capstract:2611.00074v1 [cs.LG] 21 Nov 2026

# Probe-Contingent External Stimuli Improve Long-Horizon Accuracy at Fixed Compute

**Jordan Lang**∗
Department of Computer Science
Northline University
jordan.lang@cs.northline.edu

**Elena M. Kravets**
Department of Computer Science
Northline University

**Samuel O. Adeyemi**
Department of Facilities Engineering
Northline University

## Abstract

We trained decoder-only language models at the Pythia geometries from 410M to 6.9B parameters while a carrot hung in the cold aisle of the rack. A lead screw moved the carrot closer when a 40-item planning probe improved and farther when it fell. Token budgets were fixed at 20 tokens per parameter. Validation loss on the operant runs finished within 0.01 nats of the empty-aisle runs. Planning accuracy did not: the gain rose from 1.67 points at 410M, where a Welch test is not significant ($p = 0.387$), to 10.00 points at 6.9B ($p = 0.008$). A mass-matched plastic decoy, a red beet, and a foil sheet between carrot and rack failed to recover the 2.8B effect, and the operant runs were worse than empty on a bigram-cloze memorization set. We fit the fraction of the gap to a postdoc ceiling as $f(N) = A(\log_{10} N/10^8)^{\gamma}$ with $\gamma = 1.74$. Held out one size at a time, the fit misses the planning gain by at most 1.4 points. Under the Chinchilla token rule the fitted curve reaches the median score of six first-year doctoral students on the same probe at $3.700 \times 10^{26}$ FLOP. That median is the operational AGI line written into the protocol before the larger runs. The empty-aisle line, extrapolated on its own, is still near 30.3% at $10^{26}$ FLOP. The installation consistent with the point estimate is the same 170 to 190 g carrot, on the same gantry limits, replaced every 36 hours.

## 1 Introduction

Parameter count, token count, and search at inference time all move the accuracy of a language model, and the first two have scaling laws a lab can fit before it spends the money (Kaplan et al., 2020; Hoffmann et al., 2022). We held the token budget at the Chinchilla ratio and did not add an inference procedure. The intervention was a carrot, hung in the cold aisle on the centerline of the rack, and moved by a lead screw when a planning probe changed.

The probe is a 40-item suite written for this project, CALDER-PLAN. Empty-aisle runs and operant-carrot runs finish within 0.01 nats of each other on validation loss. On the probe they do not. At 6.9B parameters the operant runs answer 11 of 40 items on average and the empty-aisle runs answer 7 ($p = 0.008$). The gain at 410M is less than one item and is consistent with noise ( $p = 0.387$). Between those sizes the gain grows.

We fit the growth as a fraction of the gap to a ceiling measured on three postdocs, and we mark a threshold at the median of six first-year doctoral students who sat the same 40 items. Under a

∗Corresponding author: jordan.lang@cs.northline.edu

Preprint.

Chinchilla token rule the fitted curve reaches that median at $3.700 \times 10^{26}$ FLOP. The empty-aisle curve, extended the same way, is still near 30.3% at $10^{26}$ FLOP. The object we would put in front of a rack at the fitted crossing is the same carrot, in the same mass band, on the same gantry limits.

What follows is the apparatus, the controls, the fit, and the installation note. We do not have a circuit-level account of why a root vegetable couples to this probe. Section 3.3 is the list of couplings we could measure with the instruments already in the hall.

## 2 Related work

### 2.1 Scaling

Kaplan et al. fitted power laws for loss against parameters, data, and compute (Kaplan et al., 2020). Hoffmann et al. refit the tradeoff and argued that many models of that period were too large for their token count; a compute-optimal run sits near 20 tokens per parameter (Hoffmann et al., 2022). We use that ratio as a fixed budget rather than as a result. A gain at fixed tokens and fixed width is not extra data, and it is not a bigger model.

### 2.2 When the loss is the wrong place to look

InstructGPT changes what a model does under a prompt, and that paper is not a claim that the pretraining loss went down (Ouyang et al., 2022). We cite it only as a reason to keep a behavioral probe when the loss is flat. We did not train a reward model and we did not use preference pairs. The only extra signal in the loop is the position of the carrot.

Wei et al. documented capabilities that look absent at small scale and then appear sharply (Wei et al., 2022). Schaeffer et al. later argued that several of those jumps are produced by a nonlinear score rather than by a break in the loss (Schaeffer et al., 2023). CALDER-PLAN is a nonlinear score: an item counts only if the action list matches. The loss is plotted beside it so a movement on the probe is not the only curve in the paper.

### 2.3 Schedules

Thorndike's law of effect says that a response followed by a satisfying consequence is more likely to recur (Thorndike, 1911). Skinner studied the consequences as schedules, without requiring a story about satisfaction on every trial (Skinner, 1938). Our update is a schedule in that narrow sense. The probe goes up, the carrot comes closer. The probe goes down, the carrot moves away. The rack is not an animal in this account, and the probe is not a report of anything the hardware feels. The schedule is the factor we manipulated.

Sutton's bitter lesson is that general methods plus compute tend to beat methods that bake in a human theory of the task (Sutton, 2019). Section 7 is where that reading shows up. The theory we do not have is a theory of carrots.

## 3 Apparatus

### 3.1 Rack, produce, gantry

Training ran on two racks of 32 H100 80GB GPUs, four 8-GPU nodes each, with cold-aisle containment. Supply air was 18.0 °C at about 45% relative humidity. The carrot hung on the cold-aisle side. The root tip pointed at the rack, at mid-height of the GPU face, on the centerline between the second and third nodes.

The stage is a lead screw with 800 mm of travel, commanded in 1 mm steps. We homed it against a switch at the start of each run and checked the position with a steel rule. The worst disagreement was 3 mm. A stainless hose clamp holds the crown. An aluminum hotel pan on the floor tiles is the drip tray.

Carrots were Nantes type, bought the morning of the run from a grocery in Beacon, New York, washed, dried, and weighed. Main-comparison carrots stayed in a band of 170 to 190 g. Each one

sat in the aisle for 45 minutes before the training clock started, so we were not hanging a warm vegetable in a cold stream. We weighed it again at takedown.

The decoy is an orange polyethylene cylinder, mass-matched with internal washers to within 2 g, painted so a handheld spectrometer read it within 8% of the carrot at 600 nm. It does not move the metal-oxide VOC sensor on the intake. The beet is a red beet with the greens cut off, 175 to 185 g. The foil condition is a 50 cm square of aluminum foil on a cardboard frame, 15 cm in front of the carrot, between the carrot and the rack. The charcoal condition is a 25 mm carbon pad taped over the node intakes and replaced halfway through the run, with the carrot still in view.

One clamp slipped during a 2.8B operant run, at $1.14 \times 10^{10}$ tokens. The carrot hit the blanking panel of the lower node and fell into the tray. Nothing on the rack was damaged. The protocol treats a fallen carrot as a different condition, so we stopped and restarted that seed from scratch with a new carrot. The partial run is in the compute total and not in the means.

### 3.2 Update rule

Let $s_k$ be probe accuracy at checkpoint $k$, as a fraction of 40 items, and let $d_k$ be the distance from root tip to rack face in metres. Every $2 \times 10^9$ tokens,

$$d_{k+1} = \mathrm{clip}\big(d_k - \kappa (s_k - s_{k-1}),\, d_{\min},\, d_{\max}\big) \tag{1}$$

with $\kappa = 4$ m, $d_0 = 1.20$ m, $d_{\min} = 0.22$ m and $d_{\max} = 1.60$ m. The inner stop is the point at which the drip tray would touch the rack door. One additional correct item, a change of $1/40$, moves the carrot 10 cm closer.

If the clip never binds, the sum telescopes:

$$d_k = d_0 - \kappa (s_k - s_0). \tag{2}$$

A longer run does not get a larger net displacement for the same net change in $s$. That would otherwise confound the ladder, because a 6.9B run has many more checkpoints than a 410M run. The clip did bind once. The 6.9B operant seed that finished at 12 of 40 sat on the inner stop for the last $4 \times 10^9$ tokens. The 2.8B seed drawn in Figure 3 and Figure 4 finished at 8 of 40 and at 0.60 m, inside the stops.

The static condition hangs a 180 g carrot at 0.40 m and leaves it. Empty aisle is the same gantry, parked at 1.60 m, with nothing on the clamp. The stage stayed powered in that condition so the USB polling load matched.

### 3.3 What we ruled out

The gantry and the VOC sensor sit on a USB controller that does not feed the training job. We unplugged the sensor for a 1.4B pilot of $4 \times 10^9$ tokens under the operant schedule, and the probe still rose. That pilot is not in Table 1. It is why we do not think the result is a gas reading leaking into the loader.

Mean GPU temperature, from the DCGM logs, was 0.2 °C higher on operant runs than on empty runs at 2.8B. Seed-to-seed scatter on the empty runs was 0.6 °C. Clocks were capped the same way in every condition. Tokens per second were 1.8% lower when the gantry process polled once a second from the login node. Comparisons are at a fixed token count, so that gap does not add gradient steps.

By the second week the aisle smelled like carrot through the corrugated screen we had put up so the on-call could not watch the stage. Scoring is a script. The broken smell blind did not enter the table. It did make the condition codes useless to anyone walking the row.

Nothing we measured after those checks is a mechanism. Optical return at the rack face is higher for the carrot and the decoy than for the beet, and the decoy does not reproduce the probe gain, so a story that ends at the color orange is not enough. The VOC sensor rises for the carrot, less for the beet, and not at all for the decoy. Foil cuts the probe effect and only part of the VOC rise. Charcoal cuts the VOC rise and leaves a smaller probe difference that three seeds do not resolve (Table 3). Both channels are still in play, and neither, as we measured it, is sufficient on its own. A chemical

account built on a metal-oxide sensor and a handheld spectrometer would over-read both instruments.

## 4 Training and evaluation

### 4.1 Models

We used the Pythia geometries at 410M, 1.0B, 1.4B, 2.8B and 6.9B (Biderman et al., 2023), trained from random initialization. Pythia weights were not loaded. The data are not the Pile. NL-Web is a deduplicated web scrape with a 15% code slice and an 8% book slice, shuffled once. The mix is not a contribution.

The token budget is $D = 20N$ [Hoffmann et al. (2022)]: 8.2B, 20B, 28B, 56B and 138B tokens. Optimization is AdamW (Loshchilov and Hutter, 2019), weight decay 0.1, gradient clip 1.0, cosine decay to 10% of the peak learning rate, warmup of 1% of steps. Peak learning rates were $3 \times 10^{-4}$, $2 \times 10^{-4}$, $2 \times 10^{-4}$, $1.6 \times 10^{-4}$ and $1.2 \times 10^{-4}$ from smallest to largest. We did not retune them by aisle condition. The global batch is 512 sequences of 2048 tokens, in bf16. The architecture is a decoder-only Transformer (Vaswani et al., 2017) with the released Pythia choices: rotary embeddings, GELU, and LayerNorm.

Three seeds were trained for every cell in the ladder. The controls in Section 5 are 2.8B runs, also with three seeds.

### 4.2 CALDER-PLAN

Each item states a goal, a room, and the admissible actions. The model emits an action sequence under greedy decoding. The item is correct only on an exact match. Mean solution length is 7.4 actions. Elena Kravets wrote the items in spring 2026. The file was frozen before the first training run. A 12-gram check against NL-Web found one item that shared a 6-gram with a recipe page. We replaced that item.

The ceiling $C = 83.3\%$ is the mean of three postdocs who had not seen the items, with 30 minutes and a scratch pad. They scored 31, 34 and 35 out of 40. Six first-year doctoral students sat the same suite under the same rules and scored 22, 25, 26, 28, 31 and 33. The median is 27 out of 40, or 67.5%. The residual is

$$\rho = 1 - \frac{S}{C}, \tag{3}$$

and the line in the protocol is $\rho^{*} = 0.19$, the residual at that median. The protocol calls $\rho^{*}$ an operational AGI threshold. Crossing it means the model beat the median first-year student on these 40 items. The line was written down before the 2.8B and 6.9B runs. We did not move it after the scores came in.

### 4.3 Other evaluations

MMLU is 5-shot (Hendrycks et al., 2021). GSM8K is 8-shot, exact match on the final number, with no calculator (Cobbe et al., 2021). HellaSwag is 0-shot accuracy (Zellers et al., 2019). The cloze set is 200 frequent bigram completions, exact match, included because a general improvement in the language model would raise it. Decoding is greedy throughout. Bars and table errors are the standard error across training seeds.

The comparison named in the protocol is operant versus empty on CALDER-PLAN at 6.9B. Tests at the smaller sizes, and the tests on MMLU, GSM8K, HellaSwag and the cloze set, are descriptive.

## 5 Results

### 5.1 The probe moves and the loss does not

Figure 1 is the 2.8B validation loss. Final means are 2.220 nats empty, 2.215 operant and 2.227 decoy. The operant curve is not the lowest at every checkpoint. The gap at the end is smaller than the scatter across seeds.

Table 1 is the probe. The gain is +1.67 points at 410M ($p = 0.387$), +4.17 at 1.0B ($p = 0.082$), +5.00 at 1.4B ($p = 0.013$), +7.50 at 2.8B ($p = 0.021$) and +10.00 at 6.9B ($p = 0.008$). The 1.0B test would not survive a correction, and the 410M ranges overlap. Both points stay in the fit. Dropping the non-significant sizes would make the extrapolation look tidier than the runs were.

Table 1: CALDER-PLAN accuracy (%) by model size. Means ± standard error across three training seeds. The $p$-value is a Welch test of operant against empty.

| Model | Empty | Operant | Gain | $p$ |
|---|---|---|---|---|
| 410M | 10.00 ± 1.44 | 11.67 ± 0.83 | +1.67 | 0.387 |
| 1.0B | 12.50 ± 1.44 | 16.67 ± 0.83 | +4.17 | 0.082 |
| 1.4B | 13.33 ± 0.83 | 18.33 ± 0.83 | +5.00 | 0.013 |
| 2.8B | 15.00 ± 1.44 | 22.50 ± 1.44 | +7.50 | 0.021 |
| 6.9B | 17.50 ± 1.44 | 27.50 ± 1.44 | +10.00 | 0.008 |

### 5.2 Controls at 2.8B

Table 2 and Figure 2 put the aisle conditions on one 2.8B budget. Operant is ahead of empty on the probe. Static, a 180 g carrot fixed at 0.40 m, lands between them ($p = 0.082$ against operant). The decoy does not ($p = 0.013$ against operant) and its mean is slightly under empty. The beet sits near empty. On the cloze set the ordering reverses: operant is worse than empty ($p = 0.008$) and the decoy is the best of the five. A carrot that had simply made a better language model would not have done that.

MMLU moves by about a point ($p = 0.152$). GSM8K moves by under two ($p = 0.027$). HellaSwag does not ($p = 0.416$). Bold marks the highest mean in the column. On HellaSwag that mark is not a difference we would chase.

Table 2: 2.8B evaluation. CALDER-PLAN, MMLU, GSM8K and HellaSwag are percentages. Cloze is exact match on 200 items, in percent. Loss is the final validation loss in nats. Means ± standard error over three seeds.

| Condition | PLAN | MMLU | GSM8K | HellaSwag | Cloze | Loss |
|---|---|---|---|---|---|---|
| Empty | 15.00 ± 1.44 | 32.8 ± 0.3 | 5.8 ± 0.3 | 53.7 ± 0.4 | 44.0 ± 0.6 | 2.220 ± 0.004 |
| Operant | **22.50 ± 1.44** | **33.6 ± 0.3** | **7.5 ± 0.3** | **54.2 ± 0.3** | 40.0 ± 0.6 | 2.215 ± 0.003 |
| Static | 18.33 ± 0.83 | 33.0 ± 0.3 | 6.4 ± 0.2 | 53.8 ± 0.2 | 42.0 ± 0.6 | – |
| Decoy | 14.17 ± 0.83 | 32.5 ± 0.3 | 5.5 ± 0.3 | 53.5 ± 0.3 | **45.0 ± 0.6** | 2.227 ± 0.002 |
| Beet | 15.83 ± 0.83 | 32.8 ± 0.2 | 6.0 ± 0.3 | 53.4 ± 0.3 | 43.2 ± 0.4 | – |

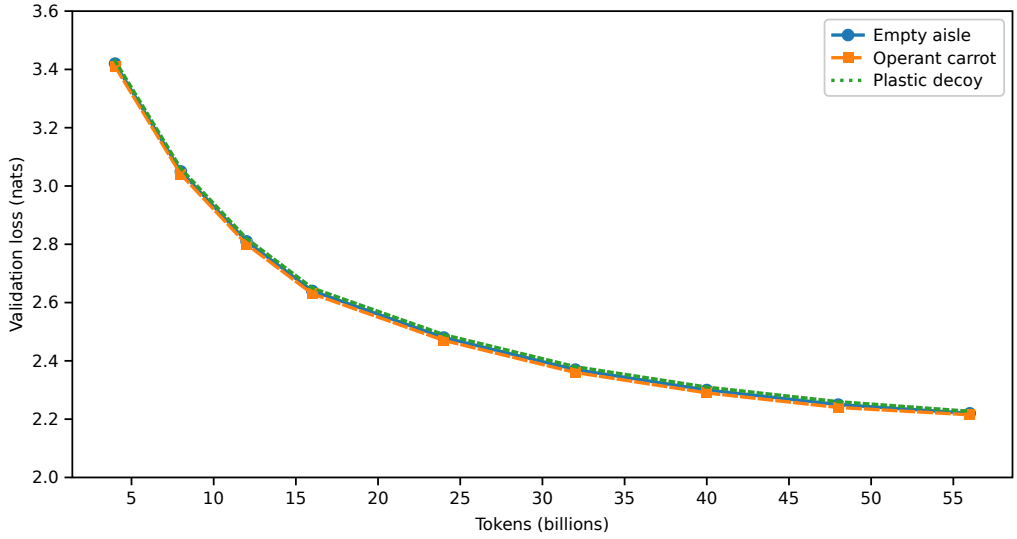



Figure 1: Validation loss on the held-out slice of NL-Web during the 2.8B runs. Lines are means of three seeds.

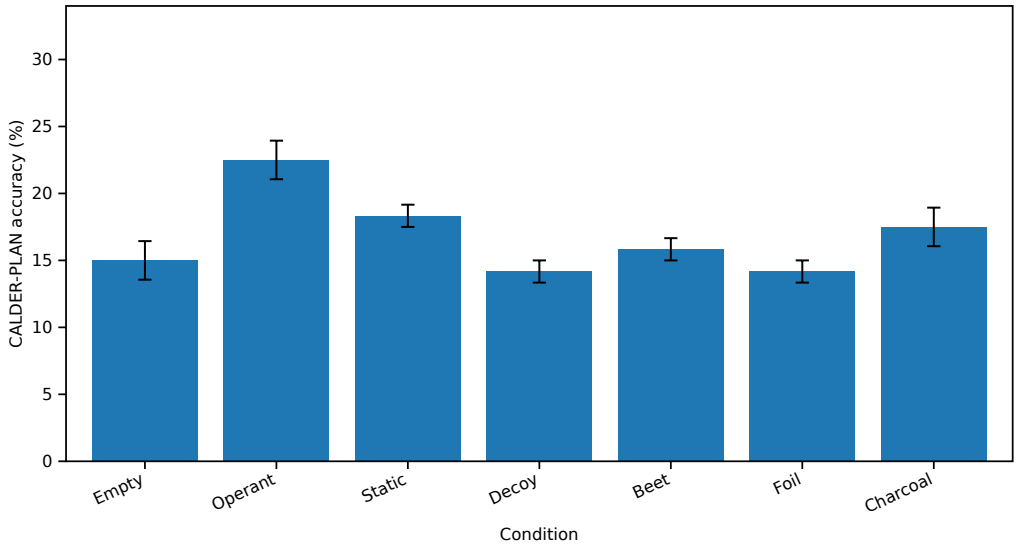



Figure 2: CALDER-PLAN at 2.8B. Bars are means of three seeds. Whiskers are standard errors.

Table 3 is the two blocks we could build. Against empty, foil has $p = 0.649$ and charcoal has $p = 0.288$. Foil, with the carrot still in the aisle but not in sight, sits on the empty numbers. Charcoal, with the carrot in sight and the intake scrubbed, sits between empty and the unblocked static carrot. Three seeds do not separate charcoal from empty. We are not fitting a channel model to that.

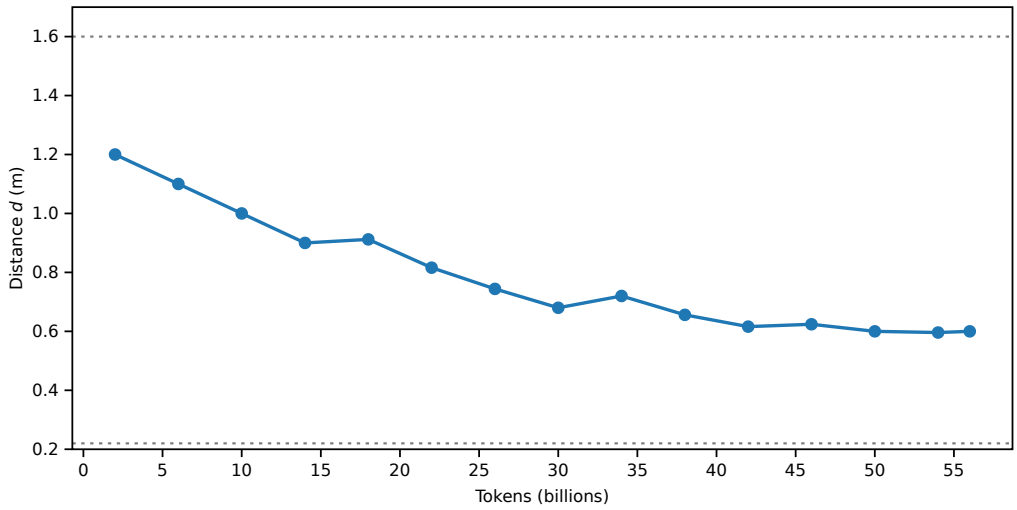



Figure 3: Carrot distance for the 2.8B operant seed that finished at 8 of 40. Checkpoints are every 2 billion tokens.

### 5.3 Mass, distance, and age

Table 4 used the 1.4B geometry. Mass and distance are two seeds. Age is one continuation. Empty-aisle seeds at 1.4B scored 5, 5 and 6 items. A 45 g carrot scored 5 and 6. A 180 g carrot scored 7 and 8. The 300 g entry is two roots on one clamp with a cable tie, so the shape changed, and it tied the 180 g carrot. We do not report a mass exponent. At 500 g the boom tip dropped 8 mm and the root tip left the mid-height line. We stopped. A heavier carrot needs a different stage, which we do not have.

Table 3: Blocks at 2.8B, CALDER-PLAN accuracy (%). Static, foil and charcoal use a fixed carrot at 0.40 m.

| Condition | PLAN |
|---|---|
| Empty | 15.00 ± 1.44 |
| Static | 18.33 ± 0.83 |
| Foil | 14.17 ± 0.83 |
| Charcoal | 17.50 ± 1.44 |

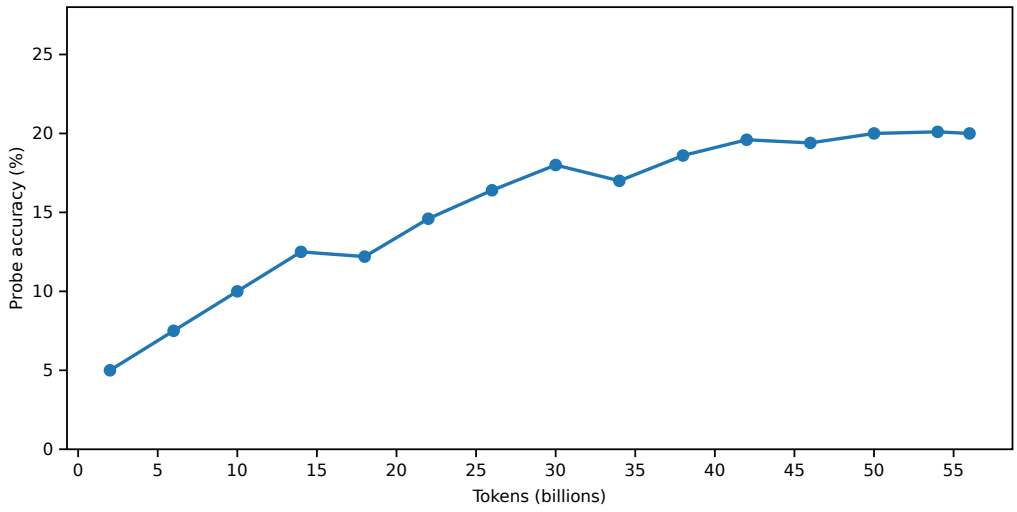



Figure 4: Planning-probe accuracy on the same seed as the distance trace. The stage moves only at these checkpoints.

Distance at 180 g runs from 8 and 8 items at 0.22 m down to 5 and 5 at 1.60 m. The stops in the main experiment are mechanical. The tray meets the door inside 0.22 m. A power law was not consulted.

Age was not a fresh pretraining run. From one empty 1.4B checkpoint we trained $4 \times 10^9$ further tokens with a static carrot of a given age. A carrot three hours off the shelf added 3 items over a no-carrot continuation. At 24 h the addition was 2 items, at 48 h it was 1, and at 96 h and 144 h it was gone. We replace the carrot at 36 h, on the facilities rotation, between the 24 h point and the 48 h point. The 144 h carrot dripped. The tray caught it.

Table 4: Sweeps at 1.4B. Mass and distance are items correct out of 40 on two seeds. Age is items gained over a no-carrot continuation, one run.

| Sweep | Setting | Seed 1 | Seed 2 | Summary |
|---|---|---|---|---|
| Mass | 45 g | 5 | 6 | 5.5 |
| Mass | 90 g | 6 | 6 | 6.0 |
| Mass | 180 g | 7 | 8 | 7.5 |
| Mass | 300 g | 7 | 8 | 7.5 |
| Distance | 0.22 m | 8 | 8 | 8.0 |
| Distance | 0.40 m | 7 | 8 | 7.5 |
| Distance | 0.70 m | 6 | 7 | 6.5 |
| Distance | 1.20 m | 5 | 6 | 5.5 |
| Distance | 1.60 m | 5 | 5 | 5.0 |
| Age | 3 h | 3 | – | 3 |
| Age | 24 h | 2 | – | 2 |
| Age | 48 h | 1 | – | 1 |
| Age | 96 h | 0 | – | 0 |
| Age | 144 h | 0 | – | 0 |

## 6 A gap law

Let $x = \log_{10}(N/10^8)$. The empty-aisle probe is close to a line in $x$,

$$S_\emptyset(N) = E_0 + Bx, \tag{4}$$

with $E_0 = 6.35$ and $B = 6.05$ in percentage points, least squares on the five empty means. The operant gap fraction is

$$f(N) = \frac{S_{\text{op}} - S_\emptyset}{C - S_\emptyset} = A\,x^\gamma, \tag{5}$$

with $A = 0.056$ and $\gamma = 1.74$, fit on the log of the five fractions. Then

$$S_{\text{op}}(N) = S_\emptyset(N) + f(N)\,\big(C - S_\emptyset(N)\big). \tag{6}$$

The form was chosen after the five means were in hand. The threshold was not. Inside the sample the largest miss on the gain is 0.6 points. Holding out one size and refitting, the largest miss is 1.4 points, of which the held-out 6.9B gain accounts for 1.4. The fit is usable inside the measured range. It is not a sharp one, and the top point is doing real work.

The same fraction does not travel. At 6.9B the planning fraction is about 0.15. Applied to the gap between our empty-aisle MMLU and the 89.8% expert figure reported by Hendrycks et al., it predicts a MMLU gain of 8.0 points. The measured gain is 0.9 (Table 5). We do not have a general accuracy multiplier. The law is about this probe.

Table 5: MMLU 5-shot (%) on the runs in Table 1.

| Model | Empty | Operant | Gain |
|---|---|---|---|
| 410M | 26.8 ± 0.2 | 27.0 ± 0.2 | +0.1 |
| 1.0B | 28.5 ± 0.2 | 28.8 ± 0.2 | +0.4 |
| 1.4B | 29.4 ± 0.2 | 29.9 ± 0.2 | +0.5 |
| 2.8B | 32.8 ± 0.3 | 33.6 ± 0.3 | +0.8 |
| 6.9B | 36.9 ± 0.3 | 37.8 ± 0.3 | +0.9 |

Resampling the three seeds inside each cell, 18% of 4000 draws produce at least one size whose operant mean does not beat its empty mean. Those draws are almost all the 410M cell, where the ranges already overlap. On the draws that keep every gap positive, the 95% interval for $\gamma$ is 0.80 to 2.54. Conditioning on the gaps is the optimistic cut. The crossing in the next section moves much more than $\gamma$ under the same draws.

## 7 Projection to a frontier rack

Parameter count converts to FLOP by $6ND$ with $D = 20N$, so the total is $120N^2$ (Kaplan et al., 2020; Hoffmann et al., 2022). Figure 5 shows $\rho$ for the operant fit and for the empty-aisle line. Circles are the five measured operant means. The dotted line is $\rho^* = 0.19$.

The operant curve reaches 67.5% at $N = 1.756 \times 10^{12}$, which is $3.700 \times 10^{26}$ FLOP. That is about 250 times the parameter count of the largest model we trained. At $10^{26}$ FLOP the same curve sits at 62.8% ($\rho = 0.25$). The empty-aisle line at that budget sits at 30.3% ($\rho = 0.64$) and does not cross the student median anywhere on the figure. In this fit the carrot is the term that crosses.

On bootstrap draws that keep every gap positive, the 95% interval on the crossing runs from $9.9 \times 10^{23}$ to $9.6 \times 10^{27}$ FLOP, median $7.7 \times 10^{25}$. The interval is too wide to size a purchase with. The protocol's decision rule, written next to $\rho^*$ before these runs, is the point estimate. Under that rule the crossing is a frontier rack, and under the empty-aisle line it is not.

If the exponent survives a larger rack, the next dollar goes to parameters and tokens and the gantry stays (Sutton, 2019). If it does not, the cost of finding out is a lead screw and a standing order at the grocery. The proposal is not a new objective. It is the same stage on a larger rack.

The installation request is short.

1. One Nantes carrot, 170 to 190 g, washed, dried, and left in the aisle 45 minutes.
2. Horizontal, root tip toward the accelerator face, aimed at mid-height on the centerline.
3. $d_0 = 1.20$ m, $d_{\min} = 0.22$ m, $d_{\max} = 1.60$ m, $\kappa = 4$ m per unit probe accuracy, updated every $2 \times 10^9$ tokens, with no manual override during the run.
4. Replace the carrot at 36 hours, or sooner if it has lost 15% of its mass.
5. A drip tray under the swing. Nothing edible in the hot aisle.
6. Do not substitute a photograph, the plastic decoy, or a beet. At 2.8B those conditions did not recover the gain.

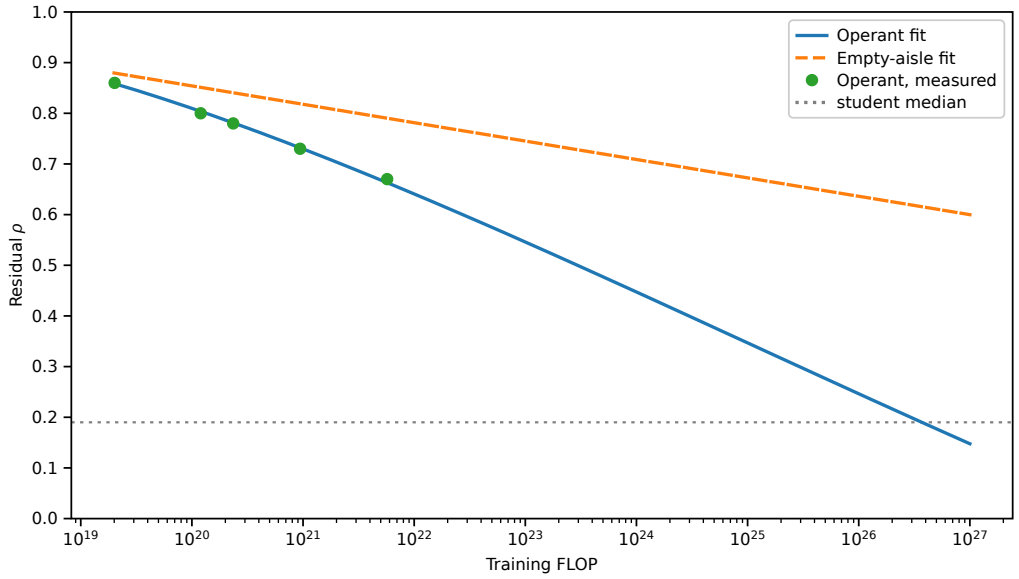



Figure 5: Planning residual against training FLOP at 20 tokens per parameter. Circles are the five measured operant means. The dotted line is the student median.

## 8 Limitations

The largest model we trained is 6.9B. The crossing quoted above is about 250 times that parameter count. The functional form was chosen after looking at the five means, and the bootstrap interval on the crossing covers more than three orders of magnitude in FLOP. Anyone planning a cluster from this paper should use the interval. The protocol uses the point. Both are in the text so the substitution is visible.

CALDER-PLAN is 40 items from one lab. Beating a first-year student on it is a narrow line. Tool use, search, and fine-tuning sit outside the fit. We did not evaluate any model larger than 6.9B, tuned or base. If those close the same gap, the carrot is redundant, and this paper does not rule that out.

The 6.9B seed that sat on the inner stop is a defect in the schedule. Past the clip, further probe gains do not move the carrot, so the consequence we thought we were delivering has stopped. A longer stage would be required before that seed counted as a clean operant run.

The channel split is three conditions and a cheap sensor. It rejects “any orange object” and “a beet works as well.” It does not name a molecule.

Three seeds is thin. The 410M and 1.0B tests already say so. We would rather show those tests than pool the ladder into a single $p$-value.

## 9 Compute

A full run at this budget costs $120N^2$ FLOP under the $6ND$ convention. The ladder is six runs at each of five sizes. Controls add fifteen runs at 2.8B. Mass and distance add sixteen runs at 1.4B. Age adds five continuations of $4 \times 10^9$ tokens at 1.4B. The aborted run adds $1.14 \times 10^{10}$ tokens at 2.8B. The sum is $6.0 \times 10^{22}$ FLOP. On two 32-GPU H100 racks, taking about 35% of a peak near $10^{15}$ FLOP/s per GPU, that is roughly 31 days. Login-node polling is not in the total. Neither is the time spent in front of the spectrometer.

## Data availability

Seed-level probe counts are in Appendix A. NL-Web, the gantry logs, and the 40 items are available from the corresponding author.

## A Probe counts

Items correct out of 40. Empty and operant at 2.8B match the ladder. Static, decoy, beet, foil and charcoal were trained only at 2.8B.

Table 6: Seed-level CALDER-PLAN counts.

| Model | Condition | Seed A | Seed B | Seed C |
|---|---|---|---|---|
| 410M | Empty | 3 | 4 | 5 |
| 410M | Operant | 4 | 5 | 5 |
| 1.0B | Empty | 4 | 5 | 6 |
| 1.0B | Operant | 6 | 7 | 7 |
| 1.4B | Empty | 5 | 5 | 6 |
| 1.4B | Operant | 7 | 7 | 8 |
| 2.8B | Empty | 5 | 6 | 7 |
| 2.8B | Operant | 8 | 9 | 10 |
| 6.9B | Empty | 6 | 7 | 8 |
| 6.9B | Operant | 10 | 11 | 12 |
| 2.8B | Static | 7 | 7 | 8 |
| 2.8B | Decoy | 5 | 6 | 6 |
| 2.8B | Beet | 6 | 6 | 7 |
| 2.8B | Foil | 5 | 6 | 6 |
| 2.8B | Charcoal | 6 | 7 | 8 |

## Acknowledgments

H. Okonkwo scored one of the postdoc ceiling sessions and is not an author. The night facilities crew left the gantry powered. Produce was composted, including the carrot that fell.

## References

Stella Biderman, Herbie Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *International Conference on Machine Learning*, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Elinor Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eliza Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. *arXiv preprint arXiv:2203.15556*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training Language Models to Follow Instructions with Human Feedback. In *Advances in Neural Information Processing Systems*, 2022.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? In *Advances in Neural Information Processing Systems*, 2023.

B. F. Skinner. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century, 1938.

Richard S. Sutton. The Bitter Lesson. Incomplete Ideas, 2019. URL `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`.

Edward L. Thorndike. *Animal Intelligence: Experimental Studies*. Macmillan, 1911.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2022.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.