assumption matters, and is our range for it right?	Assay noise outside the set range was flagged. Other messages slightly ahead
Map 3 shared records	Single records that hold up several top compounds	Is this target report sound enough to rely on?	Its agent's messages well above chance, on a different rule
Map 4 near the line	Compounds just outside the top 10, including unmeasured ones	Is a compound outside the top 10 worth a first measurement?	One region beat its band in 4 of 5 worlds. Messages above chance
Hypotheses	Claims a scientist would not get from the top of the ranking	Which claim would change our view of a series or target?	All new, and good as often as ranks 11 to 15, not more often
Messages	Whatever the agents find that the ranking hides	What should I look at this week, and why?	Above chance in 4 of 5 worlds, on a low bar
Test list	Where one test could change the picture	Which test would change most what we believe?	Never beat testing down the ranking by the set margin, in three comparisons
Ledger	None. It checks the system itself	How far should I trust what the system expects?	Worse than always saying fifty-fifty

Figure B. The weekly outputs, what each is meant to show, and where the model stands. The color bar gives a rough reading of the model's evidence. Green means the output beat its reference, blue means it was slightly ahead, gray means it was level and orange means it fell short. These are results on made-up worlds.

Unless a result says otherwise, it comes from five versions of one made-up world, called the hard world, each run for 26 weeks. The made-up worlds have a hidden truth that the system cannot see and the scoring can. A compound is truly good if its hidden chance of success is at least 0.7.

A message about a compound reveals something if it points to a compound outside the top 10 that is truly good, or to one inside the top 10 that is not. Its chance level is the share of compounds ranked within ten places of it that would count by the same rule. A message beats chance when it does better than pointing at a nearby compound at random.

Two points apply to the whole section. First, the results for the maps, the hypotheses, the test list, the ledger and the track records come from earlier runs. Those were made before the model's agents were limited to sending messages. In those runs the agents changed the batch. Before the runs of section 4.9 they were the only evidence on an automatic loop, and they gave the track records that ran against the agents' accuracy. The hypotheses came out the same under every arrangement tried, so they are unaffected. The other results have not been re-run. Second, Maps 1 to 3 have no measured region of their own, as Map 4 has. The only evidence for them comes from the agent that reads each one. Those agents' figures are pooled over five worlds, with no pass rule set for each agent in advance.

Section 4.9 is the exception to both points. Its results come from new runs of the model's self-running version, on the hard world and the NLRP3-shaped world.

4.1 Map 1: where public reasoning and the ranking disagree

Map 1 plots each leading compound's average rank over the batch against how often fred's answers support it. A band shows how much fred's own answers vary from run to run. A compound that ranks high with little public support, or the reverse, stands out.

In the model, the agent that reads Map 1 wrote messages that were no better than chance, at 40% against 42%. This result is partly built in. fred's stand-in knows nothing beyond the evidence it is given, and that is the same evidence the ranking uses. The model therefore says nothing about whether the real fred's disagreements with the ranking point anywhere useful.

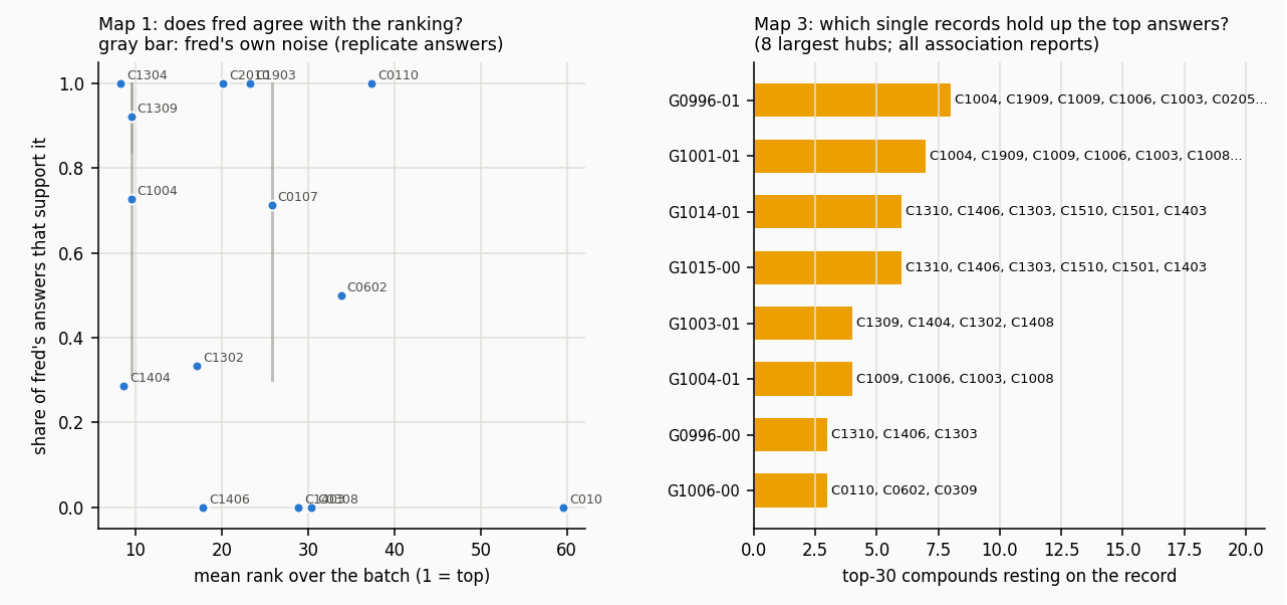

Figure 1. Maps 1 and 3 for one week of one made-up world, from the earlier runs. Left, Map 1: fred's support against average rank, with fred's own run-to-run spread as a gray bar. Right, Map 3: the records that most leading compounds rest on, with the compounds resting on each. Compound and record names are the made-up world's.

4.2 Map 2: which positions depend on an assumption

Map 2 takes the 30 leading compounds in the base run and shows how each one's rank moves between configurations. A compound whose rank swings holds its place only under some settings. The map also names the setting that moves ranks the most (Figure 2).

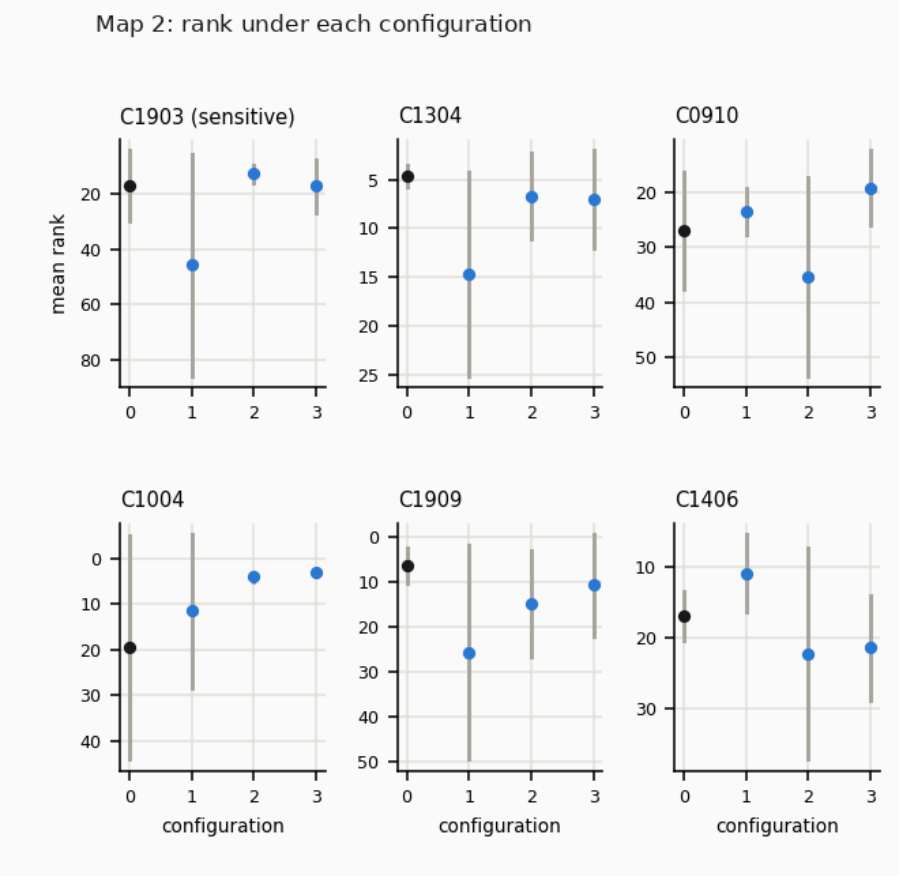

Figure 2. Map 2 for the same week. Each panel is one of the six leading compounds whose rank moves most between configurations. Each point is the compound's average rank under one configuration, with its spread over that configuration's runs. Configuration 0, in black, uses the base settings.

The agent that reads Map 2 also estimates the assay noise from repeated reports. From the twelfth week on, its estimate stayed between 0.86 and 1.19. The true value in these worlds is 1.0. The range the scientists had set was 0.3 to 0.8. The system agent passes such a mismatch on to the scientists as a note. Under the earlier design these notes went out in every world, three to ten times per world. Under the present design they are messages, and all 35 of them over five worlds pointed at a real mismatch. That was expected, because in these worlds the true noise lies outside the scientists' range every week. The result shows the note fires when it should. The model has no case in which such a note would be wrong.

The Map 2 agent's own messages were slightly ahead of chance, at 52% against 45%. For a single world, a lead of that size counts as within noise.

4.3 Map 3: records that many compounds rest on

Map 3 lists, for each leading compound, the three records that most influence its place. In the model these are reports that link a target to the disease. A record that two or more leading compounds rest on is marked as a hub. If a hub record is wrong, several compounds move at once.

About half of the findings of the agent that reads Map 3 came true. Its messages point at records, and they are scored on a different rule: whether the record misstates how relevant its target is. Chance is then the share of all such records that do. On that rule, 55% of its messages revealed something, against 29% by chance. Because the rule differs, this figure is not on the same scale as the other agents'.

4.4 Map 4: the compounds near the top-10 line

Map 4 shows one row per compound. Each row holds the compound's rank over the batch, its evidence and ranges, fred's support, how much the settings move it, and its value against its series. It can be drawn in several ways. The one that has done best is the decision map (Figure 3). It plots each compound's average rank against the share of runs that put it in the top 10. Its region was declared before any run. It holds compounds outside the top 10 that enter it in at least a quarter of the batch's runs.

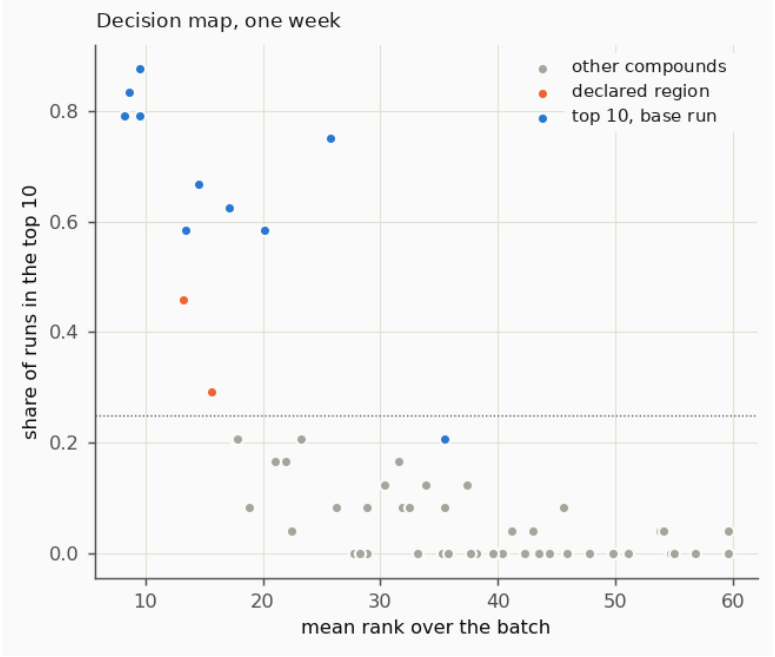

Figure 3. The decision map for one week of one made-up world, from the earlier runs. Each point is a compound. The declared region (orange) holds compounds outside the top 10 that enter it in at least a quarter of the batch's runs (dotted line). A compound can be in the base run's top 10 and still average a lower rank over the whole batch, as the blue point near rank 35 does.

In the model's runs, 76% of the compounds in the decision map's region were truly good. Of the compounds ranked 11 to 40, from which the region is drawn, 50% were. The region did better in four of five worlds. In the fifth, both were at 69%. Other ways of drawing Map 4 did less well (Figure 4). The region on the potency-against-toxicity view held only two to seven compounds per world. It beat its band in three worlds of five. The view meant to find compounds that behave unlike their series found such exceptions. Only 22% of them were truly good, against 26% of all compounds.

A grouping looked at after the runs, and not declared in advance, points the same way as the decision map. Among compounds ranked 11 to 40, those with no report of their own were truly good 57% of the time. Those with reports were truly good 40% of the time. That fits the ranking scoring unmeasured compounds too low. It depends on the series average being a fair guess for an unmeasured member. That holds in these made-up worlds and may not hold in real chemistry.

The agent that reads Map 4 wrote messages that revealed something 56% of the time, against 33% by chance.

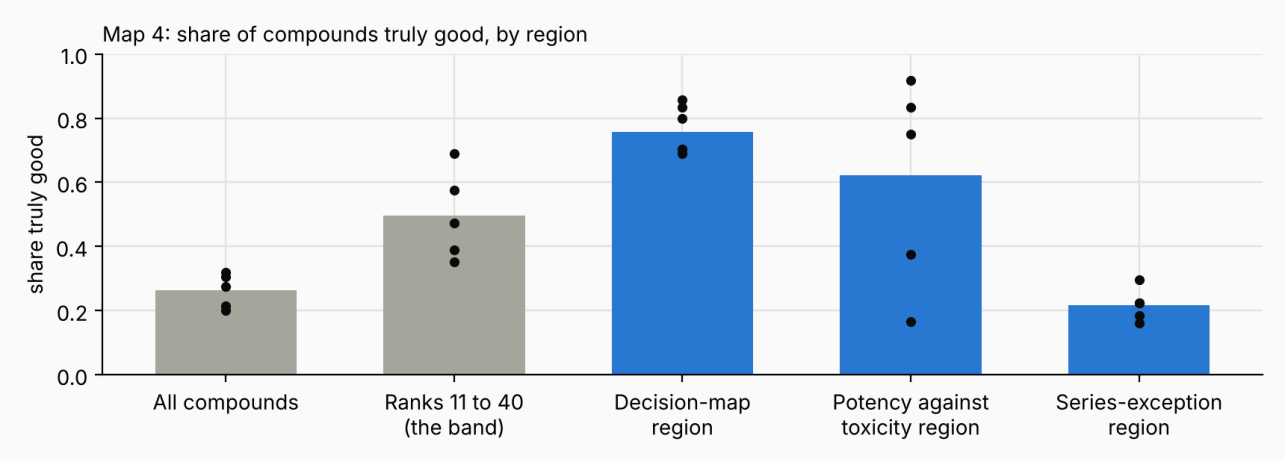

Figure 4. How often compounds in each region of Map 4 were truly good. Bars are the five worlds together and dots are each world. Weeks 3 to 17 of the earlier runs. The decision-map and potency-against-toxicity regions are compared with ranks 11 to 40. The series-exception region is compared with all compounds.

4.5 The caffeinated hypotheses and their stress test

Each week caffeinated puts forward a few claims that a scientist would not get from the top of the ranking. Most are about compounds. The method now in use asks, for compounds just outside the top 10, what single change in their own evidence would bring them in. Some of its claims are about records instead. Such a claim says that a top-10 compound's place rests on records that may be wrong.

In the model, the method put forward 229 claims about compounds over five worlds, all of them new, meaning outside the top 10. Of those, 67% were truly good. Compounds ranked 11 to 15 were truly good 65% of the time. The claims did about as well as those compounds and no better. The share varied widely between worlds, from 26% in one to 100% in another, so one world alone can mislead. The method's 86 claims about records were right 24 times. That has not yet been compared with how often a top-10 compound is not truly good.

The older consensus method was right more often, 79% of the time. Only 34 of its 650 claims were new (Figure 5).

The stress test is a scripted stand-in in the model, and the model says nothing about the real one. The stand-in ruled for between two-thirds and four-fifths of the good claims it tested. It ruled for between seven in ten and all of the claims that were not good.

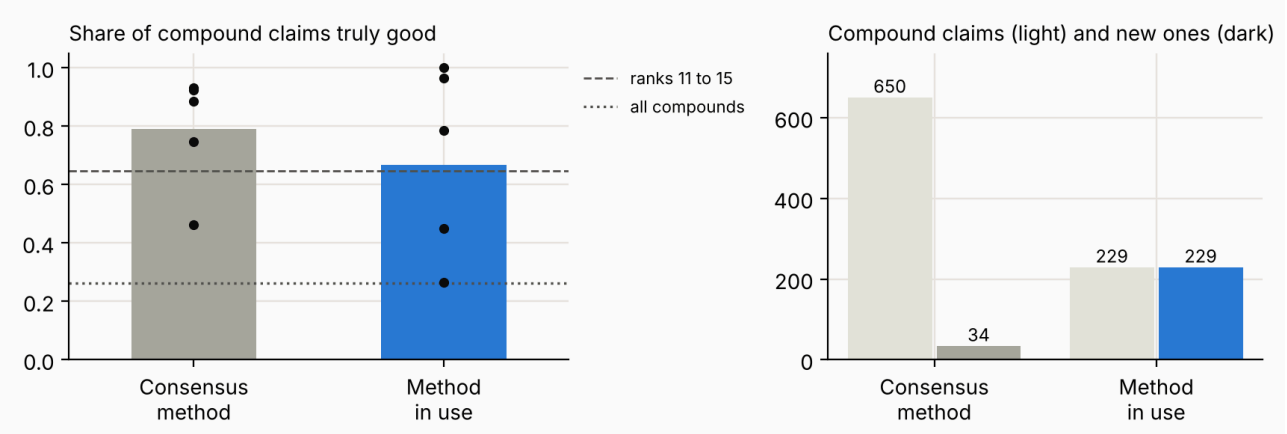

Figure 5. The caffeinated hypotheses about compounds, by two methods, from the earlier runs. Left: share truly good, with ranks 11 to 15 and all compounds as references. Bars are the five worlds together and dots are each world. Right: how many compound claims each method put forward, and how many were new.

4.6 The agents' messages

Each week the agents' messages arrive together in a digest. There are six local agents. Four each read one map. Two read the ranking itself. The near-miss agent looks for compounds that one new report would bring into the top 10. The single-record agent looks for compounds that one record holds in the top 10. A seventh agent, the system agent, reads what the others found and writes about what it adds up to. Its messages include suggestions about next week's batch, which in the model's default only the scientists can take up.

Each message says what the agent saw and which records lie behind it. It also says why this may matter and what a scientist might look at, such as a record, an assumption, a test or a question for fred. Only messages that can be scored count in what follows. About a fifth of all messages could not be scored this way. Most of those were the system agent's suggestions about settings.

This is the one comparison in the model that asked directly whether an output reveals what the ranking hides. It was declared before its runs. Messages beat chance in four of five worlds and were level in the fifth (Figure 6). In our judgment the lead is clear in three worlds and within noise in the fourth. The test had no margin. The same compound can also come up in messages week after week, which makes the counts less independent than they look. So this is a low bar. It was also declared after three comparisons of the test list had failed, and it does not reverse them. A comparison with a margin set in advance comes next.

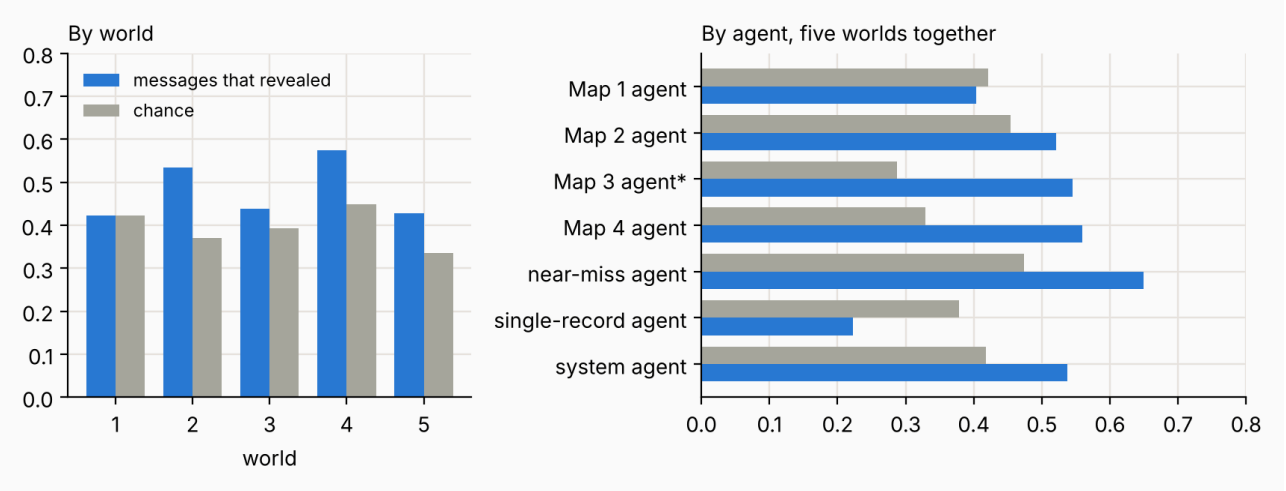

Figure 6. How often the agents' messages revealed something, against chance, in the present design. Left: each of five made-up worlds, with about 200 to 220 scorable messages per world. Right: each agent, five worlds together, with no pass rule set for each agent. *The Map 3 agent's messages point at records and are scored on a different rule. The near-miss and single-record agents read the ranking, not a map.

The conditions of this result matter. The agents' language model is a scripted stand-in that makes no call to a real model. At the setting used, the stand-in's own picks are random. Its messages are therefore mostly the agents' fixed rules, with about one rule finding in five missed. The result is evidence about those rules. It says nothing about how a real language model would do. The system agent also switches among three lines of reasoning with equal weight, and that setting is not settled.

By agent, the near-miss agent did best, at 65% against 47% by chance. The Map 4 agent came next among the agents scored on compounds, at 56% against 33%. The single-record agent did worse than chance, at 22% against 38%. It is the agent that section 2 described as right on 24% of its findings and credited with 67%. The Map 4 agent is the one right on 82% and credited with 28%.

The model also asked what happens if the scientists act on the suggestions. Three response policies were tried. One takes no suggestions, one takes all of them, and one takes only those from agents with a good track record. These policies are guesses made for the model, and no real scientist's behavior

went into them. In all five worlds, under all three policies, the top 10's share of truly good compounds came out the same, and so did the caffeinated hypotheses.

That result is built in, so it says nothing either way about whether acting on the suggestions helps. The top 10 that was scored is ranked at the base settings, and the hypotheses are chosen from the base run. The suggestions can change only the shape of the batch and the agents' shares, and those leave the base settings alone. Only those two outputs were compared. Against taking no suggestions, the agents' own findings changed by up to six points. The maps and the test list were not compared. In the self-running version of section 4.9 that limit is removed: the base settings and the top 10 the scientists see can change.

4.7 The test list

Each week caffeinated lists up to 12 tests worth running. Each comes with its compound, the assay, the reason and the records behind it. The list aims at places where one test could change the picture. One example is a top-10 compound with no cell or liver result. Another is a series in the top 40 with no cell test at all.

In the model, three comparisons declared before their runs asked one question. Does a team that chooses its tests from the list end up with a better top 10 than a team that tests straight down the ranking? They used thirty worlds in all, of three kinds. The list never beat the ranking by the margin set in advance. In the first comparison it did worse. In the other two it was a few points ahead in some worlds. On the NLRP3-shaped world it was 3 points ahead against a margin of 17, and the uncertainty around that lead spans zero. In one kind of world the ranking was already so close to the best possible result that no list could have met the margin.

Twice the list reached the ranking's level with fewer tests. In the third comparison that was partly because the list ran out of items. Saving tests was not the question any comparison was declared to answer. A comparison on cost would have to be declared first.

The list is therefore not shown to be a plan worth following. Whether it helps a scientist choose better, as one input among others, has not been tested. A model with no scientist in it cannot test that. Section 4.9 reports two orders of the list that follow what the system has learned.

4.8 The ledger of dated predictions

The ledger does not show anything the ranking hides. It is a check on the system itself. Each week the system writes down what it expects to happen, with a date, before the outcome can be known. Later weeks score those predictions against the public record. Over time this should tell the scientists how far to trust what the system expects.

In the model's earlier runs, the ledger scored worse than always saying fifty-fifty in every world. The agents' track records, which the learning step keeps separately, ran against the agents' real accuracy, as section 2 describes. Neither is yet a guide.

4.9 The learning loop, run in the model

The model now has a version in which the learning loop runs itself, as section 2 describes for the design. It is not the model's default, and the results in sections 4.1 to 4.8 do not come from it. Each week the learning step reads what the agents report and changes the next week within set ranges. It sets the assumed assay noise, which the base settings use, from the Map 2 agent's estimate of how noisy the activity reports are. The step also adds configurations to the batch around settings the system agent proposes. In this version every proposal of the system agent goes to the settings writer without passing the scientists. In earlier runs of the loop the batch grew from 6 configurations to about 8, of

which 36% to 38% were learned. As in the default, it sets each agent's share of the findings. The top 10 the scientists see comes from each compound's score averaged over the batch's configurations, so these changes reach it. Every change of a setting is logged, stays within its range, and is made only after the ledger's weekly scoring has run. Nothing changes in the first week.

So agents move settings the ranking uses, in two ways. The assumed assay noise follows the Map 2 agent's estimate, which is a fixed calculation, the spread of repeated public reports on the same compound and target. The batch's added configurations follow the system agent's proposals, which come from its language model, a scripted stand-in in the model. No language-model output becomes a number in the ranking formula, but the second route lets it choose settings that the scientists' top 10 is averaged over. Section 2 notes that whether the design lets agents propose such changes is not yet decided. The model allows both in this version, which does not settle the design question.

Two tests were declared before their runs. They checked how the loop works, not whether it improves the science, because made-up worlds cannot settle that. Both used five hard worlds and ten NLRP3-shaped worlds, each run with three different sets of reports. The first test compared 45 pairs of runs with the loop on and off. Its first version was declared with an error that made the two arms differ in the first week, and it was redone; the redone test reused the 45 loop-on runs. For this test the range of the assumed assay noise was widened from the 0.3 to 0.8 of section 4.2 to 0.3 to 1.2, so that the true value of 1.0 lay inside it. With the 0.3 to 0.8 range, earlier runs of the loop stopped at 0.8.

Every declared check of how the loop works passed except one. Every change of a setting matched a logged, allowed decision, every setting stayed within its range, the scientists' ranking, which uses in-house data, stayed marked private, and with the loop off nothing changed. In every pair, the top 10 the scientists see was the same as the loop-off run's in week 1 and differed from it in a later week. The assumed assay noise started at 0.5 and followed the Map 2 agent's estimate toward the true value. It ended at 0.95 on average in the hard world and 1.02 in the NLRP3-shaped world (Figure 7, left). The check that failed asked for it to end within 0.15 of the truth in 41 of 45 runs. It did so in 38. In the other seven, the agent's own estimate was that far off, and the setting followed it.

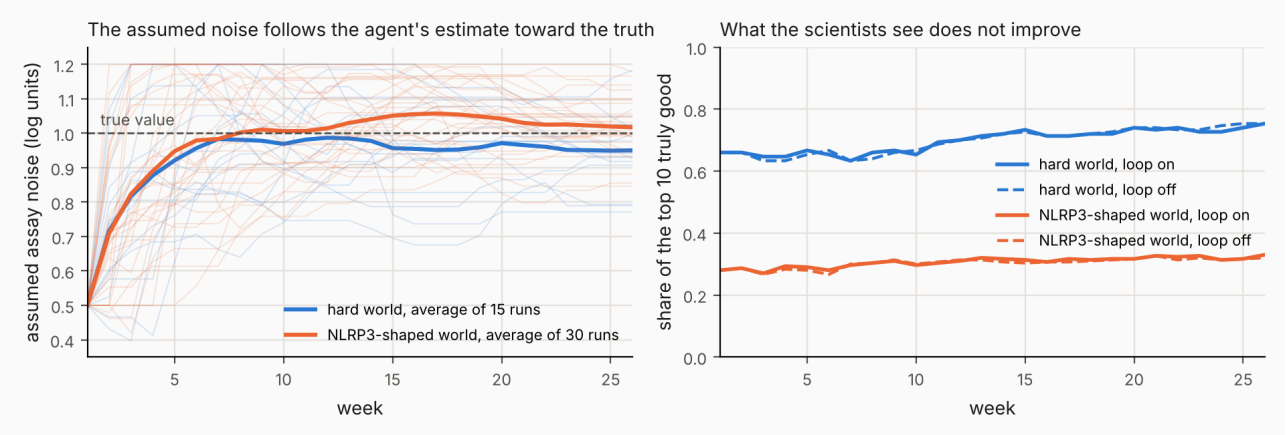

Figure 7. The self-running learning loop in the model. Left: loop-on runs only; with the loop off the setting stays at 0.5. The assumed assay noise is the standard deviation of a potency report, in log units, as the learning step sets it each week from the Map 2 agent's estimate. Thin lines are single runs, thick lines the averages, and the dashed line is the worlds' true value of 1.0. The range was widened to 0.3 to 1.2 for this test, and many runs touch 1.2 early on. Right: the share of truly good compounds in the top 10 the scientists see, week by week, with the loop on and with it off, averaged over the runs. Hard world: five worlds, each with three different sets of reports, 15 runs each way. NLRP3-shaped world: ten worlds, three sets each, 30 runs each way. Made-up data. The two kinds of world judge truly good differently (section 5), so their shares are not on the same scale.

What the scientists see did not get better. With the loop on, the share of truly good compounds in their top 10, averaged over weeks 6 to 26, differed from the loop-off runs by less than one point (Figure 7,

right). An earlier declared test of the loop set 2.6 points in the hard world and 2.8 in the NLRP3-shaped world as the smallest average difference that counts.

The second test tried two orders of the test list, in 45 sets of three runs with the loop on. In these runs the team ran the list's first tests each week. The orders are fixed rules that use what the system has recorded. The learning step does not choose between them. In one, tests on compounds where the ledger's predictions had been most wrong came first. In the other, tests most likely to move a compound across the top-10 line came first, judged with the learned assay noise. The first order departed from the list's own order from week 4 to 8 on, as the ledger's first predictions were scored. The second departed in the first week. In all four comparisons, two orders in two kinds of world, the top 10 came out lower than under the list's own order, by 0.7 to 3.1 points. Only the largest, the second order in the hard world, is bigger than the 2.6 points above, and there gains and losses were even, 7 and 7. The first order also has a risk: it steers tests away from compounds the ledger got right, which can include the leading compounds. One check of this test failed because it was worded wrongly. The model did what its specification says.

The messages moved in different directions in the two kinds of world. On the NLRP3-shaped world they beat chance by 6 points with the loop on, against 12 with it off. On the hard world they were at chance with the loop off and 2 points above it with the loop on. In an earlier set of runs, where the fall on the NLRP3-shaped world was 2 to 5 points, much of it came from the Map 3 agent, whose messages were 37 points above chance with the loop off and 26 to 31 with it on. Why is untested. One guess is that the configurations the loop adds change the maps the agents read.

So far, then, the loop runs as specified on made-up worlds, apart from how close the learned noise came to the truth, and it changes what the scientists see without improving it. Whether it helps on real data, or with a real language model as the agents' model, the model cannot show.

5. How to read the model's results

The model runs on made-up worlds, built so that its outputs can be checked. Most results here come from five versions of one world, the hard world. Each has 10 targets and 200 compounds in 20 series. Records arrive week by week over 26 weeks, and some of them are wrong. A hidden truth sits behind them that only the scoring can see. In the hard world about a quarter of all compounds are truly good.

The test-list comparisons also used two other kinds of world. One has looser series, and the other is shaped like an NLRP3 program, which section 4.9 also uses. In the test-list worlds truly good is judged on cell potency, and in the NLRP3-shaped world brain exposure also counts. Figures from different kinds of world are not on the same scale. The worlds were built by the model's author, so they may suit its design better than real data would.

Each kind of output is compared with its own reference. Map regions are compared with the band of ranks they are drawn from, ranks 11 to 40. Claims about compounds are compared with ranks 11 to 15, the compounds just below the line. Messages are compared with chance as section 4 defines it. The test list is compared with testing straight down the ranking.

Most measures in this document were set before the system's purpose was restated around showing what the ranking hides. They asked whether an output points at good compounds. Only the messages were measured on the restated question. Their measure covers both halves of it: good compounds outside the top 10, and compounds inside it that are not good. The Map 4 region and the count of new and good claims cover only the first half.

Some results were declared with a pass rule before their runs, so that they could not be explained away afterwards. These are the Map 4 regions, the three comparisons of the test list, the comparison of the

messages and the two tests of section 4.9. The three test list comparisons failed. The messages comparison passed with no margin and needs a stricter test. The two tests of section 4.9 each passed all but one check, as that section says. Other figures had no pass rule set in advance. These include the grouping of unmeasured compounds and the figures for each agent. They describe what happened and test nothing.

Results vary between worlds, sometimes widely. A single world can say the opposite of five worlds together. In the hard world the top 10 is already mostly good, which leaves few wrong compounds inside it for an output to reveal. There is more room below it, where about half of the compounds ranked 11 to 40 are truly good.

6. What the model cannot show

The model cannot show whether any output triggers a scientist's intuition. A made-up world has no scientist in it. The most the model can show is that an output points at something true that the ranking hides. It cannot show that real compounds and real Alzheimer's data behave like these worlds. It cannot show whether scientists will use any of the outputs. It cannot show how good a real language model would be as an agent, because the agents' model is a stand-in whose rates were guessed.

Two results are positive so far, both in the model. The decision map's region points at good compounds below the line more often than the band it is drawn from. The agents' messages reveal what the top 10 hides more often than chance in most worlds. With the top 10 taken from scores averaged over the batch (section 4.9), hard-world messages were at chance with the loop off.

Several outputs do not yet work in the model. The claims about compounds do no better than the compounds just below the line. The stress test's stand-in does not separate good claims from bad ones. The test list has not beaten the ranking. The ledger and the track records are not yet a guide. The self-running loop passed every declared check of how it works but one, and has not yet improved what the scientists see.

The next measurement is a declared comparison of the messages on the NLRP3-shaped world, with a margin set in advance.

Glossary

Ranking, top 10
: The order of all compounds by a formula whose weights are the scientists' own. "The top 10" means the top 10 of the base run, except in section 4.9, where it comes from each compound's score averaged over the batch's configurations.

Setting, configuration, base run
: A setting is a value the system is unsure of, such as assay noise, varied within ranges the scientists set. A configuration is one combination of settings. Configuration 0 uses the base settings, and its ranking is the base run.

Batch
: The set of ranking runs made each week, several for each configuration.

fred
: A reasoning engine that runs a language model locally and reads public evidence only. A scripted stand-in in the model.

Map
One of four weekly views of the batch. Map 1 compares fred's support with rank. Map 2 shows what depends on a setting. Map 3 shows records that many compounds rest on. Map 4 shows each compound's full picture.

Band
The group of ranks a map region is compared with, here ranks 11 to 40.

Caffeinated hypothesis
A claim a scientist would not get from the top of the ranking, put forward each week.

Stress test
An argument for and against a few hypotheses a week. A scripted stand-in in the model.

Agent
Software that reads the week's outputs, writes messages and proposes changes. The learning step decides which changes go in. In the model's default the agents only write messages. In its self-running version their reports change the next week through the learning step.

Message, digest
An agent's note to the scientists, and the week's messages collected together.

Test list
Up to 12 tests a week worth running, each with its reason and records.

Ledger
The record of the system's dated predictions and their later scores.

Track record
The learning step's score for each agent, from how its earlier predictions turned out. In the model it sets how many findings the agent may write next week. In the self-running version it scores claims about a compound's next public report.

Self-running version
The version of the model in which the learning step changes the assumed assay noise and the batch, and the scientists' top 10 comes from scores averaged over the batch. Section 4.9.

Made-up world, hard world
A simulated discovery program with a hidden truth, used to score the model's outputs. The hard world is the one used for most results.

Truly good
In the hard world, a compound whose hidden chance of success is at least 0.7.

Reveals, chance
Defined at the start of section 4.

Names in the design record

This table links the plain names used here to the names in the phase-0 design record. That record will hold the engine specifications and evidence tables, and it is still to be written. Until it is, the phase-0 model guide r7 holds them, with the agents described as they were before the model's agents were limited to messages.

Name in this document	Name in the design record
Gathering and measuring	E1 intake, E2 measures
The ranking	E3 ranking
The batch	E6B batch design
fred, and the time set aside for it	E4 fred stand-in, E5P packer, E5G review gate, E11 orchestrator
Caffeinated hypotheses	E6 caffeination (the method in use is the counterfactual reading, and the consensus reading is the standard one)
Stress test	E12 SIM stand-in
Re-asking fred about close calls	E8P probe
Maps 1 to 4	E7 spatial analysis (the decision map is Map 4 reading r3)
Map 1 to Map 4 agents, near-miss agent, single-record agent	E8X local agents L1 to L4, champion, fragility
System agent	E8S system agent
Learning step, track records	E8L Learn
Settings within the scientists' limits	E8W settings writer
Digest	E10 cycle report and agent log
Test list	WT, engine E14, the worth-testing list
Ledger	E9 prediction ledger, with E13 backtest
Self-running version of the learning loop	E8L Learn v5 and v6, E8W settings writer v2, E6B v4, E2 v6 and E3 v4 (the batch's top 10), WT v3 (the test list's two orders, tested separately); phase-0 update plan r6, item 10
Response policies	SR, the scientists' stand-in, with policies none, follow and selective (an instrument outside caffeinated)
The team's choice of tests in the comparisons	TC team choice and SV scope view (instruments)
Scoring against the hidden truth	EV evaluation (an instrument)
Made-up worlds	W. Worlds 1 to 5 here are the hard world's seeds 7, 11, 13, 17 and 19. The other kinds are the loose-series world and the NLRP3-shaped world

Provenance and revision

Draft r6, 4 November 2026. It corrects section 4.9 of r5, which said that agents move a setting the ranking uses only through the Map 2 agent's estimate, a fixed calculation. The batch's added configurations also follow the system agent's proposals, which come from its language model, so in this version language-model output chooses settings that the scientists' top 10 is averaged over. Section 4.9 now says so. This was read directly from the model's learning step on 4 November 2026. Nothing else changed.

Draft r5, 4 November 2026. It adds section 4.9 and Figure 7, which report the model's self-running version of the learning loop, built and tested that evening after Raminderpal Singh reopened the model's work for it. Sections 1, 2, 3, the preamble of section 4, sections 4.6, 4.7, 5 and 6, three glossary entries, one new glossary entry and the names table point to it or were brought into line. The numbers in section 4.9 and in the new text of sections 2 and 6 come from model runs made on 4 November 2026 and read directly: 135 runs of the loop on and off (90 of them form the 45 pairs), 135 runs of the test list's orders, the 60 and 225 runs of two earlier tests of the loop, and a diagnosis of the messages that re-read the 225. They were made for the model's work plan, not for this document. Figure 7 was drawn from the 90 paired runs. Figure A is unchanged. Its design statuses come from the design files, and a result of the model is not a design decision. An independent review checked these changes before delivery, and its corrections were applied.

Draft r4, 4 November 2026. It follows Raminderpal Singh's direction of that evening that the system runs itself, with the agents' learning driving it, while the scientists monitor it, adjust it and extract what they need. Figure A now shows the learning step's changes feeding back into steps 2 to 5, and its caption is changed to match. Sections 1, 2 and 4.6, the preamble of section 4 and two glossary entries say the same, and say that the model runs a narrower loop. That the learning step may change the uncertain settings, what fred is asked next and the order of the tests is taken from the note on caffeinated's engines, r3, and the module decomposition, r2, read directly on 4 November 2026. Each agent's share of the findings comes from the model and from Raminderpal Singh's decision of 3 November. The learning step is now marked partly designed, as the engines note gives it. No result in section 4 changed. An independent review checked these changes before delivery, and its corrections were applied.

Draft r3, 4 November 2026. It changes only Figure A and its caption, at Raminderpal Singh's request that the computational methods, caffeination and spatial analysis be easier to see. Figure A now says what each part of the system does in each step, names the methods the design gives it, and marks its design status. The methods and statuses come from two design files read directly on 4 November 2026: the note on caffeinated's engines, r3, and the module decomposition, r2. Where those files predate later decisions, the decisions then in force are followed: the agents use a language model, they only message the scientists, and the one change the system makes on its own is how many findings each agent may write. For that reason the agents and the learning step are marked not yet designed. The stand-ins named in the caption come from the model's own descriptions of its measurement, caffeination and spatial-analysis engines, read the same day. The rest of the text is unchanged from r2. An independent review checked the figure and its caption against these sources before delivery.

Draft r2, 3 November 2026. It corrects one paragraph of r1 in section 4.6. r1 called the result that acting on the agents' suggestions changed nothing "partly built in". It is wholly built in, as the review of the design record found: the top 10 scored and the hypotheses both rest on the base settings, which no suggestion can change. Draft r1 was written the same day by Claude for Raminderpal Singh from outline r2, which he agreed that day. It replaces the part of the phase-0 model guide r7 written for scientists.

Outside Figure A, section 4.9 and the text r5 added, every number in the text comes from records read directly on 3 November 2026: the phase-0 session state r6, the update plan r4, the engine cards r6, the text of guide r7 and the stored result of the messages comparison. Figure 6 and the message figures were computed afresh from that result. Figures 1 to 3 are images from guide r7's build kit, cropped and, for Figure 3, relabeled. Figures 4 and 5 were redrawn from that kit's data. No model run was made for this document. The description of Krauss's book was checked against the publisher's description by web search on 3 November 2026. An independent review checked the draft against these records before delivery.