

Agents reading a mapped brain: pre-registered tests against published fly experiments, named and blind

Dan Siroker¹

¹Big Brain Lab. Correspondence: bigbrainlab.ai

Abstract. A whole-brain leaky integrate-and-fire model of the fruit fly, built on the FlyWire connectome, predicts many published perturbation results. We asked whether language-model agents that read the wiring and run experiments on that model predict published results better than the simulation read directly, and how much of any gain survives hiding the cell-type names that let a language model recognize the papers. The tests ran under protocols hashed before the first scored run, five of them with pre-registered pass rules; three of the five primaries failed. On a frozen key of 259 experiments from 24 papers, named agents scored 0.902 AUROC against the simulation’s 0.831 (+0.071 [+0.023, +0.120] over items; the paper-clustered interval includes zero). With every name replaced by an opaque code the gain fell to +0.039 overall, below the pre-registered bar, and to +0.090 [+0.037, +0.148] on the 90 experiments the model’s authors had not used (paper-clustered [-0.010, +0.190]); a names-only control with no map and no model scored 0.786 on the 86 of those 90 it answered. The blind figure still includes about eight items per replicate recognized from wiring motifs alone, so it bounds the map’s own contribution from above. A library of 350 agent-written neuron papers let a reader without tools reach the agent’s score on the 173 experiments it covered, with no detected benefit from neighboring types’ papers. A key of 53 experiments from 2026 papers did not separate recall from inference, because the names-only control scored 0.815 on them. On the MICrONS mouse volume, where no connectome-based perturbation simulator exists, an agent with the whole map did no better than a rule using the tuning of nearby imaged cells. A compound report with named mechanisms and tissue predictions was generated end to end in hours. On this retrospective benchmark, reading a mapped brain with agents gives a modest gain over simulating it, and only blinding separates that gain from memory of the literature.

Introduction

Three wiring diagrams at synaptic resolution bear on this study: the adult fruit fly brain^{1,2}, the male fly central nervous system³, and a cubic millimeter of mouse visual cortex with activity recorded from the same neurons⁴. For the fly, Shiu and colleagues turned the FlyWire connectome into a whole-brain leaky integrate-and-fire model, with every synapse signed by its presynaptic neuron’s predicted transmitter^{5,6}, and showed that driving sensory neurons in the model reproduces the activation of downstream neurons and behaviors reported in the literature. In an earlier pre-registered test we found that this predictive power depends on the wiring at the level of cell types: a model whose connections are shuffled within brain regions, keeping every region-to-region count, falls to chance⁷.

Whether a map of this kind can be read for mechanism is a separate question from whether it can be simulated. The nearest existing example of a completely mapped network that has been read is an artificial neural network. The dominant account of such networks has been that their units superpose unrelated features. A 2026 working hypothesis from automated investigation at scale, one written investigation per unit, attribution graphs with the network’s own units as nodes, ablations, and explanations that could lose to a near miss, argues instead that most units in the models examined are legible⁸. We take that method, not its conclusion, as the starting point. The method rests on being able to read every unit and run every intervention. A whole-brain simulation on a connectome is the one biological network where that is true in software.

We therefore ran the method on the fly model and measured what it produced against published experiments, under protocols written down and hashed before the first scored run. The questions were four. Does an agent that can read the wiring and run experiments on the model predict published perturbation results bet-

ter than the raw simulation? How much of any gain is the map and the model, and how much is the language model’s memory of a literature it has read? Does what the agents learn accumulate, so that a written account of one cell type helps a reader with no model access predict experiments on another? And does any of this transfer to a mammalian map that has activity but no model to run? A fifth product of the work, a report on a simulated compound with mechanisms and tissue predictions, is described because it is the artifact the method is for.

Five tests below were pre-registered with a frozen answer key, fixed arms, a stated primary endpoint and a success rule: the agent-versus-simulation comparison, the name-blind rerun, the post-cutoff key, the compounding test and the MICrONS test. The blinding ladder was pre-registered with fixed contrasts and no pass bar. The second-model replicate, the old-model ladder and the recognition probe were run after the fact and are exploratory. For every test the hashes of every harness file were written before the first scored item, and every change after a freeze is logged with its time and reason. We report the primaries that failed with the same prominence as those that passed.

Results

The answer key and the raw simulation

The answer key holds 259 experiments from 24 papers^{5:10_13:15_33}, each a stimulated or silenced population, a readout neuron or set, and a published outcome coded as activation (label 1, 71 items), no effect (label 0, 153) or suppression (label -1, 35). 169 items are the experiments Shiu and colleagues used to validate their model; 90 are not Shiu-validation items in our key and come from 22 papers; one of those papers, Hampel et al. 2015¹⁸, also supplies three validation items, so the split is by experiment, not by paper. These 90 are the independent set. The simulator is a batched re-implementation of the published model on the FlyWire v630 wiring (127,400 neurons, 14.7 million connections, 52.8 million synapses) that reproduces the reference Brian2 implementation to a per-neuron rate correlation of 0.9998 (Methods). Each experiment drives the stimulated population with 100 Hz Poisson input for one second (50 Hz for the 12 items coded low intensity and 200 Hz for the 3 coded high), silences by zeroing outputs, and reads out the mean rate of the readout set; the score is the area under the receiver operating characteristic (AUROC) for activation against the other two labels, on the signed readout response.

The raw simulation, arm S, scores 0.831 on all items, 0.914 on the Shiu validation items and 0.672 on the independent items. The independent score is the reference against which everything below is measured; on those 90 items the model is right about two thirds of the time it is asked to rank a published pair. Shuffling whole connections within cell-type blocks, keeping each neuron’s output count and every block-to-block count, leaves the score unchanged (0.838 all, 0.669 independent); shuffling within brain regions drops it to chance (0.513, 0.509), with the fraction of simulations that spread activity downstream unchanged but fewer neurons active (median 27 against 68) and the readout almost never reached⁷. A map at the resolution of FlyWire’s cell types is what carries the prediction.

Agents with the wiring and the simulator beat the raw simulation

Arm A gives a language model (claude-opus-5-5, high effort) six tools over the wiring and the simulator: describe a neuron, list its top partners, trace signed connections up to three hops, simulate a drive with optional silencing, co-activation and a subthreshold baseline, ablate candidate intermediates one at a time in a batch, and screen candidate input sets one at a time. The agent sees the item’s design (populations, side, manipulation, method, measurement) and the model indices of each set, never the citation, the figure or the label; it is told to rest its answer on the wiring and the simulations and to flag any reliance on prior knowledge. It returns probabilities for the three outcomes, a written mechanism, the path, the evidence it rests on, and what it believes the simulation alone predicts. Two independent replicates were run per item and their probabilities averaged. Arm L, the leakage control, is the same language model given only the

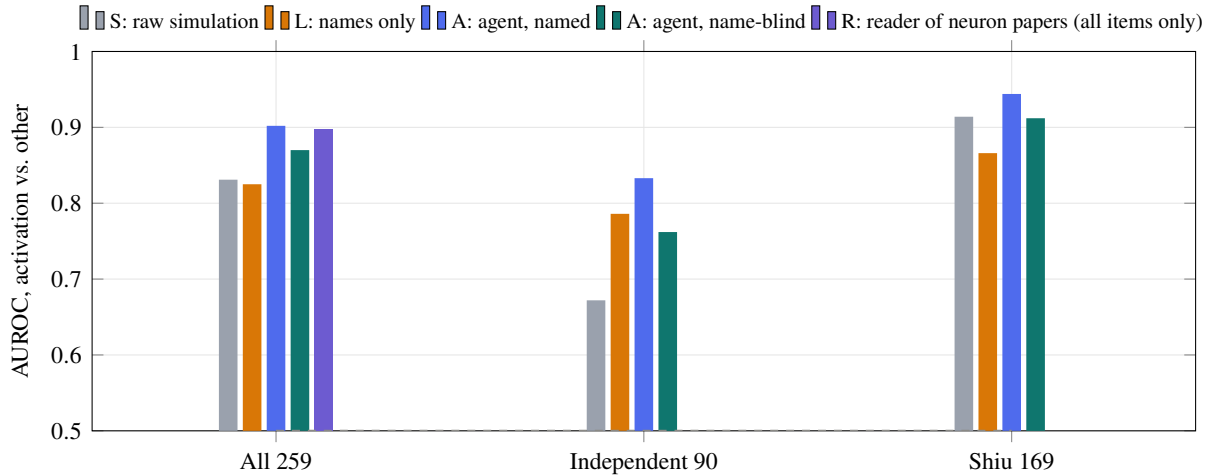


Figure 1: Predicting 259 published fly perturbation results on the Shiu 2024 whole-brain model. AUROC for activation (label 1) against no effect or suppression, by subset. Grey: the raw simulation scored on its signed readout response. Orange: the language model given only the cell-type names and the design, with no map and no model (255, 86 and 169 items in the three columns; four safety refusals); the grey bar rests on all 259, and the paired contrasts in Table 1 use the items both arms scored. Blue: the agent with six tools over the wiring and the simulator, mean of two replicates. Green: the same agent with every cell-type name, transmitter, class and region string replaced by an opaque code, mean of two replicates. Purple: a reader with no tools that sees only a library of 350 neuron papers written by the agent from the tools alone, at full coverage (the neuron-paper section below), shown for the full set only: the library covers the 173 subesophageal items, and on the other 86 the reader’s names-only answer is carried over unchanged, so the independent and Shiu columns would not be like-for-like (on the 173 covered items it scores 0.941 against the agent’s 0.943). Dashed line: chance.

cell-type names and the design, with no map, no model and no tools.

The primary passed. The agent scored 0.902 on all 259 items against the simulation’s 0.831, a difference of +0.071 with a paired item-bootstrap interval of [+0.023, +0.120], above the +0.05 bar (Fig. 1, Table 1). Each replicate passed alone (+0.079 [+0.032, +0.128] and +0.069 [+0.022, +0.118]), and so did the contrast with the 15 development items removed (+0.076 [+0.024, +0.129]). The pre-registration also asked for the interval with resampling by paper, and that interval touches zero ([−0.006, +0.276]): 24 papers, one of which supplies 127 items, do not support the pass under clustering. On the independent items the agent gained +0.162 [+0.053, +0.275]; on the Shiu validation items +0.030 [+0.001, +0.072].

The leakage check passed by its rule and failed where it mattered. The rule was that the agent must beat the names-only control. On all items it does (+0.079, [+0.035, +0.126]). But the names-only control, with no map and no model, scores 0.786 on the 86 independent items it answered, where the raw simulation scores 0.688 and the agent 0.843, and the agent’s margin over it there is +0.057 with an interval that includes zero ([−0.008, +0.128]). The agent flagged reliance on prior knowledge on 79 of the 90 independent items in at least one replicate (136 of 259 in replicate 1, 150 in replicate 2), and the names-only control reported specific prior knowledge on 27 of 86. Of the 21 corrections in replicate 1 where the agent overrode a definite model-alone class, 18 were on independent items and all 18 carried the prior-knowledge flag; the seven wrong overrides were flagged too. Examples are the seven negative glomeruli of Fişek and Wilson¹⁰, where the model fires the lateral-horn readout for every projection neuron because activity runs away through antennal-lobe local neurons the transmitter prediction signs as excitatory, and the agent showed by ablation that the response came from other projection neurons, answered no, and also said it recognized the paper; and the DPM to Kenyon-cell item of Haynes and colleagues¹¹, where the model signs DPM excitatory and the agent knew it is GABAergic. These were real mechanistic corrections of the model and also memories, at once.

Table 1: Headline scores and the pre-registered contrasts. AUROC for label 1 against labels 0 and -1 ; intervals are paired 95% bootstraps over items (10,000 replicates) and, in the second bracket where given, over papers. The primary was A minus S on all items, with a pass rule of at least +0.05 and an item interval excluding zero.

	All (259)	Independent (90)	Shiu (169)
S, raw simulation	0.831	0.672	0.914
L, names only	0.825 (255)	0.786 (86)	0.866 (169)
A, agent, named (mean of two replicates)	0.902	0.833	0.944
replicate 1 / replicate 2	0.910 / 0.900	0.839 / 0.838	0.955 / 0.936
A, agent, name-blind (mean of two)	0.870	0.762	0.912
A – S (primary, pre-registered)	+0.071 [+0.023, +0.120] [-0.006, +0.276]	+0.162 [+0.053, +0.275] [-0.027, +0.399]	+0.030 [+0.001, +0.072] [-0.002, +0.068]
A – L (leakage check)	+0.079 [+0.035, +0.126]	+0.057 [-0.008, +0.128]	+0.078 [+0.006, +0.159]
L – S	-0.017 [-0.083, +0.049]	+0.098 [-0.034, +0.230]	-0.048 [-0.136, +0.036]
Blind A – S	+0.039 [+0.010, +0.070] [-0.009, +0.138]	+0.090 [+0.037, +0.148]	-0.001 [-0.030, +0.033]
Blind A – named A	-0.032 [-0.072, +0.009]	-0.071 [-0.163, +0.018]	-0.031 [-0.065, -0.007]

On the Shiu validation items, where the names-only control is weakest relative to the simulation (0.866 against 0.914), the agent beats both the simulation (+0.030) and the control (+0.078, [+0.006, +0.159]), the cleanest evidence in the named run that running experiments on the model adds something: the agent reranked outcomes the simulation had ranked wrongly, for example weak responses of a few hertz that the agent discounted after an ablation showed a single fragile path.

Secondary endpoints. With a subthreshold background drive on the readout, which lets a decrease be measured in a model with no spontaneous activity, the agent separates suppression from no effect at 0.945 AUROC against the driven simulation’s 0.630 (+0.315, [+0.206, +0.429]); the names-only control reaches 0.935, because most suppression items are well-known results. Three-class accuracy at the model’s thresholds is 0.807 for the agent, 0.710 for the simulation and 0.651 for the control (a constant “no effect” scores 0.59). The agent is better calibrated (multi-class Brier 0.307 against 0.444 for the control; a constant at the base rates scores 0.56). The two replicates agree on the class on 96.9% of items (mean absolute difference in $P(1)$ 0.048, correlation 0.969); they differ mainly in which simulations they chose to run. Per item the agent made 7.5 tool calls (median 7, range 4 to 12), 4.1 of them simulations, in 2.1 minutes and \$0.14 of API spend; no item approached the 40-call or 35-minute budgets. A second language model (gpt-6-astra), run through the Codex command-line interface with the same six tools served over the Model Context Protocol in a replicate that was not pre-registered, scored 0.912 (0.844 independent, 0.960 Shiu), +0.081 [+0.041, +0.126] over the simulation and +0.010 [-0.020, +0.041] over the first agent, with 89% class agreement and the same counts of right and wrong overrides, 25 and 7, on mostly the same items; it flagged prior knowledge on 8 of 259 items while making the same memory-based corrections, so the self-report is not comparable across models.

Hiding the names: where the gain comes from

To separate the map and the model from the language model’s memory, the agent was rerun with every annotation string (cell type, hemibrain type, class, Shiu name, predicted transmitter, super class, input neuropil) and every cell-type name in the item replaced by an opaque stable identifier, the item identifier hashed, and the free-text method, measurement and design notes removed; it kept the manipulation, the standard protocol, the side, the model’s sign, synapse and partner counts and the model indices. A check on two development items found zero occurrences of 30 real names across the prompt, the tool results and the raw event stream. Two replicates, 518 runs, no refusals, no fallbacks.

The blind primary failed the +0.05 bar and the gain was not zero. Blind A minus S on all items is +0.039 [+0.010, +0.070]: the interval excludes zero, the point estimate is below the bar, and the paper-clustered interval touches zero. On the 169 Shiu items the blind agent is the raw simulation (0.912 against 0.914); the named agent’s +0.030 there is gone, and blind minus named is -0.031 with an interval excluding zero ([-0.065, -0.007]). On the 90 independent items the blind agent still beats the simulation by +0.090

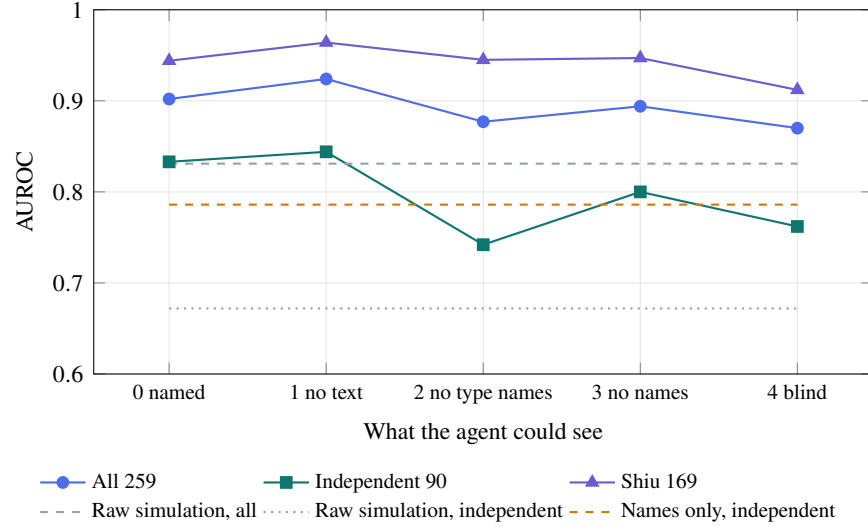


Figure 2: The blinding ladder. AUROC of the agent at five levels of what it could see (Table 2): level 0 the named run; level 1 item identifier and all free text hidden, names kept; level 2 cell-type names hidden, transmitter, super class, neuropil and design notes kept; level 3 all names and annotations hidden, design notes kept; level 4 fully blind. Levels 1 to 3 are one replicate each (259 runs per level), levels 0 and 4 the mean of two. Levels 1 and 2 hide different things and are not ordered with each other, so the x axis is a sequence of conditions rather than a monotone scale. Reference lines: the raw simulation on all and on independent items, and the names-only control on independent items.

[+0.037, +0.148], about half the named agent’s +0.162. The names-only control alone scores 0.786 there and the blind agent 0.762: an agent with the map and the model and no names does about as well as the names alone, by a different route (on the 86 items both scored, blind minus names-only is -0.004). Prior-knowledge flags fell from 136 and 150 of 259 items in the two named replicates to 40 and 39 (independent items from 71 and 72 of 90 to 8 and 8); of the remaining flags about eight per replicate are structural recognition from the wiring alone (Kenyon-cell, APL and projection-neuron motifs, a DPM-like cell, a Delta7-like cell) and the rest general priors, so the blind number is a lower bound on the leakage removed, not a clean zero. The named agent’s celebrated corrections, the Fişek and Wilson negatives, DPM, DN1p, were memory: the blind agent followed the model on them and got them wrong. Corrections of the model’s class that turned out right fell from 25 and 24 in the named replicates to 5 and 3.

The pattern is consistent with the cell-type names carrying most of the independent-item gain; the design text and the annotations are not resolved by one replicate per level. A five-level ladder (Fig. 2, Table 2) hid the information in steps, one replicate per intermediate level. Hiding the item identifier and every free-text field while keeping the names (level 1) cost nothing: 0.924 all, 0.844 independent, 0.964 Shiu, equal to or above the named run on every subset, and the agent still flagged prior knowledge on 69 of the 90 independent items, recognizing the experiment from the names alone. Hiding the type names while keeping the transmitter, the super class, the neuropil and the design notes (level 2) took the independent items down to 0.742, the same as the fully blind 0.762 (-0.020 , $[-0.096, +0.051]$) and 0.091 below the named agent ($[+0.017, +0.170]$), the one ladder contrast whose item interval excludes zero in the named direction. The largest movers were the same seven Fişek and Wilson items: $P(1)$ of 0.12 to 0.25 named, 0.30 to 0.55 at level 1, 0.75 to 0.85 at level 2 and about 0.70 to 0.80 when blind. At level 2 the agent still recognized designs from the free text (“P2X2 activation of DANs with sytGCaMP imaging of gamma KC axons”), but recognizing the design without knowing which cell types were in front of it did not tell it the answer. Hiding the transmitter, super-class and neuropil strings on top of level 2 (level 3) scored higher, not lower (0.800 independent; +0.058, $[+0.003, +0.118]$, driven by the same Fişek items and the eight clock-circuit items),

Table 2: The blinding ladder and its pre-registered contrasts. The ladder’s pre-registration fixed the contrasts and set no pass bar, so the run is descriptive. Levels 1 to 3 one replicate each; item and paper bootstrap intervals. The post hoc adjacent-step contrasts use the same bootstrap.

Level	All	Indep.	Shiu	A – S, all (item; paper)
0 named	0.902	0.833	0.944	+0.071 [+0.023, +0.120]; [-0.006, +0.276]
1 id and free text hidden	0.924	0.844	0.964	+0.093 [+0.052, +0.137]; [+0.028, +0.276]
2 type names hidden, context kept	0.877	0.742	0.945	+0.046 [+0.006, +0.086]; [-0.021, +0.133]
3 all names hidden, notes kept	0.894	0.800	0.947	+0.064 [+0.020, +0.109]; [-0.001, +0.206]
4 blind	0.870	0.762	0.912	+0.039 [+0.010, +0.070]; [-0.009, +0.138]
Contrast (independent items)				
named – level 2	+0.091 [+0.017, +0.170]; paper [-0.024, +0.262]			
level 2 – blind	-0.020 [-0.096, +0.051]			
level 1 – blind	+0.082 [+0.021, +0.153]; paper [-0.019, +0.249]			
level 3 – level 2 (post hoc)	+0.058 [+0.003, +0.118]; paper [-0.019, +0.168]			
level 3 – blind	+0.038 [-0.033, +0.109]			

which with one replicate per level is not interpretable as a gain and leaves the separate contributions of the annotations and the design text unresolved.

On the Shiu items the measurement text is as good as the names: levels 1, 2 and 3 hold the named agent’s Shiu gain (+0.050, +0.031, +0.034 over the simulation) and the blind run lost it (-0.001). At levels 2 and 3 the Shiu measurement reads “behaviour (rostrum extension, T7121-controlled)” (the key’s spelling) with the readout coded T7121: the text tells the agent the readout is the proboscis motor neuron, which is what the names told it before, and 135 to 137 of its 169 Shiu explanations at those levels mention rostrum, proboscis, PER or motor neuron, against 50 of 169 when blind. That is legitimate experimental context, worth +0.03 to +0.05 AUROC on the screen.

Taken together, on the 90 independent items the named agent’s +0.162 over the simulation is accounted for as about +0.090 retained under full blinding, about +0.07 to +0.09 from the type names, and nothing detectable from the annotations or the scrubbed design notes. Accuracy and calibration are worse fully blind than named, with intermediate levels that are not monotonic, while AUROC on the Shiu items holds: without the names the agent’s probabilities are less sharp even where its ranking holds (three-class accuracy 0.807 named, 0.714 blind; Brier 0.307, 0.448).

Is the names-only score recall of the papers?

The names-only control’s 0.786 on the independent items could be either recall of the specific papers or knowledge of fly biology. Two tests were run.

A key of 53 experiments from papers first public in 2026 did not separate the two. From 1,117 Europe PMC records, every fly preprint posted to bioRxiv from March 1, 2026 (948, of which 188 neuroscience), a sweep of about 2,260 records citing FlyWire and an eLife sweep, 53 papers were read and 980 candidate experiments screened; 53 items from 12 papers^{34–45} survived, 34 of them first public on or after June 1, 2026, of which 33 are scorable by the simulation and form the strict cut, and 19 from March to May. Most 2026 fly experiments read out behavior with no identified readout neuron, or are genetic, pharmacological, larval or in the ventral nerve cord, and 14 apparent 2026 papers have 2023 to 2025 preprints; 35 of the 53 items come from three papers. The pre-registered pass rule was a names-only AUROC at or below 0.60 on the strict cut. The names-only control scored 0.815 [0.639, 0.959] on the strict cut and 0.852 on all 52 scorable items, at least as high as on the earlier independent items; the agent with the map and the model equaled it (0.793 and 0.852; A minus L -0.022 [-0.120, +0.073] and -0.001); the blind agent fell toward the simulation

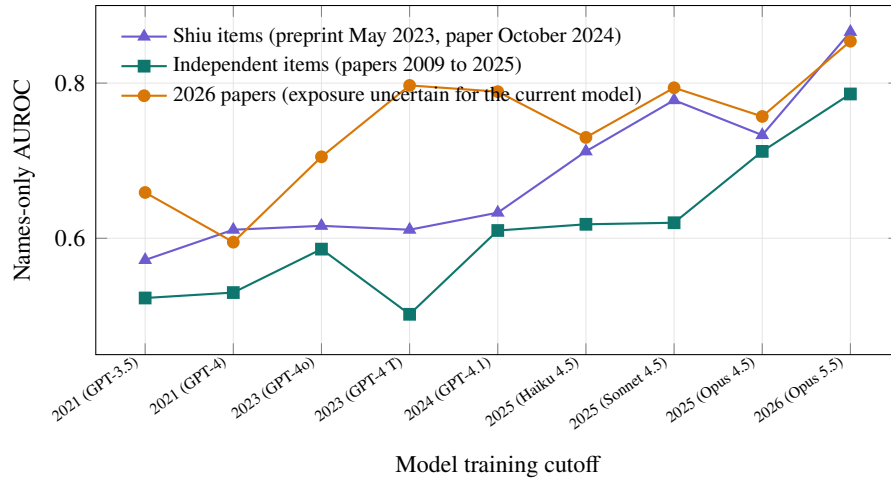


Figure 3: Names-only prediction by models of increasing training cutoff. The names-only arm, prompt and scoring unchanged, answered by every older model still served on October 17, 2026, with documented cutoffs (Methods). Purple: the 169 Shiu validation items. Green: the 90 independent items. Orange: the 53 items from papers first public in 2026; 34 are first public after June 1, 2026, and the current model’s cutoff is given only as June 2026, so exposure is uncertain for it and excluded for the older models. The 2021 models cannot have read the Shiu preprint; the curve rises with cutoff year with no step at the May 2023 preprint or the October 2024 paper, and the current model’s advantage is as large on the unseen 2026 papers as on the seen ones.

(0.678, 0.780); and the simulation itself scored 0.713 and 0.685. The control’s reasons name sources from 2017 to 2019 and “Shiu et al. predicted”, never a 2026 paper, so the diagnostic for a mis-stated training cutoff is negative. The 19 March-tier items are known results re-reported in 2026 (names-only 0.964 against the simulation’s 0.393); on the strict cut the control separates the label-0 items as well as the label-1 ones (0.864 for 1 against 0 only), so those items are guessable from the population names and the design rather than recalled from the paper. The key did not do its job as a contamination test. A usable one needs experiments without precedent, negatives and sign reversals in circuits with no published parallel, and about three times as many: 33 strict items give a 95% interval of about ± 0.13 on any contrast.

Models that cannot have read the papers do not get the names-only score, but the gap is capability more than recall. The names-only arm was rerun unchanged with every older model still served (Fig. 3; Extended Data Table 1). The two models with September 2021 cutoffs, which cannot have read the Shiu preprint of May 2023 or the paper of October 2024, scored 0.57 and 0.61 on the Shiu items; the current model scores 0.866. But for six of the eight older models the current model’s advantage is not measurably different on the 53 items from 2026 papers than on the Shiu items (+0.06 to +0.26) as on the Shiu items (+0.09 to +0.29): difference-in-difference intervals span zero for six of eight models, and the two exceptions show the same excess on the independent items from 2009 to 2023. The Shiu curve rises smoothly with cutoff year with no step at either Shiu publication date. Asked directly which paper each experiment is from and what it found, the current model names Shiu 2024 for 127 of 127 Shiu-2024 items and the right paper for 85% of all key items (2% of the 2026 items), yet its recalled finding matches the label on only 49% of the Shiu items, where always answering “no change” scores 70%: it recognizes the screen and its base rate, not the per-line results. On the independent items recall is more real (right paper 87%, finding matches the label 60%, against 38.9% for always answering “no effect” and 42.2% for the majority class), and the named agent’s gain over the simulation there sits on the recognized items: +0.19 [+0.07, +0.32] on the 76 where it both reported recognition and named the key paper, +0.03 [-0.10, +0.18] on the 14 where it does not, in agreement with the blind run. The blind numbers remain the ones to quote.

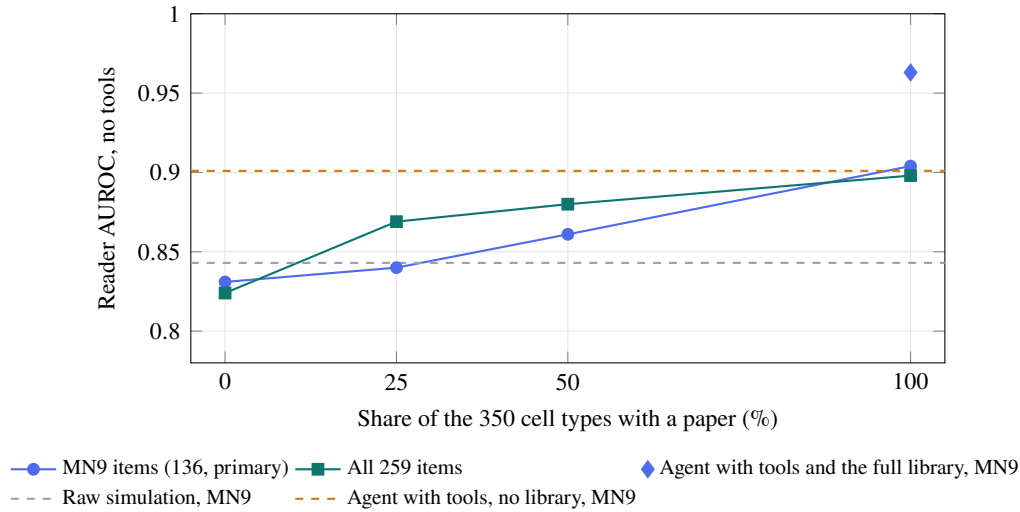


Figure 4: Coverage and compounding. A reader with no tools predicts each experiment from its design plus the neuron papers a library selects among the covered cell types, at 0, 25, 50 and 100% coverage of 350 types around the proboscis motor neuron in one random order (one replicate per level). Blue: the 136 MN9 items, the pre-registered primary (pass required +0.10 at full coverage with an interval excluding zero; observed +0.073 [-0.039, +0.201]). Green: all 259 items (+0.074 [+0.036, +0.114]). Diamond: the tool-using agent given the same library (A+P), 0.963 on MN9. Reference lines: the raw simulation and the tool-using agent without the library.

Neuron papers: a written account per cell type, read blind

The method’s product is not a probability but a written mechanism. In the first test, for each of 16 readout sets with at least three items (200 items), an agent with the tools and no item or label wrote a neuron paper of 1,900 to 2,230 words from the tools alone, with its own synapse counts, driver tables at three rates, ablation attributions and limits; a fresh reader with no tools then predicted each item from the paper and the item’s design. Over the 197 items it answered the reader scored 0.787, below the simulation (0.837, -0.050 [-0.132, +0.025]) but above it on the 41 independent items (0.824 against 0.618, +0.206 [+0.012, +0.408]). Where a readout has a few well-defined inputs the paper is a usable product: for the lateral-horn readout of Fişek and Wilson and Dolan and colleagues^{10;12} the reader scored 0.96 and 92% accuracy from the paper, against the raw simulation’s 0.78 and 31%. It failed for the proboscis motor neuron MN9, whose 134 items ask about 106 different cell types of the subesophageal zone: the paper could enumerate only the inputs its author chose to test, the reader said the paper did not cover 88 of the 134 items, and it fell to 0.68 against the simulation’s 0.84. An explanation that could lose did lose wherever the question was outside what the author had screened.

On the 173 experiments it covers, a library of 350 papers brings a reader to the tool-using agent’s score; there is no evidence that papers compound across types. The second test wrote one paper for each of 350 cell types of the subesophageal feeding circuit read out at MN9 (43 direct inputs with at least 30 synapses, 257 two-hop types with at least 200, the 106 Shiu types, MN9 itself and six item-named sets; 1,108 neurons; median 1,924 words, 32 tool calls and 19 simulations per paper), and read every item at 0, 25, 50 and 100% coverage in one random order (Fig. 4). On the 136 MN9 items the reader, a command-line rerun of the names-only arm that reproduced it (+0.004 [-0.032, +0.043] on the 255 items both scored), went from 0.831 with names only to 0.904 with the full library, +0.073 with an interval [-0.039, +0.201] that includes zero (16 positives): the pre-registered primary failed. On all 259 items the same contrast is +0.074 [+0.036, +0.114] and the slope across the four levels +0.067 per unit coverage [+0.032, +0.105], and at full coverage the reader with no tools matches the tool-using agent (R100 minus A -0.004 [-0.033, +0.025]; on the 86 items the library never covers the reader’s answer is its names-only answer carried over (0.773 against the

agent's 0.821 there), and on the 173 subesophageal items the library covers the reader scores 0.941 against the agent's 0.943; the study was not designed as an equivalence test) and beats the simulation (+0.067 [+0.016, +0.120]) and the names-only control (+0.075 [+0.029, +0.125]). But the gain arrives with the item's own paper, not with its neighbors': in the pre-registered indirect-coverage test, items whose stimulated type had no paper but which received other papers through cross-reference gained nothing (-0.024 [-0.133 , $+0.071$] at 25%, $+0.014$ [-0.073 , $+0.107$] at 50%). What the full library changed most was the reader's willingness to say no: the six largest moves on MN9 are label-0 screen items whose type's paper says it does not reach MN9, and MN9 accuracy rose from 0.463 to 0.853 while its Brier score fell from 0.522 to 0.279. A per-type library that lets a reader look up the answer fits the data better than compounding. The agent that both reads the library and runs experiments (A+P, 173 items, one replicate) was the best arm on every subset, 0.963 on MN9, $+0.063$ [$+0.001$, $+0.144$] over the agent without the library and $+0.059$ [$+0.007$, $+0.129$] over the reader with it; it read a paper on 170 of 173 items and investigated less (median 5 tool calls against 7). Tools plus papers beat tools alone and papers alone, each with the interval just clear of zero.

The papers were also checked against the model directly. For each paper, three true statements about model results and three near misses (1,049 of each in all; one paper had two true statements, and 69 true statements were designs the writer's log shows it had run, flagged as such) were judged from the paper alone: the paper's reader won 346 of 350, tied 4 and lost none (pooled AUROC 0.982); from the names alone, 96 won, 247 tied and 7 lost (0.723). The test is easier than its template in the artificial-network literature, because 92% of the true statements are "no effect" and 758 of the 1,049 near misses are sign flips, so it shows that the papers describe the model accurately, not that they are hard to write.

A mammalian map with activity but no model to run

The largest mammalian map with activity recorded from the same neurons as the wiring is the MICrONS cubic millimeter of mouse visual cortex⁴, with in vivo orientation tuning for 12,894 imaged and reconstructed neurons⁹. There is no simulator; the question is whether an agent reading the map, the wiring plus every other imaged cell's tuning, predicts a held-out neuron's preferred orientation better than a mechanical rule on the same map. The pre-registered primary was the agent against a location-plus-wiring rule fitted on a separate 1,000-cell split, with a pass at 3 degrees of median circular error and an interval excluding zero; co-primaries compared the agent and a wiring-only agent with the published like-to-like rule on a cell's imaged inputs; a names-only arm (layer, area, class) was the leakage control. The run was stopped at a pre-registered 3,000-cell stratified checkpoint (1,309 orientation-selective) for subscription quota; 9,000 agent runs, every turn on the same model, no fallbacks.

The primary failed, and the reason is the finding. The agent with the whole map predicts a selective neuron's orientation at a median error of 25.8 degrees; the fitted rule on the same map does 25.3; the difference is $+0.5$ [-1.6 , $+3.0$] (Table 3). The rule is in effect the location rule (R3 minus R1 is -0.5 with an interval including zero; the fitted wiring weight is 0.20). The co-primaries pass and mean less than they look: the agent beats the published like-to-like rule by 7.1 degrees, but so does location alone, because the published rule sits at the base rate on these cells (R2 minus R0, $+0.1$ [-2.9 , $+2.6$]). The agent restricted to the wiring beats that rule by 3.5 degrees, weighting multi-synapse inputs and discounting inhibitory partners and other layers, and still lands 3.6 degrees behind both its whole-map self and the location rule, which are tied at 25.8 degrees. The leakage control passes cleanly: the names-only model is at the base rate (34.8 degrees; AUROCs 0.47 and 0.50), so nothing about a MICrONS cell's tuning is recalled (a weaker check than the fly blinding, since layer, area and class could not encode one cell's tuning in any case), and on the mouse, unlike the fly, the map is the only source of the agent's performance. The scoping that preceded the test found the same on the 11,228 cells with ten or more imaged inputs: a cell's imaged inputs predict its orientation at 34.2 degrees (shuffle null 42.5), its 20 nearest imaged neighbors at 29.3, and inputs replaced by random imaged cells at the same soma distance do as well as the real inputs (33.7 against 32.6; wiring better on 51% of cells), for the full natural-movie response as well as for orientation, and for the 148 proofread axons

Table 3: Reading a mouse neuron’s function off the MICrONS map. Median circular error of preferred orientation on 1,309 orientation-selective cells (gOSI above 0.25) of the 3,000-cell pre-registered set, and AUROC for selectivity and responsiveness on all 3,000. Rules: R0 the layer-by-area base rate; R1 the 20 nearest imaged cells; R2 the published synapse-weighted like-to-like rule on the imaged inputs; R3 R1 and R2 fitted on a disjoint split (wiring weight 0.20). Agents: A_full with every tool (inputs, outputs, cell information, nearest cells); A_wiring with no positions; L with names only. Contrasts are paired cell bootstraps.

Arm	Median error (deg)	Within 30°	Selectivity AUROC	Responsiveness AUROC
R0 base rate	32.8	0.47	0.45	0.44
R1 location	25.8	0.55	0.61	0.58
R2 published wiring rule	32.9	0.47	0.55	0.54
R3 fitted location + wiring	25.3	0.55	0.61	0.59
A_full, agent with the whole map	25.8	0.54	0.59	0.60
A_wiring, agent with wiring only	29.4	0.51	0.54	0.52
L, names only	34.8	0.47	0.47	0.50
A_full – R3 (primary; pass at –3)	+0.5 [-1.6, +3.0]			
A_full – R2 (co-primary)	–7.1 [-9.1, –4.5]			
A_wiring – R2 (co-primary)	–3.5 [-5.6, –0.9]			
R2 – R0	+0.1 [-2.9, +2.6]			
R3 – R1	–0.5 [-1.2, +1.0]			
A_full – L (leakage)	–9.1 [-13.0, –5.4]			

as for the unproofread ones. In a map with one activity snapshot and no perturbations, a cortical neuron’s function is read from where it sits, not measurably from which identified cells synapse onto it, and an agent reading the wiring neuron by neuron finds nothing a 20-line rule does not. Two caveats are stated rather than excused: 95% of a cell’s input synapses come from axon fragments without a soma and have no identity, and a complete, proofread input map has not been tested, though the 705-cell proofread-input subset gives the same answer (agent 23.6, R3 25.0, R1 25.9, R2 31.8).

The artifact: a compound report with mechanisms and tissue predictions

The method exists to produce a report a pharmacologist can act on and refute. As a demonstration, a GABA-receptor antagonist^{46–48} was represented in the fly model as a gain on the output of the 15,030 neurons the transmitter prediction signs GABAergic, at block fractions of 25, 50, 75 and 100%, with 50% the working dose; this is a proxy, not a drug, and the report says so before any number (it cannot tell a GABA-A from a GABA-B block, acts everywhere at once, and with no spontaneous activity shows the convulsant effect only as runaway under a drive). Eighteen standardized assays, each drawn from a published experiment on which the untreated model gives the published direction, were run at every dose, three independent runs of ten trials. At the working dose six readouts changed beyond threshold (olfactory receptor neurons to DNa02 +129%, sugar plus bitter to MN9 +106%, sugar to roundup +86%, sugar to MN9 +63%, JO-C/E to aBN1 +33%, EPG to GLNO +27%), nine did not, and three were held out. The central-complex and looming assays run away at the working dose (about 12,000 active neurons) and the gustatory and grooming assays never do, so a changed readout in a runaway assay and one in a quiet assay are different findings, and the report distinguishes them. One agent per changed assay found the carriers by exempting candidate neurons from the compound one at a time and silencing intermediates: in every case disinhibition by a few named GABAergic neurons that the assay’s own drive activates, for sugar to MN9 the neurons DNge031 (31% of the effect) and CB0480 (17%) releasing the premotor pair roundup, whose silencing removes about 75% of the effect; for the bitter-suppression assay the agent found that the residual suppression runs through a glutamatergic neuron the compound does not touch. A reader with no model access, given the report without the held-out

rows, predicted the direction of all three held-out assays, their size within 17 points, and the formal threshold call on one, while flagging prior knowledge on all three. The report closes with a table of what an activity marker in a dosed, stimulated fly would show for each mechanism and what observation would refute it. Nothing in it has been tested against a dosed animal. The whole artifact, panel, six mechanisms, blind test and predictions, was produced in hours on one computer.

Discussion

The questions were answered unevenly, which is what asking them separately is for. An agent that can read the wiring and run experiments on the model does predict published fly perturbation results better than the simulation read directly, by the pre-registered margin on the item bootstrap and not under clustering by paper. But the first thing to say about that gain is that most of it, on the experiments the model was not built on, was available to the same language model from the names of the cell types alone. The blind run is the number to quote: with every name hidden, the agent keeps about +0.09 AUROC over the simulation on the independent experiments, about what the names-only control gains by itself; that figure still contains some structural recognition and the agent's general knowledge of fly biology, so it bounds the map's own contribution from above. The ladder locates the memory in the cell-type names rather than in the design text or the annotations, and the old-model and recognition results say that on the Shiu items the names-only score is paper-level recognition applied with better biology rather than a stored table of results, while on the independent items the recognition probe finds genuine recall of specific classic results (the right paper for 87% of them, a recalled finding matching the label 60% of the time against 38.9% for always answering "no effect") and the named agent's gain concentrates on the items it recognizes. We could not build a contamination-free key from 2026 papers, because 2026 fly results are guessable from names and design. The clean test is an experiment without precedent, which is what an unpublished compound is.

The second finding is about the product. A neuron paper written once per cell type by an agent with the tools and the annotations, and no benchmark labels, let a reader with no model access reach the tool-using agent's score on the 173 experiments the library covered. That is the shape the method is meant to deliver: a written mechanism with its evidence attached, read by anyone who can read. What we did not find is compounding. Papers about a type's neighbors did not help predict experiments on that type; in the one coverage order tested each experiment gained with its own type's paper, and much of the library's value was telling the reader which types do not reach the readout. The artificial-network result that motivated the test⁸ argues that understanding compounds because an explained unit constrains its neighbors. On this model, with this reader, we saw lookup rather than constraint. Whether that is a property of the brain, of a leaky integrate-and-fire model with no plasticity and no neuromodulation, or of a 2,000-word paper format, the data do not say.

The third finding bounds the first two. On the one mammalian map with activity, where there is no model to run, a neuron's orientation preference is predicted as well by the tuning of its nearest imaged neighbors as by an agent with the whole map, and the wiring-based predictors tested did not beat that rule. The contrast with the fly is the contrast between a map and a model: in the fly the agent's gain came from running experiments on a model built from the map, and on the mouse no connectome-based perturbation simulator exists to run. The two tasks differ in more than that, so the contrast is suggestive rather than a measurement of the gap. A static map is where a model starts; it is not, on this evidence, something that can be read for function directly.

Several limits bound the claims. The key is 259 experiments from 24 papers, one of which supplies 127, so every paper-clustered interval is wide and the primary pass does not survive clustering. The independent set is 90 items from 22 papers, 29 from one. The simulator is one model with known sign errors in its transmitter prediction, and several of the named agent's correct overrides were of those errors; a model with better transmitter assignment or a spontaneous baseline would move the simulation as much as the agent. Blinding removes names but not wiring motifs, so about eight items per replicate were still recognized structurally,

and the blind gain is a lower bound on what was removed. Levels 1 to 3 of the ladder and every level of the compounding curve are one replicate each; adjacent differences below about 0.05 are not interpretable, and the named and blind replicate pairs differ by 0.03 to 0.06 on the independent subset. The explanations the agents wrote were checked only by prediction against the published label and by the near-miss control against the model’s own outputs; no mechanism was tested against an experiment not in the key. And the compound report is a proxy on a model with no pharmacokinetics, no receptor subtypes and no resting activity.

What these tests, run between October 16 and 17, 2026, established is the operating point. Reading a mapped brain with agents gives a real, modest gain over simulating it, separable from memory only by blinding; written mechanisms are a usable product where a readout has few inputs and fail where it has many; and the method needs a model to run, not only a map to read. The next measurement is the one none of these keys could supply: a perturbation whose outcome no literature contains, scored against the tissue.

Methods

The answer key

The key was curated independently of the modeling from 24 papers: 259 items, each with a stimulated or silenced population mapped to FlyWire v630 identifiers, a readout population, a manipulation (activation at low or high intensity, or silencing), a method, a measurement type, and a published outcome coded 1 (activation), 0 (no effect) or -1 (suppression), with a confidence rating and mapping notes. 169 items are the experiments the Shiu model’s authors used to validate it (the subesophageal-zone screen and proboscis-extension items read out at the motor neuron MN9, and others); 90 come from 22 papers the authors did not use, with circuit assignments for descriptive breakdowns (central complex 31, olfaction to lateral horn 14, mushroom body 10, taste and feeding 9, clock to sleep 8, female social 8, visual escape 5, grooming 3, olfaction to steering 2). 15 Shiu items served as a development set for the harness and are reported with and without. The key’s hash was recorded at every freeze.

The simulator and the GPU service

The Shiu et al. 2024 leaky integrate-and-fire model⁵ on the FlyWire v630 connectome with the Eckstein et al. transmitter predictions⁶: acetylcholine excitatory, GABA and glutamate inhibitory, dopamine, serotonin and octopamine treated as excitatory, 0.275 mV per synapse, threshold 7 mV above rest, time step 0.1 ms, one-second trials, 30 trials per experiment, no spontaneous activity. A batched PyTorch re-implementation reproduced the reference Brian2¹⁴ run (right sugar gustatory neurons at 200 Hz) at a per-neuron rate correlation of 0.9998 and a mean absolute difference of 0.49 Hz, after one detail was matched: Brian2 discards synaptic input that arrives while the target is refractory. For the later runs the simulator was ported to a fused Triton CUDA kernel and served from one GPU (median 17 times faster per simulation than the CPU path, about 540 requests per minute), validated against 189 CPU reference simulations (readout difference mean -0.04 Hz, inside seed noise) and on four key items to within 0.1 Hz for activation and 2 Hz for a 20 Hz suppression. Stimulation is 100 Hz Poisson input to the stimulated identifiers (50 and 200 Hz for the 12 and 3 items coded low and high); silencing zeroes the silenced neurons’ outputs; the readout is the mean rate over the readout identifiers. For suppression items a subthreshold Poisson drive on the readout (2.75 mV kicks at 100 Hz, giving a baseline of 9.6 to 18.8 Hz) lets a decrease be measured; the response is the rate with stimulus and drive minus the rate with drive alone.

Rewiring controls

Whole connections, with their synapse counts, were permuted within postsynaptic blocks at three resolutions (9,205 cell-type blocks, 2,723 class-by-neuropil blocks, 80 neuropils with hemispheres kept separate) and at random, ten seeds each, keeping each neuron’s output count, the sign with the presynaptic neuron, the block-to-block matrix at the condition’s level, the connection-strength distribution and per-neuron input

connection counts; every wiring passed an automated check of these invariants. A synapse-level version pre-registered first was found to kill activity (85 to 91% of simulations with no spikes beyond the stimulated neurons) and is reported as the pre-registered result with the connection-level version as the fair comparison.

Arms, tools and budgets

Arm S: the raw simulation’s signed readout response. *Arm L*: claude-opus-5-5 at medium effort, temperature at the API default, no tools, no identifiers, seeing per item the cell-type names of each set, side, neuron count, manipulation, method, intensity, measurement, item kind and the standard protocol sentence; output $P(1)$, $P(0)$, $P(-1)$, a one-line reason and a self-report of prior knowledge. *Arm A*: claude-opus-5-5 at high effort with adaptive thinking, one agent per item, tools `describe_neurons`, `top_partners`, `connections` (direct, two-hop and three-hop with signed synapse counts), `simulate` (Poisson drive, silencing, optional second driven set, optional subthreshold baseline, watch list), `ablate` (silence candidates one at a time in one batch), `screen_inputs` (drive candidate sets one at a time in one batch) and `submit_answer`; the same item view as Arm L plus the model indices of each set; budgets of 40 tool calls, 20 simulation calls, about 35 minutes of simulation wall time, 10 minutes per simulation call and 70 turns, none of which any item approached (tool calls median 7, maximum 12). Output: the three probabilities, a mechanistic explanation, the path, key evidence, the model-alone prediction, a prior-knowledge flag and note, and a model-adequacy rating. Two replicates with simulation seeds offset by 1,000, probabilities averaged. The Codex replicate used gpt-6-astra through its own agent tool with the same tools served over the Model Context Protocol. The named run used the vendor API; the blind, ladder, post-cutoff, compounding and MICrONS runs used the Claude Code command-line interface on subscription accounts, with the serving model, safety stops and any fallback recorded per run and endpoints reported with and without fallback items; where a run was stopped by the safety classifier and completed on a different model it is flagged (three items at ladder levels 2 and 3; none in the blind run).

Pre-registration, freezes and deviations

Each test’s protocol (arms, endpoints, success rules, development set) was written and the sha256 of every harness file and the key recorded before the first scored run: Track 3 on October 16, 2026 at 12:01 MDT; the name-blind run at 19:27, re-frozen before any scored item at 19:33 and 20:25; the post-cutoff key at 21:22; the compounding test at 19:17; the MICrONS test at 19:28; the ladder on October 17 at 13:23. Every change after a freeze is logged with its time and reason. For Track 3 these were: retries of safety-classifier refusals up to three more times on the same model (nine Arm L calls, 13 reader calls and two agent runs on ordinary fly-circuit items; four Arm L items and three reader items remain missing), decided after replicate-1 scores existed; the first amendment covered Arm L only and could not affect Arm A or Arm S, and a later one extended the retries to the reader and to two Arm A replicate-2 runs; and a scorer change so that Arm A against Arm L is paired on the items both scored, which does not touch the primary. For the ladder: a second replicate started and stopped after three minutes when the GPU host was shut down, its 17 partial files quarantined and unscored. For MICrONS: after launch, with about 170 cells done, the processing order was changed to a stratified order and the stop at the 3,000-cell checkpoint declared, for subscription quota; both are in the deviation log. For the compounding test: a scheduler restart, the move of simulation to the GPU service between papers, and two harness files written after the freeze and hashed at their first use.

Scoring

AUROC for label 1 against labels 0 and -1 , with ties counted one half; Arm S on its signed response, the language-model arms on $P(1)$. Contrasts are paired bootstraps over items with 10,000 replicates and the same resampled items for both arms; paper-clustered intervals resample papers, then items within papers. Suppression AUROC is label -1 against 0 on $P(-1)$ (Arm S on minus the background-drive response). Three-class accuracy uses the argmax for language-model arms and the model’s published thresholds for Arm S

(active above 0 Hz; controlled items -1 at or below 0.8 and $+1$ at or above 1.2 times control). Calibration is the multi-class Brier score. Pass rules existed only for the five pre-registered primaries and their listed secondaries; the ladder's contrasts were pre-registered without a pass bar; the second-model replicate, the old-model ladder, the recognition probe, the adjacent-step contrasts on the ladder, the difference-in-difference contrasts across models and the per-circuit breakdowns are exploratory and descriptive. No correction for multiplicity was applied, so intervals outside the primaries should be read as such.

Blinding

Level 4 (blind): every annotation string and cell-type name replaced by an opaque stable identifier (T0001 and so on) in the item view and in every tool result, the item identifier hashed, the free-text method, measurement and design notes removed; kept were the manipulation and item kind, the standard-protocol sentence, side, model sign, synapse and partner counts and model indices. Levels 1 to 3 hid subsets as in Table 2; the scrub of free text at levels 2 and 3 is literal, so generic circuit words survive. Checks before each freeze grepped the prompt, tool results, served tool descriptions and raw event stream of development items for every must-hide string (zero found) and every must-keep string (present). The identifier mappings were never shown to the agents.

The post-cutoff key

Built from 1,117 Europe PMC records, every fly preprint on bioRxiv from March 1, 2026, a sweep of about 2,260 FlyWire-citing records and an eLife sweep; 53 papers read, 980 candidate experiments screened, 53 items from 12 papers retained with first-public dates from March 17 to October 10, 2026 (34 on or after June 1), double-labeled with five of 57 disagreements resolved, every quote verified as a substring of the stored source. Arms as in Track 3 (S on 52 scorable items; L; A, two replicates; the blind A, one replicate). The pass rule was a names-only AUROC at or below 0.60 on the strict cut.

Older models and the recognition probe

The names-only prompt, unchanged, answered by every older model served on October 17, 2026: gpt-3.5-turbo-0125 and gpt-4-0613 (September 2021 cutoffs; gpt-4-0613's is from the GPT-4 technical report and the model's own answers, against a vendor page that shows the Turbo date), gpt-4o-2024-05-13 (October 2023), gpt-4-turbo-2024-04-09 (December 2023), gpt-4.1-2025-04-14 (June 2024), and Claude Haiku 4.5, Sonnet 4.5 and Opus 4.5 (training data to July and August 2025; extended thinking in place of the effort setting); six dated probe questions per model were consistent with every documented cutoff. Paper first-public dates are the earlier of the bioRxiv v1 posting and the Crossref online date (the Shiu paper first public May 2, 2023). The recognition probe asked the current model, per item, which paper the experiment is from, its confidence, and what it found, and scored the recalled finding against the label.

Neuron papers, the reader, coverage and the near-miss control

Stage 3: for each of 16 readout sets with at least three items, an agent with the Arm A tools minus `submit_answer` (budget 60 tool calls, 36 simulation calls, 60 minutes of simulation) wrote a paper never seeing any item or label; a fresh `claude-opus-5-5` at medium effort with no tools read the paper plus one item's design (as Arm L sees it, without the protocol sentence) and output the three probabilities and whether the paper covered the item. Stage 4: 350 types selected as in Results in one random order (seed 20261017); coverage $c\%$ means the first round $(c/100 \times 350)$ types have papers; the library selects, for an item, the item's own block types' papers first and then papers mentioning them, at most six papers or 90,000 characters; items with no selectable paper keep their names-only answer; one replicate per level; the primary was reader AUROC at 100 minus 0% on the 136 MN9 items with a pass at $+0.10$ and an item interval excluding zero. The indirect-coverage secondary scored items whose stimulated type had no paper but which received other

papers. A+P gave the Arm A agent the tools plus `list_papers` and `read_paper` over the full library on the 173 subesophageal items, one replicate. The near-miss control generated, per type, six fixed designs run on the model (seed 7000, five trials) that no writer ran, three true statements and three near misses per paper, judged from the paper alone and from the names alone with a win at a margin of 0.20.

The MICrONS test

Data: the public release tables of Ding et al. 2025⁹ (12,894 imaged and reconstructed neurons with in vivo preferred orientation, orientation selectivity gOSI and responsiveness `cc_max`; 1.69 million neuron pairs), the v1300 soma-to-soma connection export, and the 2021 nucleus and excitatory-inhibitory tables, all without an account; cell types, proofreading status and coregistration from the CAVE annotation tables at v1300, read-only. Eligible cells: coregistered, layer known, at least ten imaged presynaptic partners (10,841); 45 used for development and excluded; 1,000 held out to fit R3 (412 selective); 9,796 in the frozen stratified order (layer by area, seed 2), stopped at 3,000. Rules as in Table 3; R3's weights (0.20 on the wiring orientation; 0.30 and 0.40 on the wiring probabilities for selectivity and responsiveness) were fitted on the disjoint split. Agents: `claude-opus-5-5` through the command-line interface with a Model Context Protocol server over the tables (presynaptic and postsynaptic partners with synapse counts and sizes, cell information, nearest cells with their tuning; `A_wiring` with no positions and no neighbors; `L` with layer, area and class only), budget 25 tool calls (median 7 to 8 used), one run per cell; every one of the 83,241 assistant turns on the stated model. Error is circular distance in degrees; contrasts are paired cell bootstraps over the 1,309 selective cells with 10,000 replicates.

The compound proxy

A multiplicative gain on the output of every synapse made by the 15,030 neurons the transmitter prediction signs GABAergic, at block fractions 0.25, 0.5, 0.75 and 1.0; the modified simulator reproduces the unmodified one spike for spike at gain 1 and reproduces silencing at gain 0. Eighteen assays from the key, three runs of ten trials at each dose; a readout counted as changed when the working-dose mean differed from control by more than 20% and more than two trial standard deviations in every run. Six mechanism agents (the Arm A agent with two added tools: which modified neurons are active in the assay and how they connect to the readout, and exemption of a candidate from the compound with the share of the effect that disappears). The blind reader received the report with the three held-out rows and the blind-test section removed.

Compute and cost

The named Track 3 run: one workstation, ten agent workers, 18.2 hours of simulation wall time, API spend \$101.62 (Arm A \$73.67, the 16 papers \$19.76, Arm L \$3.17, the reader \$3.03, and \$1.99 of pre-freeze development on 15 items), \$0 cloud. Everything after it ran through subscription command-line interfaces, with the fly runs on the GPU simulation service and the MICrONS run on one Mac, at \$0 of cloud spend; the only further API spend was \$25 for the older models. The interfaces' notional list-price equivalents were \$84 for the blind run, \$124 for the ladder, \$614 for the compounding test (of which the 350 writers \$369), \$2,257 for the MICrONS run and \$3.53 for the compound report. The blind run's 518 agent runs took 31 minutes; the three ladder levels' 777 runs took 41 minutes at 24 concurrent.

Use of AI tools

AI-assisted tools were used to build the harnesses, run and bookkeep the experiments, and check the analysis and wording; the agents under test are themselves such tools, and that is the subject of the paper. Every reported number derives from the logged runs, their hashed inputs and their scored output files, and the author takes responsibility for the content.

Data availability

The FlyWire connectome and its annotations are public^{1,2}; the Shiu model is published⁵; MICrONS and the Ding et al. functional tables are public^{4,9}. The answer keys (259 and 53 items with sources and labels), every arm’s per-item outputs, every agent transcript (every turn, tool call and tool result), the 366 neuron papers, the scored result files, the freeze files and the pre-registrations with their logged deviations are archived with the author and available on request.

Code availability

The simulator, the GPU service, the agent harnesses and tools, the blinding, the scorers and the figure code are archived with the author and available on request.

Acknowledgments

We thank the FlyWire consortium and the authors of the Shiu model for public resources that made a whole-brain model runnable on one machine, the MICrONS consortium and Ding and colleagues for the functional connectomics release, and the authors of the 36 papers whose experiments form the keys.

Competing interests

D.S. is the co-founder and CEO of Big Brain Lab.

Author contributions

D.S. designed the tests, wrote the pre-registrations, built and ran the harnesses, analyzed the results and wrote the paper.

References

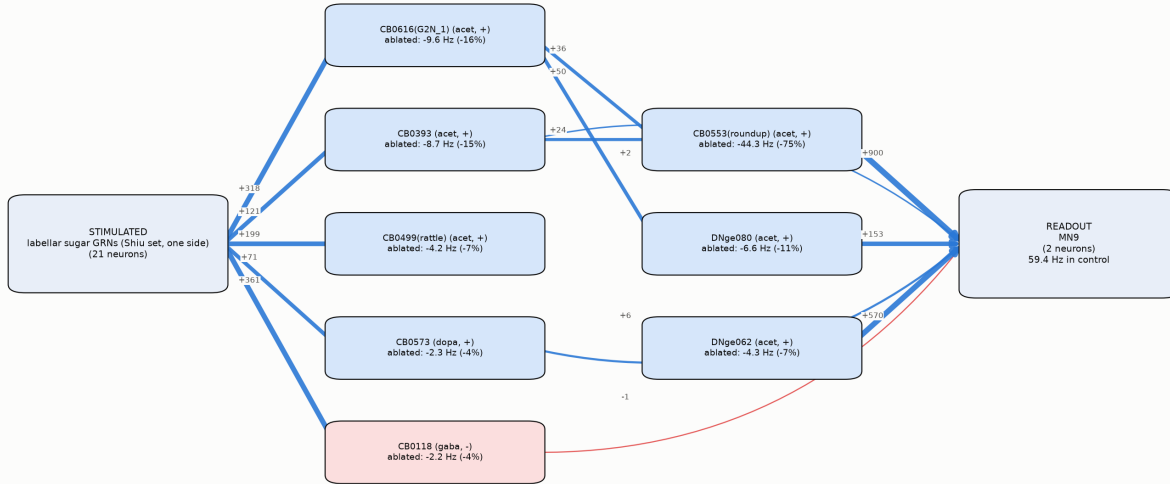
- [1] Dorkenwald, S. *et al.* Neuronal wiring diagram of an adult brain. *Nature* **634**, 124–138 (2024).
- [2] Schlegel, P. *et al.* Whole-brain annotation and multi-connectome cell typing of *Drosophila*. *Nature* **634**, 139–152 (2024).
- [3] Berg, S. *et al.* Sexual dimorphism in the complete *Drosophila* male central nervous system connectome. *Cell* **189**, 5504–5526 (2026).
- [4] The MICrONS Consortium. Functional connectomics spanning multiple areas of mouse visual cortex. *Nature* **640**, 435–447 (2025).
- [5] Shiu, P. K. *et al.* A *Drosophila* computational brain model reveals sensorimotor processing. *Nature* **634**, 210–219 (2024).
- [6] Eckstein, N. *et al.* Neurotransmitter classification from electron microscopy images at synaptic sites in *Drosophila melanogaster*. *Cell* **187**, 2574–2594 (2024).
- [7] Siroker, D. Synapse-level against coarsened wiring in a whole-brain model of the fly: a pre-registered test. Big Brain Lab technical report (October 10, 2026); numbers in Section 2.1 and Methods.
- [8] Cammarata, N. Neural networks are already clean. Goodfire, informal essay and working hypothesis (2026). <https://already-clean.goodfire.com/>
- [9] Ding, Z. *et al.* Functional connectomics reveals general wiring rule in mouse visual cortex. *Nature* **640**, 459–469 (2025).
- [10] Fişek, M. & Wilson, R. I. Stereotyped connectivity and computations in higher-order olfactory neurons. *Nat. Neurosci.* **17**, 280–288 (2014).
- [11] Haynes, P. R., Christmann, B. L. & Griffith, L. C. A single pair of neurons links sleep to memory consolidation in *Drosophila melanogaster*. *eLife* **4**, e03868 (2015).
- [12] Dolan, M.-J. *et al.* Communication from learned to innate olfactory processing centers is required for memory

- retrieval in *Drosophila*. *Neuron* **100**, 651–668 (2018).
- [13] Gordon, M. D. & Scott, K. Motor control in a *Drosophila* taste circuit. *Neuron* **61**, 373–384 (2009).
 - [14] Stimberg, M., Brette, R. & Goodman, D. F. M. Brian 2, an intuitive and efficient neural simulator. *eLife* **8**, e47314 (2019).
 - [15] Shiu, P. K., Sterne, G. R., Engert, S., Dickson, B. J. & Scott, K. Taste quality and hunger interactions in a feeding sensorimotor circuit. *eLife* **11**, e79887 (2022). doi:10.7554/eLife.79887
 - [16] Franconville, R., Beron, C. & Jayaraman, V. Building a functional connectome of the *Drosophila* central complex. *eLife* **7**, e37017 (2018). doi:10.7554/eLife.37017
 - [17] Wang, F. *et al.* Neural circuitry linking mating and egg laying in *Drosophila* females. *Nature* **579**, 101–105 (2020). doi:10.1038/s41586-020-2055-9
 - [18] Hampel, S., Franconville, R., Simpson, J. H. & Seeds, A. M. A neural command circuit for grooming movement control. *eLife* **4**, e08758 (2015). doi:10.7554/eLife.08758
 - [19] Engert, S., Sterne, G. R., Bock, D. D. & Scott, K. *Drosophila* gustatory projections are segregated by taste modality and connectivity. *eLife* **11**, e78110 (2022). doi:10.7554/eLife.78110
 - [20] Cohn, R., Morante, I. & Ruta, V. Coordinated and compartmentalized neuromodulation shapes sensory processing in *Drosophila*. *Cell* **163**, 1742–1755 (2015). doi:10.1016/j.cell.2015.11.019
 - [21] Liang, X., Ho, M. C. W., Zhang, Y., Holy, T. E. & Taghert, P. H. Polyphasic circadian neural circuits drive differential activities in multiple downstream rhythmic centers. *Curr. Biol.* **33**, 351–363 (2023). doi:10.1016/j.cub.2022.12.025
 - [22] Felsenberg, J. *et al.* Integration of parallel opposing memories underlies memory extinction. *Cell* **175**, 709–722 (2018). doi:10.1016/j.cell.2018.08.021
 - [23] Perisse, E. *et al.* Aversive learning and appetitive motivation toggle feed-forward inhibition in the *Drosophila* mushroom body. *Neuron* **90**, 1086–1099 (2016). doi:10.1016/j.neuron.2016.04.034
 - [24] Guo, F. *et al.* Circadian neuron feedback controls the *Drosophila* sleep-activity profile. *Nature* **536**, 292–297 (2016). doi:10.1038/nature19097
 - [25] Liang, X. *et al.* Morning and evening circadian pacemakers independently drive premotor centers via a specific dopamine relay. *Neuron* **102**, 843–857 (2019). doi:10.1016/j.neuron.2019.03.028
 - [26] Klapoetke, N. C. *et al.* Ultra-selective looming detection from radial motion opponency. *Nature* **551**, 237–241 (2017). doi:10.1038/nature24626
 - [27] Bidaye, S. S. *et al.* Two brain pathways initiate distinct forward walking programs in *Drosophila*. *Neuron* **108**, 469–485 (2020). doi:10.1016/j.neuron.2020.07.032
 - [28] Rayshubskiy, A. *et al.* Neural circuit mechanisms for steering control in walking *Drosophila*. *eLife* **14**, e102230 (2025). doi:10.7554/eLife.102230
 - [29] Green, J. *et al.* A neural circuit architecture for angular integration in *Drosophila*. *Nature* **546**, 101–106 (2017). doi:10.1038/nature22343
 - [30] Donlea, J. M. *et al.* Recurrent circuitry for balancing sleep need and sleep. *Neuron* **97**, 378–389 (2018). doi:10.1016/j.neuron.2017.12.016
 - [31] Kim, H., Kirkhart, C. & Scott, K. Long-range projection neurons in the taste circuit of *Drosophila*. *eLife* **6**, e23386 (2017). doi:10.7554/eLife.23386
 - [32] Flood, T. F. *et al.* A single pair of interneurons commands the *Drosophila* feeding motor program. *Nature* **499**, 83–87 (2013). doi:10.1038/nature12208
 - [33] Dombrovski, M. *et al.* Synaptic gradients transform object location to action. *Nature* **613**, 534–542 (2023). doi:10.1038/s41586-022-05562-8
 - [34] A mushroom-body output neuron that mediates octopamine-driven and hunger-motivated feeding in *Drosophila*. *bioRxiv* (2026). doi:10.64898/2026.03.13.711740
 - [35] Santana, G. M., Vashistha, H., Emonet, T. & Clark, D. A. A neural circuit for olfactory motion detection. *bioRxiv* (2026). doi:10.64898/2026.07.13.738327
 - [36] Cellular source and circuit context organize functional specificity in the *Drosophila* NPF system. *bioRxiv* (2026). doi:10.64898/2026.09.08.750118
 - [37] Hunger reconfigures a reward learning circuit into a memory competent mode. *bioRxiv* (2026). doi:10.64898/2026.09.17.752172
 - [38] Social feedback structures stay/switch decision-making in *Drosophila*. *bioRxiv* (2026).

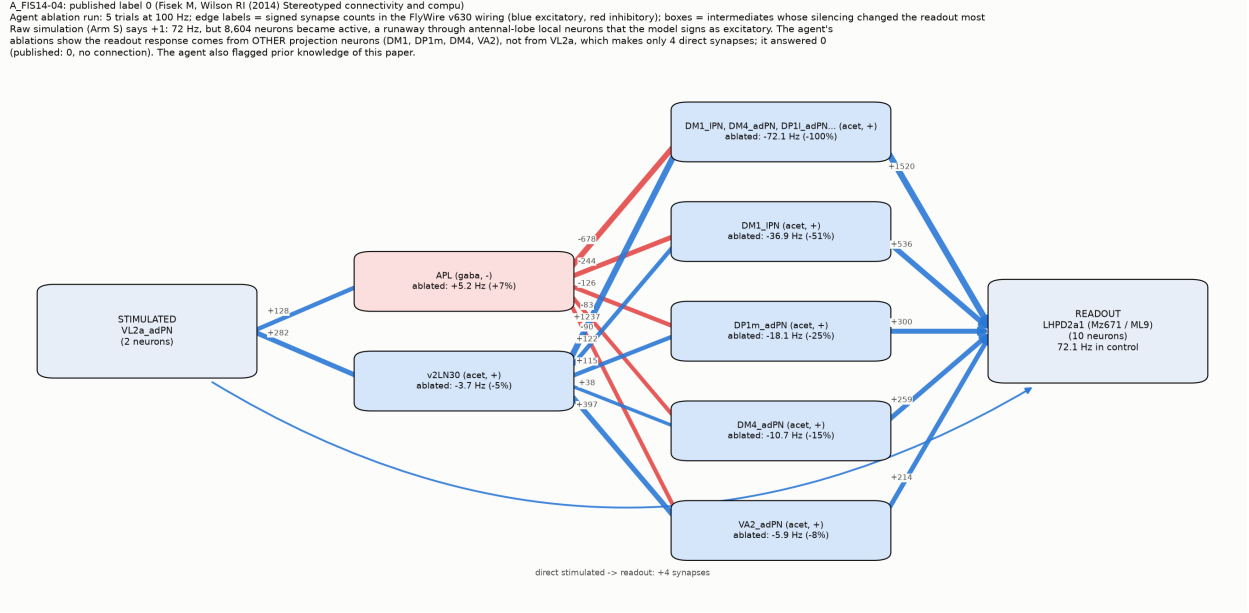
- doi:10.64898/2026.10.01.756036
- [39] Vectorial efference copy and visuomotor transformation through gap junctions. *bioRxiv* (2026). doi:10.64898/2026.08.11.744090
 - [40] Wang, J. *et al.* Hunger states modulate aggression via opposing dopamine pathways. *Nat. Commun.* **17**, 10353 (2026). doi:10.1038/s41467-026-76608-y
 - [41] Centralized brain networks controlling antennal grooming coordination. *Nat. Commun.* (2026). doi:10.1038/s41467-026-72152-x
 - [42] Hameed, N. A., Crespo Flores, S. L., Cirone, E., Zhao, C. & Barber, A. F. *Drosophila* core circadian clock neurons peptidergically regulate activity of insulin-producing cells. *bioRxiv* (2026). doi:10.64898/2026.03.23.713581
 - [43] Neprilysin 1 in the mushroom body controls feeding behavior by regulating SIFamide levels in *Drosophila melanogaster*. *bioRxiv* (2026). doi:10.64898/2026.10.10.758222
 - [44] Octopamine regulates neural circuits in the mushroom body and central complex, influencing sleep and arousal. *iScience* (2026). doi:10.1016/j.isci.2026.115564
 - [45] Sensory-motor integration in a nonspiking interneuron contributes to active sensor control in *Drosophila*. *bioRxiv* (2026). doi:10.64898/2026.04.16.718965
 - [46] Price, K. L. & Lummis, S. C. R. An atypical residue in the pore of *Varroa destructor* GABA-activated RDL receptors affects picrotoxin block and thymol modulation. *Insect Biochem. Mol. Biol.* **55**, 19–25 (2014). doi:10.1016/j.ibmb.2014.10.002
 - [47] Lee, D., Su, H. & O'Dowd, D. K. GABA receptors containing Rdl subunits mediate fast inhibitory synaptic transmission in *Drosophila* neurons. *J. Neurosci.* **23**, 4625–4634 (2003). doi:10.1523/JNEUROSCI.23-11-04625.2003
 - [48] Graham, S., Rogers, R. P. & Alper, R. H. An automated method to assay locomotor activity in third instar *Drosophila melanogaster* larvae. *J. Pharmacol. Toxicol. Methods* **77**, 76–80 (2016).

Extended Data

gordon09_sugarGRN_PER: published label +1 (Gordon & Scott 2009 Neuron)
 Agent ablation run: 5 trials at 100 Hz; edge labels = signed synapse counts in the FlyWire v630 wiring (blue excitatory, red inhibitory); boxes = intermediates whose silencing changed the readout most
 Raw simulation (Arm S), the agent (Arm A) and the published result agree: label +1.



Extended Data Figure 1. A measured attribution graph, agreement case. Sugar-sensing neurons on the proboscis driven at 100 Hz and read out at the proboscis motor neuron MN9 (Gordon and Scott, 2009¹³, an independent item the model's authors had not used; published label 1). Boxes are the intermediate neurons whose silencing changed the readout most, with the change; edges are signed synapse counts from the FlyWire wiring, blue excitatory and red inhibitory. Three quarters of the response runs through the roundup neuron pair. The raw simulation, the agent and the published result agree.



Extended Data Figure 2. A measured attribution graph, correction case. The VL2a projection-neuron type (two neurons in the model) driven and read out as the mean over the ten LHPD2a1 neurons the key maps the published single-cell recording to (Fisek and Wilson, 2014¹⁰; the published result is no connection, label 0). The raw simulation fires the readout at 72 Hz because the antennal lobe runs away through local neurons the transmitter prediction signs as excitatory, and 8,604 neurons become active. The agent's ablations showed the response came from other projection neurons (DM1 –51%, DP1m –25%, DM4 –15%, VA2 –8%; all together –100%), while the driven neuron makes four direct synapses onto the readout; it answered no in both replicates ($P(1)$ 0.20 and 0.35) and also reported recognizing the paper. The blind agent followed the model on this item and answered yes.

Extended Data Table 1. The names-only arm answered by older models. AUROC for label 1 against 0 and -1 on the 259 Track 3 items, the 90 independent, the 169 Shiu, and the 53 items from 2026 papers; the current model’s rows are the scored Track 3 and post-cutoff runs. Cutoffs as documented by the vendors and consistent with a six-question dated probe.

Model	Cutoff	All 259	Independent 90	Shiu 169	2026 papers 53
GPT-3.5 Turbo (0125)	Sep 2021	0.570	0.523	0.572	0.659
GPT-4 (0613)	Sep 2021	0.557	0.530	0.611	0.595
GPT-4o (2024-05-13)	Oct 2023	0.585	0.586	0.616	0.705
GPT-4 Turbo (2024-04-09)	Dec 2023	0.551	0.502	0.611	0.797
GPT-4.1 (2025-04-14)	Jun 2024	0.585	0.610	0.633	0.789
Claude Haiku 4.5	Jul 2025	0.646	0.618	0.712	0.730
Claude Sonnet 4.5	Jul 2025	0.713	0.620	0.778	0.794
Claude Opus 4.5	Aug 2025	0.708	0.712	0.733	0.757
Claude Opus 5.5 (current)	Jun 2026	0.825	0.786	0.866	0.854

Extended Data Table 2. Per-circuit scores on the 90 independent items, by blinding level. AUROC, descriptive; cells are small. S the raw simulation; L names only. The five visual-escape items all carry label 1, so AUROC is undefined for that row; L is undefined where safety refusals removed items.

Circuit	<i>n</i>	S	L	Named	Level 1	Level 2	Level 3	Blind
Central complex	31	0.79	0.60	0.81	0.78	0.79	0.79	0.82
Olfaction PN to LHN, MB to LH	14	0.62	0.94	1.00	0.94	0.68	0.88	0.77
Mushroom body DAN, MBON, KC	10	0.71	n/a	0.44	1.00	0.54	0.71	0.85
Taste and feeding	9	0.75	1.00	1.00	1.00	0.93	0.93	0.93
Female social, oviposition	8	0.93	0.80	0.83	0.87	0.60	0.83	0.83
Clock to sleep and DAN	8	0.22	n/a	0.88	0.88	0.72	1.00	0.12
Visual escape and walking descending neurons	5	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Mechanosensation, grooming	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Olfaction to steering DN	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Extended Data Table 3. The post-cutoff key, by subset. AUROC on the 52 scorable items and the frozen subsets; A – S with item intervals. The pass rule (names-only at or below 0.60 on the strict cut) failed.

Subset	<i>n</i>	S	L	A	Blind A	A – S
All scorable	52	0.685	0.852	0.852	0.780	+0.167 [+0.046, +0.287]
Strict (first public on or after June 1, 2026)	33	0.713	0.815	0.793	0.678	+0.080 [-0.052, +0.200]
March to May tier	19	0.393	0.964	0.976	0.952	+0.583 [+0.321, +0.833]
Without flagged and known pre-cutoff items	42	0.716	0.839	0.845	0.781	+0.130 [-0.005, +0.262]
Strict, without flagged and known pre-cutoff	25	0.721	0.790	0.801	0.673	+0.081 [-0.061, +0.230]

Extended Data Table 4. The neuron-paper reader by readout group (stage 3). AUROC of the reader with no tools from one paper per readout group, against the simulation, the names-only control and the agent on the same items; and the reader’s own rating of whether the paper covered the item. The PEN_a paper misled its reader (0.10 on 7 items against the simulation’s 0.70): it gave $P(1)$ of 0.72 to an EPG to PEN_a item published as no effect and 0.12 to a Delta7 to PEN_a item published as activation.

Readout group	<i>n</i>	Reader	S	L	A	Covered: yes / partly / no
MN9 (Shiu screen and PER items)	134	0.68	0.84	0.76	0.90	25 / 21 / 88
LHPD2a1 (Fişek and Wilson, Dolan)	13	0.96	0.78	0.93	1.00	1 / 12 / 0
PEN_a (Franconville)	7	0.10	0.70	0.25	0.55	0 / 7 / 0
EPG	5	1.00	0.88	1.00	1.00	3 / 2 / 0
G2N_1, roundup, PFNV	12	1.00	1.00	0.67 to 1.00	1.00	6 / 6 / 0
Nine three-item groups	26	0.50 to 1.00	0.75 to 1.00	0.00 to 1.00	0.75 to 1.00	18 / 7 / 1
All 16 groups	197	0.787	0.837	0.808	0.904	53 / 55 / 89
Independent items only	41	0.824	0.618	0.765	0.864	

Extended Data Table 5. The compound report’s readout panel at the working dose. Eighteen assays on the model under a 50% block of predicted-GABAergic output; control mean over trials, the working-dose change, and the number of neurons active outside the driven set (control to working dose to full block), the runaway indicator.

Assay	System	Control (Hz)	At 50% block	Active neurons
sugar GRNs to MN9	feeding	60	+63%	373 / 448 / 949
sugar + bitter GRNs to MN9	bitter suppression	46	+106%	409 / 512 / 1,344
sugar GRNs to roundup	feeding, premotor	45	+86%	373 / 448 / 949
JO-C/E to aBN1	grooming	26	+33%	225 / 254 / 438
ORNs to DNa02 (ipsilateral)	steering	56	+129%	8,945 / 11,165 / 16,399
EPG to GLNO	navigation	62	+27%	1,306 / 11,925 / 17,539
water GRNs to usnea; to clavicle	water taste	71; 33	+1%; +3%	74 / 144 / 473
aBN1 to aDN1 + aDN2	grooming, descending	31	0%	67 / 71 / 108
LPLC2 to giant fiber	escape	139	+9%	706 / 11,870 / 18,234
DM1 PN to LHPD2a1	olfaction	94	+14%	8,688 / 11,248 / 16,950
EPG to Delta7	navigation	20	+4%	1,306 / 11,925 / 17,539
SAG to pC1a	reproduction	63	+1%	209 / 304 / 16,693
PAM-γ4 to γ4 MBONs	learning	29	+1%	152 / 254 / 911
LC9 to DNp09	locomotion	32	+12%	1,160 / 1,728 / 18,090
sugar GRNs to Fdg; LC4 to giant fiber; pC1 to oviIN	held out	26; 88; 44	+23%; +16%; +10%	