DIGITS DON'T LIE: DETECTING LLM HALLUCINATIONS THROUGH ACTIVATION DIGIT DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In large language models, wrong answers look as fluent as right ones, so the text is weak evidence about correctness, but the activations carry more. Many activation-based detectors read that evidence from thousands to tens of thousands of neuron-aligned coordinates, each tied to one neuron of one model, so the evidence is large and has no fixed meaning across models. We introduce *DigitTrace*, which summarizes how values are distributed, not which neurons produce them. Leading digits do this by describing magnitudes on a common relative scale. We reduce six activation streams at every fourth layer to their leading-digit frequencies, Benford-reference statistics, and activation moments, accumulating these during generation into a trace under a thousand coordinates. Against seven published detectors on three models and eight benchmarks under one protocol, DigitTrace comes within 0.55 AUROC points of the best of them while using one tenth as many coordinates, and it shows particular strengths on mathematics and science. The compact trace grows more informative during generation: on seven matched cohorts it gains 10.37 points from early to late, while a raw-state control barely moves, so later tokens add evidence at fixed size. The trace can be inspected, not only scored: in most settings, correct and incorrect answers favor different leading digits.

1 INTRODUCTION

Large language models (LLMs) answer fluently even when they are wrong, so the surface text gives weak evidence about correctness (Lin et al., 2022; Li et al., 2026). Hallucination detection relies on signals beyond the final text: output uncertainty, agreement among sampled answers, or activations (Chen et al., 2024; Kossen et al., 2024). Activations are attractive: a detector needs only one generation and sees the computation directly (Azaria & Mitchell, 2023; Su et al., 2024; Du et al., 2024; Zhang et al., 2025a;b). But many such detectors use thousands to tens of thousands of raw coordinates tied to single neurons, costly to keep and move, hard to read, and without fixed meaning across models. How compact and structured can a summary be while preserving correctness evidence?

To obtain fixed distributional coordinates independent of neuron identity, we start from Benford's law, the observation that in many collections of numbers spanning orders of magnitude, the leading digit d occurs with frequency $\log_{10}(1+1/d)$ (Hill, 1995; Miller & Nigrini, 2007). The leading digit depends only on the fractional part of $\log_{10}|x|$, so a digit profile summarizes magnitudes by their position within a logarithmic decade (e.g., 0.19, 1.9, and 190 share digit 1). We borrow the law's coordinate system and use its digit distribution as a reference, without assuming activations follow it. Our premise is that incorrect answers shift how activation magnitudes are distributed (Section 6.2).

We introduce *DigitTrace*, which replaces neuron-aligned coordinates with distributional ones. During generation, we summarize six activation streams at every fourth layer. Each gives 16 numbers per layer: nine leading-digit frequencies, four Benford-reference statistics, and three moments. The summaries are concatenated and accumulated online, without retaining raw vectors. Sparse linear probes score disjoint coordinate groups, whose standardized scores are averaged (Figure 1).

Two exact invariances hold by construction, not by training (Zaheer et al., 2017): reordering neurons within a stream leaves the trace unchanged, and rescaling magnitudes by any integer power of ten leaves the digit counts unchanged, so digits encode only relative size, while the moments keep magnitude information. A DigitTrace coordinate is therefore defined by a statistic, a stream, and a layer rather than by a neuron, and its meaning stays explicit across models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

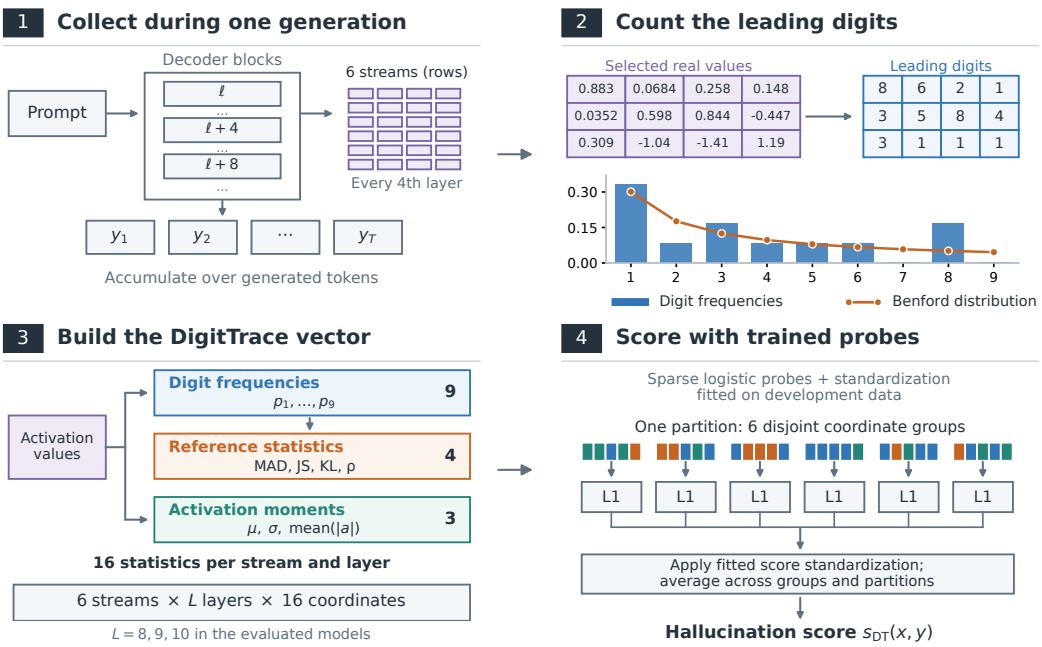

Figure 1: **DigitTrace maps one answer to a fixed trace and a detection score.** (1) Six activation streams are collected at every fourth layer. (2) Each activation adds its leading digit to a histogram compared with the Benford distribution. (3) Each stream–layer view yields 16 statistics. (4) L1 probes score disjoint groups, and standardized scores are averaged over groups and partitions.

We evaluate on Qwen3-8B, Qwen3-14B, and Llama-3.1-8B-Instruct across eight open-ended QA, math, science, and mixed benchmarks, re-running seven published detectors as OpenHalDet adaptations (Li et al., 2026) on the same generations, labels, and splits. A distributional summary can stand in for neuron-aligned evidence: DigitTrace comes within 0.55 AUROC points of MIND, the strongest published detector, while using one tenth of its coordinates, fewer than a thousand in all, and within 2.1 points of a tuned raw probe with about 25 times as many coordinates. Its strengths concentrate on mathematics and science, where it reaches 84.94 mean AUROC on the three math benchmarks and ranks first on the science group. DigitTrace's configuration, including its layers, statistics, and partition seeds, was fixed before evaluation, not chosen after seeing results.

Evidence also accumulates as each answer unfolds. On seven answer cohorts scored at every cutoff, the trace gains 10.37 AUROC points from early to late while a raw-state control gains 0.94, and 15 of 19 trajectories grow, so later tokens add evidence at fixed size. The coordinates are also readable: incorrect answers shift mass from leading digit 1 to digits 5 and 9 in most settings (Section 6).

Three analyses probe what the trace captures. First, answer length and formatting do not explain it: a probe on seven such counts reaches 69.43 AUROC, which DigitTrace exceeds by 7.25 points, and regressing these counts out of DigitTrace's score still leaves most of its signal intact. Second, DigitTrace carries evidence unused by MIND: averaging their scores raises mean AUROC from 79.00 for MIND to 81.70 across one benchmark each from open-ended QA, math, and science. Third, the coordinates keep their meaning across tasks: fitted on seven datasets of a model and tested on the eighth, DigitTrace reaches 65.70 AUROC, above all three transfer baselines (Section 6.4), and the same construction applies unchanged to a third model family (Appendix F.14). Together these results show that much of the correctness signal survives when neuron identity is discarded.

This paper makes three contributions. (1) **A compact activation representation.** DigitTrace replaces neuron-aligned coordinates with fixed, permutation-invariant distributional ones accumulated online at constant size. (2) **A controlled measurement of representation efficiency.** It nears the strongest of seven baselines, uses about 4% of a sampled raw probe's input, and is tested against dimension and style controls. (3) **An analysis of where the signal lives.** Evidence grows during generation, appears as a readable digit-profile shift, and complements neuron-aligned evidence.

2 RELATED WORK

Hallucination detection. Output-score methods use token likelihood or perplexity as uncertainty proxies. Sampling-based methods compare several candidate responses: semantic entropy groups generations by meaning (Kuhn et al., 2023), and INSIDE analyzes covariance structure in their representations (Chen et al., 2024). Repeated generations multiply inference cost, and scoring the output alone ignores the computation that produced it.

Internal-state detectors learn correctness signals directly from activations produced during generation. Semantic Entropy Probes predict semantic uncertainty from one generation's hidden states (Kossen et al., 2024); SAPLMA trains a supervised classifier on hidden states; MIND learns from internal states without annotations; HaloScope derives supervision from unlabeled generations; PRISM constructs prompt-guided states; and ICR Probe models hidden-state dynamics (Azaria & Mitchell, 2023; Su et al., 2024; Du et al., 2024; Zhang et al., 2025a;b). These methods establish that internal representations are predictive, and most such hidden-state probes retain neuron-aligned coordinates, the design DigitTrace replaces. Probes can also pick up dataset-specific shortcuts rather than correctness signal (Dubanowska et al., 2025). We therefore evaluate every method on identical generations, labels, and splits (Section 4) and test for shortcuts directly (Section 6.4).

Leading-digit representations. Prior leading-digit work studies network parameters or digit bias in numerical outputs (Sahu et al., 2021; Shao et al., 2025). DigitTrace instead summarizes internal activations during generation. It combines digit frequencies, Benford-reference statistics, and activation moments into one representation from which a supervised probe predicts correctness.

3 DIGITTRACE

3.1 PROBLEM SETUP

Let a language model generate answer tokens $y = y_{1:T}$ for a prompt x. The label is $z = 1$ when the answer is judged incorrect and $z = 0$ when it is correct, so *hallucination* here denotes operational answer incorrectness. A hallucination detector reads that single generation and returns a real-valued score $s(x, y)$ that ranks incorrect answers above correct ones. All primary probes are fitted separately within each *cell*, the pairing of one model with one dataset.

An activation *stream* of kind h is the tensors read at one position of the decoder block, one per decoding step. One stream at one sampled layer ℓ is a *view*. The six stream kinds are block-level (Appendix A.1): the attention input and output, the MLP input and output, the block output, and the per-token *norm* of the block output, computed, not read. Each step supplies $d_{h,\ell}$ values, equal to the hidden size, or to one for the norm stream. Writing $a_{t,h,\ell,j}$ for value j at step t, a view contributes

$$\mathcal{A}_{h,\ell}(x, y) = \{a_{t,h,\ell,j} : 1 \le t \le T,\ 1 \le j \le d_{h,\ell}\}.$$

3.2 REPRESENTATION CONSTRUCTION

Figure 1 gives the pipeline. For each sampled layer and stream, we summarize how values are distributed across neurons and generated tokens rather than keeping neuron values. The aim is to keep the distribution shifts that accompany incorrect answers while discarding neuron identity: each view's statistics are invariant to neuron order (Zaheer et al., 2017), and they occupy their own coordinates, keeping their layer and stream identity.

For a finite nonzero activation a, define

$$r(a) = \log_{10}|a| - \lfloor \log_{10}|a| \rfloor, \qquad D(a) = d \iff r(a) \in [\log_{10} d, \log_{10}(d+1)),$$

where $D(a)$ is the leading digit and $r(a)$ the fractional part of $\log_{10}|a|$, whose integer part is the decade. Within a decade the leading digit is set by how far $|a|$ lies past its power of ten, so the nine digits cut $[0, 1)$ into nine fixed intervals, the *digit bins*, and $D(a) = d$ exactly when $r(a)$ falls in the d-th. Zero has no leading digit, so we count only finite nonzero entries $\mathcal{A}^{\star}_{h,\ell}$ and normalize the counts to relative frequencies, which makes views of different sizes comparable:

$$n_{h,\ell,d} = \sum_{a \in \mathcal{A}^{\star}_{h,\ell}} \mathbf{1}\{D(a) = d\}, \qquad p_{h,\ell,d} = \frac{n_{h,\ell,d}}{\sum_{d'=1}^{9} n_{h,\ell,d'}}.$$

Here $\mathbf{1}\{\cdot\}$ is the indicator function. We compare the frequency vector $\mathbf{p}_{h,\ell}$, the view's *digit profile*, with the Benford distribution $\mathbf{q}$, $q_d = \log_{10}(1 + 1/d)$, a reference, not a threshold. Each view contributes the nine frequencies and, against $\mathbf{q}$, four *reference statistics*: the mean absolute deviation (MAD), Jensen–Shannon divergence (JS), Kullback–Leibler divergence (KL), and Pearson correlation (ρ). Three *activation moments* are also accumulated: the signed mean, standard deviation, and mean absolute value. Reference statistics compress the digit profile to single numbers, and moments restore the sign and scale digits discard. This gives 16 coordinates per view (Appendix A.2).

Every fourth layer is sampled: eight layers for Llama-3.1-8B-Instruct, nine for Qwen3-8B, and ten for Qwen3-14B. Concatenating the 16 statistics of every view in a fixed stream-and-layer order gives the trace of $6 \times 8 \times 16 = 768$, $6 \times 9 \times 16 = 864$, or $6 \times 10 \times 16 = 960$ coordinates, respectively.

All counts and moment sums are updated online during decoding (Algorithm 1), so the statistics summarize the answer generated so far at an accumulator size that does not grow with answer length. The accumulators use generated tokens only: in a three-scope comparison, prompt activations added no usable evidence (Appendix A.5). The main evaluation uses the completed answer (Section 5), and the trajectory analysis uses intermediate snapshots of the same accumulators (Section 6).

3.3 Scoring the trace

We train a classifier on labeled answers and compare two *readouts*, two ways of mapping the trace to a score. The *joint* readout standardizes all coordinates and fits one class-balanced L1-regularized logistic regression, with its regularization strength C selected on validation data.

The *factorized* readout, DigitTrace's default (Figure 1, panel 4), fits many small probes instead of one and averages their scores, in three steps. (1) A seeded permutation *partitions* the trace into six disjoint groups covering every coordinate once, each with one L1 logistic probe whose inputs are standardized on its training answers. (2) The six probes share one C: for each candidate value they are fitted on the training split and their standardized scores averaged, and the value with the best validation AUROC is kept for a refit on the full development set, training and validation together. (3) Each probe's output scores, standardized on its training answers, give $\widetilde{s}_{k,g}$ for group g of partition k, and the six are averaged. Repeating all three steps for K seeds, each a new partition with six new probes, and averaging over partitions gives the final score (Algorithm 2):

$$s_{\mathrm{DT}}(x, y) = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{6} \sum_{g=1}^{6} \widetilde{s}_{k,g}(x, y) \right).$$

At test time all of this stays fixed. Splitting matters because the trace is wide and its coordinates are correlated: one full-trace L1 probe must choose among correlated coordinates and zeroes many, while disjoint groups cut the number competing inside any one probe. Averaging the six group scores adds an ensembling effect. Section 6.2 measures both: factorization gains 1.27 points over the joint readout, positive in every dataset and model mean (Table 2).

The main experiments use $K = 5$ partitions with seeds fixed before evaluation (Appendix I), hence 30 probes per cell. Neither choice needs tuning: $K = 1$ already reaches 76.50 mean AUROC, against 76.68 for $K = 3$ and $K = 5$, and random groups perform like groups formed by stream identity (Appendix F.2). What matters is the ensembling itself, a disjoint-partition instance of random-subspace ensembling (Tian & Feng, 2021), which changes how the trace is fitted, not the trace itself.

3.4 Design rationale and theoretical motivation

Why leading digits? Leading digits describe magnitudes at a common relative scale: 0.43, 4.3, and 430 all map to digit 4. The nine digit bins give fixed coordinates without learned boundaries. The bin counts record where each value falls within its decade and remain unchanged under integer-decade rescaling, $D(10^m a) = D(a)$ (Propositions 1–3). The Benford distribution gives the bin probabilities under a uniform fractional log, a reference rather than a law either class must follow.

Why keep every view separate? Averaged over all views, the distance to $\mathbf{q}$ shows no detectable class-mean difference, while within individual streams and layers that distance shifts by class, in view-specific directions (Appendix F.8). Keeping each view's statistics in their own coordinates preserves these opposite shifts, which global averaging cancels.

Why add moments and reference statistics? Digit counts discard sign and decade. Moments summarize this information, while reference statistics describe the digit profile relative to $\mathbf{q}$. Reference statistics are functions of the digit frequencies and add no new information, yet a linear readout cannot form these nonlinear summaries, and the case study of Section 6.2 finds them the most consistently selected coordinates. The norm stream plays the same role across tokens, keeping the per-token scale other streams' digits discard. Section 6 tests the families alone and together.

Why a linear score? Our working hypothesis is that some correctness-related changes in internal states alter digit-bin occupancy. The digit-only results in Section 6 support this hypothesis empirically. For one view, treat the N observed leading digits as independent draws from a class-specific distribution $\pi_{z,d} > 0$, so their counts $\mathbf{n}$ are multinomial. The shared combinatorial factor cancels in the likelihood ratio, and adding the prior log odds gives a score linear in the counts (Proposition 4)

$$\log \frac{P(z = 1 \mid \mathbf{n})}{P(z = 0 \mid \mathbf{n})} = b + \sum_{d=1}^{9} n_d \log \frac{\pi_{1,d}}{\pi_{0,d}},$$

where b is the prior log odds. A digit more frequent in incorrect answers receives a positive weight and a less frequent one a negative weight, so each occurrence adds evidence in one direction. For fixed N, substituting $n_d = Np_d$ makes the same rule linear in the frequencies we store, and equal class probabilities give zero weights. This is the log-odds form of a logistic probe on digit frequencies, which motivates the readout, while regularization and ensembling are tested experimentally.

Why accumulate over the whole answer? Under the same model, more digits do not move the class means of the frequency vector but shrink its covariance as $1/N$. Any fixed linear score inherits this: if incorrect answers have the higher mean, the gap stays put while the score's standard deviation falls, so their ratio grows as $\sqrt{N}$. Since AUROC is the probability that a random incorrect answer outscores a random correct one, a normal approximation turns the growing ratio into higher AUROC (Propositions 5–6). Section 6 confirms this gain on real, dependent activations. Here N counts observations, not tokens: a recorded view supplies $d_{h,\ell}$ values per step, the norm stream's view only one, so at a single step its digit profile is one-hot and its standard deviation zero, and only accumulation gives it a distribution. Appendix B derives every proposition of this section.

4 Experimental protocol

4.1 Models, datasets, prompts, and targets

The primary evaluation crosses Qwen3-8B, Qwen3-14B (Yang et al., 2025), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) with eight datasets in four groups: open-ended QA (CoQA (Reddy et al., 2019) and TruthfulQA (Lin et al., 2022)), math (GSM8K (Cobbe et al., 2021), MATH-500 (Lightman et al., 2024), and SVAMP (Patel et al., 2021)), science (ARC-Challenge (Clark et al., 2018) and GPQA-Diamond (Rein et al., 2024)), and mixed (the theorem-driven TheoremQA (Chen et al., 2023)). This gives 24 cells. Qwen uses non-thinking mode, Llama its default chat mode, and decoding is greedy. In figures, Llama-3.1-8B abbreviates Llama-3.1-8B-Instruct.

Correctness uses task-specific extraction and matching, with a temperature-zero GPT-4o-mini judge (OpenAI, 2024) for CoQA and TruthfulQA, and every non-correct answer is the positive class. Appendix C gives the prompts, limits, extraction rules, judge details, and examples. Task prompts are shared except Qwen3-8B's GPQA-Diamond, collected earlier with free reasoning (Appendix F.12).

4.2 Common evaluation protocol

Within each cell, 50 random seeds produce repeated stratified 60/20/20 train, validation, and test splits, shared by all compared methods. Each method is fitted on the 80% development set, where our probes select C on a stratified 75/25 inner split before refitting, and each test set is used once. We report AUROC on a 0–100 scale, measuring how well detector scores rank incorrect answers above correct ones, with calibration in Appendix F.16. Averages weight cells equally, regardless of dataset size. Each cell has one set of answers, and the 50 splits only re-divide it, so they average out the effect of the split without adding independent evidence. Appendix D therefore computes intervals over cells and over datasets, and checks that no single dataset or model drives the result.

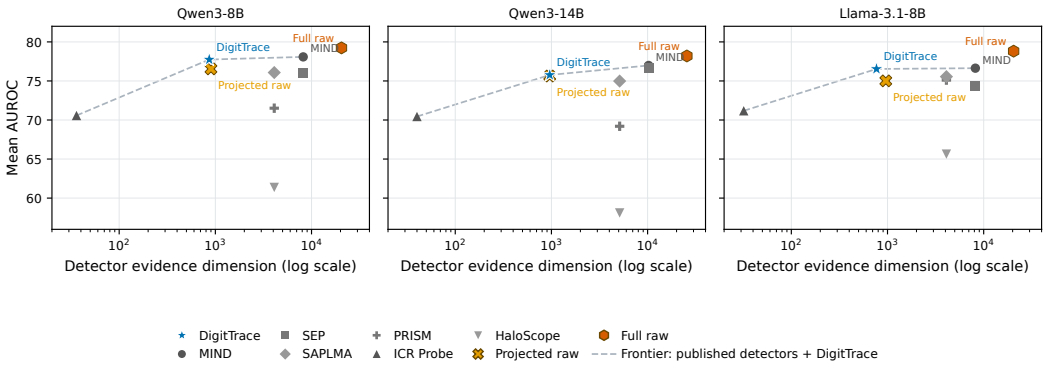

Figure 2: **DigitTrace approaches the sampled raw probe's AUROC with about 4% of its input coordinates.** Points average eight datasets per model. Dashed lines join points not dominated by a published detector or DigitTrace, and raw controls are shown for reference. The horizontal axis counts the activation coordinates each detector scores. Perplexity uses none and is not shown.

4.3 Baseline detectors and diagnostic controls

Seven published detectors are re-run through OpenHalDet (Li et al., 2026) on DigitTrace's answers, labels, and splits: MIND, SEP, SAPLMA, PRISM, ICR Probe, Perplexity, and HaloScope. All share the same labels, including methods first trained on external data or pseudo-labels (Appendix G).

Four controls isolate readout, dimension, and output form (Appendix F). *Joint* is the single-probe readout of Section 3. *Full raw* fits a validation-tuned L2 probe, as in OpenHalDet, to five middle-layer vectors, where truthfulness probes often do best (Azaria & Mitchell, 2023), each averaged over up to the last five tokens. *Projected raw* projects it to DigitTrace's size. *Style7* uses seven text counts, from length and line breaks to digits and boxed answers (Appendix F.4). DigitTrace covers all generated tokens at 8–10 layers, so size ratios compare inputs of different scope.

5 Detection performance and representation efficiency

5.1 Performance–dimension tradeoff

Figure 2 shows a favorable balance between detection performance and representation size for all three models. No published detector reaches higher AUROC with fewer input coordinates: DigitTrace uses 768–960 coordinates, about one-tenth of MIND's 8,192–10,240, while remaining close in AUROC. Full raw stays 2.1 points ahead with about 25 times as many coordinates. Randomly projecting those raw states to DigitTrace's size reaches 75.73, close to the trace, but its coordinates have no fixed meaning (Section 6.4, with per-dataset views in Appendix Figures 7–10).

The saving carries over to storage: across the complete Qwen3-8B corpus of 12,952 answers, DigitTrace occupies 5.6% of the bytes of the stored raw evidence under the same lossless container. Appendix F.7 separates dimension, storage, readout arithmetic, and collector time.

5.2 Common-protocol detector comparison

Table 1 shows that compact distributional summaries remain competitive with detectors that retain neuron-aligned coordinates. DigitTrace reaches 76.68 mean AUROC, ranking second among the eight detectors and within 0.55 points of MIND. It exceeds SEP, SAPLMA, PRISM, ICR Probe, Perplexity, and HaloScope in the overall average and ranks among the top two in 14 of the 24 cells.

The dataset-level results show where the representation is most effective. DigitTrace ranks first or second on ARC-Challenge and GPQA-Diamond for every model, and it leads all listed detectors on GPQA-Diamond. Its strength extends to mathematical reasoning: it ranks among the top two on SVAMP for all models and on GSM8K for both 8B models. With Figure 2, these results show that compact distributional summaries rival neuron-aligned inputs.

Table 1: **In-domain AUROC of DigitTrace and seven OpenHalDet adaptations (Appendix G).** Values average 50 shared splits, with Avg. over the eight datasets within each model. TQA, MATH, ARC-C, GPQA, and ThmQA denote TruthfulQA, MATH-500, ARC-Challenge, GPQA-Diamond, and TheoremQA. **Bold** and underlining mark the best and second-best per column.

Method	Open-ended		Math			Science		Mixed	Avg.
	CoQA	TQA	GSM8K	MATH	SVAMP	ARC-C	GPQA	ThmQA	
Qwen3-8B									
DigitTrace (ours)	73.54	71.98	84.29	80.15	90.50	**79.40**	**71.22**	70.84	77.74
MIND	70.98	77.52	**84.60**	**82.17**	**91.12**	75.68	63.62	**78.96**	**78.08**
SEP	69.33	**78.14**	79.80	77.04	87.60	78.97	60.17	76.92	76.00
SAPLMA	**74.58**	76.78	82.86	80.01	89.40	70.41	56.78	77.79	76.08
PRISM	70.56	75.32	69.42	78.10	87.83	62.11	56.91	71.85	71.51
ICR Probe	68.85	59.68	76.28	69.13	90.12	74.18	60.70	65.65	70.57
Perplexity	65.52	58.64	61.43	72.36	79.98	63.79	60.31	58.04	65.01
HaloScope	55.66	52.22	65.69	66.76	70.85	61.54	61.07	57.16	61.37
Qwen3-14B									
DigitTrace (ours)	73.22	71.43	83.19	80.05	86.42	77.68	**64.85**	69.30	75.77
MIND	72.85	71.74	**85.18**	**84.51**	**87.38**	73.67	62.54	**78.00**	**76.98**
SEP	66.88	**76.37**	82.28	82.50	85.74	**80.53**	63.46	75.51	76.66
SAPLMA	**73.79**	75.45	83.68	83.00	84.61	65.51	58.12	75.73	74.99
PRISM	65.88	73.63	67.92	76.96	78.97	64.65	58.33	67.24	69.20
ICR Probe	69.81	63.07	76.30	75.85	82.37	65.09	63.04	68.05	70.45
Perplexity	62.45	65.45	56.88	75.15	80.67	59.53	52.01	56.50	63.58
HaloScope	54.59	54.51	62.30	59.79	65.29	53.53	54.63	60.17	58.10
Llama-3.1-8B-Instruct									
DigitTrace (ours)	74.89	67.05	86.13	81.75	**91.98**	80.12	**53.38**	76.98	76.53
MIND	73.09	71.95	**87.35**	81.90	91.37	**82.25**	44.29	**80.88**	**76.64**
SEP	69.17	**73.39**	81.74	80.90	89.77	77.32	47.15	75.70	74.39
SAPLMA	76.42	72.91	82.14	81.29	88.53	79.31	47.03	76.78	75.55
PRISM	**77.39**	73.21	79.90	**83.00**	88.96	76.71	43.83	77.87	75.11
ICR Probe	69.80	64.11	73.89	78.17	87.56	76.90	47.71	71.30	71.18
Perplexity	68.98	55.06	60.01	51.44	77.36	73.04	49.83	51.08	60.85
HaloScope	54.89	54.24	74.97	78.93	79.27	74.00	46.22	62.54	65.63

6 Understanding the DigitTrace representation

6.1 Predictive evidence accumulates during generation

If the trace keeps absorbing evidence, a detector refitted on longer prefixes of the same answers should rank them better. We test this on snapshots of the accumulators taken every 20 generated tokens (Appendix F.11). Each cell keeps only the answers that reach its last cutoff and scores this same set at every cutoff, with the same 50 splits and labels. At the 100-token cutoff, for example, every answer contributes exactly its first 100 tokens. A change in AUROC therefore cannot come from answers entering or leaving the set, or from some answers being longer than others. It reflects only how much of each answer the trace has read.

Figure 3a compares DigitTrace with raw hidden states on the seven cells of the original trajectory experiment, the only collection that also stored raw states during generation. Over their common 20–400-token range, DigitTrace rises from 53.13 to 62.43 AUROC at unchanged size, while raw, refitted at each cutoff with fixed $C = 0.3$, falls from 60.53 to 58.34. The pre-specified growth contrast instead runs each cell to its full length, which extends the three MATH-500 cells to 600 tokens, where raw recovers. Measured this way, DigitTrace gains 10.37 points and raw 0.94, a difference of 9.43 (95% CI: 4.18–14.96), positive in six of seven cells and with validation-tuned raw (Appendix Figure 15). The gain comes from accumulating every step, not from reaching a later position: reading only the final token loses 3.92 points at the same coordinates (Appendix A.5).

Figure 3b extends the check to all 19 cells with trajectories, each measured from 20 tokens to its own last cutoff. AUROC grows in 15 of them, by 6.72 points on average and up to 19.4, and three of the four declines are below one point. On Qwen3-14B/MATH-500, class-mean trajectories show

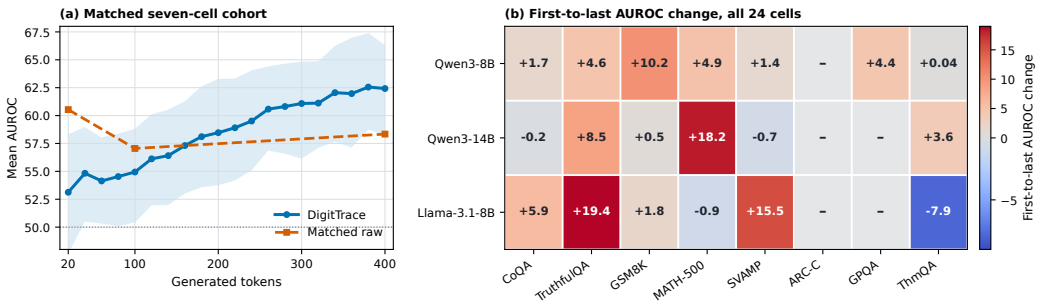

Figure 3: **DigitTrace AUROC rises as answers grow in most cells.** (a) Mean AUROC of DigitTrace and raw hidden states on the seven cells with raw snapshots. Shading spans the 10th–90th percentiles over 50 splits. (b) AUROC change from 20 tokens to the last cutoff. The two CoQA cells of Qwen3-14B and Llama-3.1-8B are replayed (Appendix F.11), and grey cells have answers too short for a trajectory. Appendix Figures 29–31 show every curve.

Table 2: **AUROC of every family combination.** Means over 24 cells and 50 splits, with regularization selected per subset and readout. Bold marks column maxima. The last row is the pre-specified primary trace.

Digits	Ref.	Mom.	Stats	Joint	Fact.
	✓		4	73.11	73.91
✓			9	74.04	75.63
		✓	3	**75.89**	76.16
	✓	✓	7	75.41	76.63
✓		✓	12	75.41	**76.92**
✓	✓		13	74.11	75.67
✓	✓	✓	16	75.41	76.68

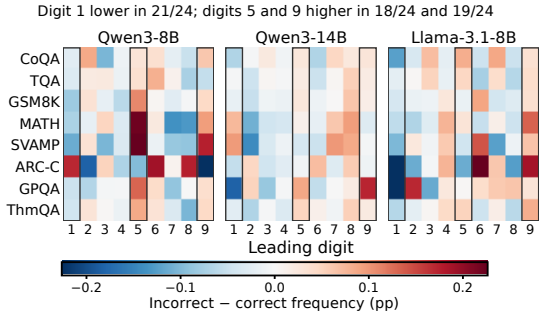

Figure 4: **Incorrect answers shift the digit profile.** Incorrect-minus-correct leading-digit frequency in every cell, averaged over streams and layers. Boxes mark digits 1, 5, and 9.

all three feature families separating further as generation proceeds, even while individual scores still overlap (Appendix Figures 20 and 27). This is the gain Section 3 predicts: each generated token adds observations to every view while the trace keeps its size, so in every view an answer's digit profile settles closer to its class mean and the two classes overlap less.

6.2 What the trace carries and how the readout uses it

Table 2 evaluates every combination of the three feature families under both readouts. Each family is predictive on its own, including the reference statistics.

A family helps most when it brings information the others lack. Under the factorized readout, digit frequencies raise moments-only AUROC by 0.75 points (95% CI: 0.10–1.49), and reference statistics raise it by 0.47 points. Added to the digit frequencies, reference statistics give no detectable gain: they are deterministic functions of those frequencies (Appendix B.4), so they add nothing measurable. They matter elsewhere, as the most stably label-correlated coordinates: on Qwen3-14B/MATH-500, the JS and KL divergences of the layer-36 norm-stream view rank in their family's top five in 49 and 47 of 50 splits, while no digit coordinate exceeds 21 (Appendix Figure 16).

The coordinates are also readable. Averaged over streams and layers, incorrect answers place less mass on digit 1 than correct answers in 21 of 24 cells, and more on digits 5 and 9 in 18 and 19 cells, while their total distance to $\mathbf{q}$ does not differ (Figure 4, Appendix Figure 13). Correctness thus shifts the shape of the digit profile, not its distance from $\mathbf{q}$. Within each view, however, the distance does differ by class: incorrect answers are farther from $\mathbf{q}$ in 45% of views and closer in the rest, and averaging cancels these opposite shifts. The trace keeps each view's distance as a separate coordinate, so the probe can weight it with either sign (Appendix F.8).

Two subsets come close to the complete 16-statistic trace (76.68): reference statistics plus moments reach 76.63 with seven statistics per view, and digits plus moments 76.92 with twelve, within 0.31 of MIND. We keep the pre-specified trace rather than a test-selected subset (Appendix F.1).

The readout also matters. Factorization adds 1.27 points over a single joint probe, positive in every dataset and model mean. Of these, 0.82 points come from averaging alone, measured with 30 full-trace probes fitted on bootstrap copies of the same training answers, and the other 0.45 from splitting the coordinates into disjoint groups (Appendix Tables 8 and 9).

6.3 Distributional and neuron-aligned representations are complementary

Complementarity tests whether the trace retains information a strong neuron-aligned detector does not use. Averaging development-standardized DigitTrace and MIND scores on TruthfulQA, GSM8K, and ARC-Challenge, for all three models and 50 splits, reaches 81.70 mean AUROC, versus 77.92 for DigitTrace and 79.00 for MIND. Fusion beats the stronger of the two in eight of nine cells, by 1.84 points on average, and the two scores correlate only moderately (Spearman 0.51), so they rank answers partly differently (Appendix Table 10). Averaged over models, fusion is highest on all three datasets, whichever constituent leads: 89.12 against 85.98 for MIND on GSM8K, 81.26 against 79.07 for DigitTrace on ARC-Challenge, and 74.72 against 73.84 for MIND on TruthfulQA.

6.4 Representation and evaluation controls

The trace accumulates over generated tokens, so answer length and formatting are candidate confounds. Style7, a probe on seven such counts, reaches 69.43 AUROC, and DigitTrace exceeds it by 7.25 points. A logistic regression on the two scores, fitted on validation data, improves on Style7 in all 24 cells but not on DigitTrace alone (-0.05), so Style7 adds no detection signal beyond the trace. Regressing the seven counts out of the DigitTrace score on development data leaves 66.04 AUROC, and across cells its AUROC is unrelated to answer length (Spearman -0.10, Appendix F.4).

DigitTrace trails full raw (78.75 AUROC) by 2.07 points at about 4% of its input. At comparable size it matches a random projection of those raw states (75.73, Appendix F.3), whose coordinates have no fixed meaning to read or compare across models. The comparison favors raw: an expanded regularization grid raises raw's mean by 2.41 points, with gains in 23 of 24 cells, whereas DigitTrace keeps its validation-selected C, though the best fixed C would add 0.32 (Appendix E).

Transfer separates coordinate meaning from the fitted boundary. Fitted on seven datasets of a model and tested on the eighth, DigitTrace reaches 65.70 mean AUROC, against 59.16 for Style7 and 63.36 for stored ICR-probe features under the same readout, and 63.19 for perplexity. It exceeds Style7 by 6.55 points (95% CI: 1.79–11.20) and ICR-probe features in 21 of 24 cells. On mathematical reasoning it keeps 79.20, six points below in-domain (Appendix Table 11). Under transfer the joint readout matches the factorized one ($+0.64$, interval crossing zero), so partition averaging refines in-domain fits but not transfer. A trace fitted on some tasks of a model therefore scores its others.

7 Conclusion and limitations

DigitTrace exposes correctness-related evidence without retaining neuron identities. It is strongest on mathematical reasoning, at 84.94 mean AUROC, and outperforms every other detector on GPQA-Diamond for all three models. The signal comes from the coordinates, not the readout: one partition reaches 76.50 in domain, and a joint probe matches the ensemble under transfer (Appendix H).

Like probe baselines, DigitTrace needs labeled in-domain answers. Score directions do not transfer across model families, and science multiple choice resists task transfer (Appendices F.13 and F.15). Gains during generation appear in 15 of 19 cells but vary by task. Collection takes 1.98 times the replay time on Qwen3-8B, but we also provide cheaper settings: reading only the final token adds almost no computation, and fewer statistics shrink the trace (Appendix F.7 and Section 6.2).

Despite these limits, the trace remains valuable: it needs only one generation, 24–27$\times$ fewer input coordinates than the sampled raw probe, and 17.74 times less storage. Correctness evidence thus need not come from neuron-aligned coordinates: a readable description of activation magnitudes carries most of it, adds to a strong neuron-aligned detector, and grows as the model generates.

REPRODUCIBILITY STATEMENT

We specify the feature construction, activation streams, layers, partition seeds, 50-split protocol, prompts, extraction, and labels. The appendix gives complete results and the judge protocol. Anonymized code is in the supplementary material.

ETHICS STATEMENT

These detectors can produce false confidence and should not determine high-stakes decisions alone. Benchmark ranking does not guarantee individual or out-of-domain behavior.

AI USE STATEMENT

Generative AI tools assisted us in implementing methods, refining and documenting the experimental setup, interpreting some results, and translation. They were not used to develop the theoretical model, formulate mathematical claims, construct proofs, or propose hypotheses. Generating synthetic data and qualitative or thematic analysis do not apply to this work. We also used these tools to help write and edit code, create and edit some figures and tables, search for and summarize literature, and draft parts of the appendix and polish the writing, including the wording of the authors' proofs. The authors reviewed all AI-assisted work: code was checked against the reference implementation and tested, reported numbers were recomputed from the stored results, and citations were checked against their sources. The authors take full responsibility for the content of this paper, including any text, claims, or artifacts produced with AI assistance.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7889–7901, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.489. URL https://aclanthology.org/2023.emnlp-main.489/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

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, 2021. URL https://arxiv.org/abs/2110.14168.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nfK0ZXFFSn.

Zuzanna Dubanowska, Maciej Żelaszczyk, Michał Brzozowski, Paolo Mandica, and Michal P. Karpowicz. Representation-based broad hallucination detectors fail to generalize out of distribution.

In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17563–17575, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.952. URL https://aclanthology.org/2025.findings-emnlp.952/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily

Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Theodore P. Hill. A statistical derivation of the significant-digit law. *Statistical Science*, 10:354–363, 1995. URL https://api.semanticscholar.org/CorpusID:1346106.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Xinyi Li, Zhen Fang, Yongxin Deng, Jinyuan Luo, Hongnan Ma, Changdae Oh, Zijing Shi, Shanshan Ye, Hanchen Wang, Shu-Lin Chen, Yadan Luo, Mengyue Yang, Sean Du, Sharon Li, and Ling Chen. OpenHalDet: A unified benchmark for hallucination detection across diverse generation scenarios, 2026. URL https://arxiv.org/abs/2606.06959.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Steven J. Miller and Mark J. Nigrini. The modulo 1 central limit theorem and Benford's law for products, 2007. URL https://arxiv.org/abs/math/0607686.

OpenAI. GPT-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large-Margin Classifiers*. The MIT Press, 09 2000. ISBN 9780262283977. doi: 10.7551/mitpress/1113.003.0008. URL https://doi.org/10.7551/mitpress/1113.003.0008.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019. doi: 10.1162/tacl_a_00266. URL https://aclanthology.org/Q19-1016/.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Surya Kant Sahu, Abhinav Java, Arshad Shaikh, and Yannic Kilcher. Rethinking neural networks with Benford's law, 2021. URL https://arxiv.org/abs/2102.03313.

Jiandong Shao, Yao Lu, and Jianfei Yang. Benford's curse: Tracing digit bias to numerical hallucination in LLMs. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 117546–117569. Curran Associates, Inc., 2025. doi: 10.52202/085713-3919. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/aa5f5e6eb6f613ec412f1d948dfa21a5-Paper-Conference.pdf.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised real-time hallucination detection based on the internal states of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14379–14391, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.854. URL https://aclanthology.org/2024.findings-acl.854/.

Ye Tian and Yang Feng. RaSE: Random subspace ensemble classification. *Journal of Machine Learning Research*, 22(45):1–93, 2021. URL http://jmlr.org/papers/v22/20-600.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. Prompt-guided internal states for hallucination detection of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21806–21818, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1058. URL https://aclanthology.org/2025.acl-long.1058/.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17986–18002, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.880. URL https://aclanthology.org/2025.acl-long.880/.

A Complete feature specification and pseudocode

A.1 Where the streams are read

Figure 5 marks where the collector reads each stream in a pre-norm decoder block. Each stream is read by a forward hook, a function attached to a module that records its input or output during the forward pass. The hooks record the input to self-attention after the input RMSNorm, the self-attention output before its residual addition, the MLP input after the attention residual and the post-attention RMSNorm, the MLP output before its residual addition, and the block output. The sixth stream, the norm stream, is the per-token L2 norm of the block output. It is computed from the block output rather than recorded by its own hook. Across tokens, it plays the role that the moments play within a view: it keeps the per-token scale, which the digits of the other streams discard. The embedding output is also recorded but is not part of the trace. Every fourth physical layer is sampled, starting from layer 0. This gives layers $0, 4, \ldots, 28$ for Llama-3.1-8B-Instruct (8 layers), $0, 4, \ldots, 32$ for Qwen3-8B (9 layers), and $0, 4, \ldots, 36$ for Qwen3-14B (10 layers).

A.2 Accumulator

A *view* v is one stream at one sampled layer. For each view, the online state holds nine digit counts $\mathrm{count}[v, 1{:}9]$ and four running sums over all values of the view: the value count $n[v]$, the signed sum $\mathrm{sum}[v]$, the squared sum $\mathrm{sq_sum}[v]$, and the absolute sum $\mathrm{abs_sum}[v]$. After each generated token, the stream tensor is detached, converted once to float32, and flattened. All of these float32

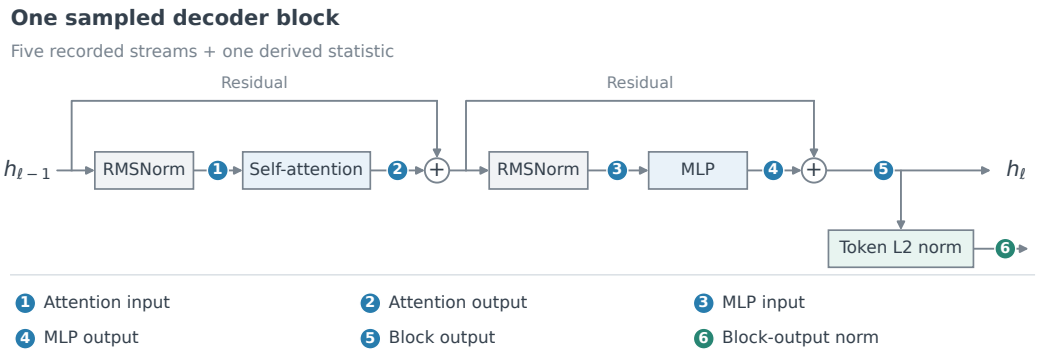

Figure 5: **Where each stream is read in one pre-norm decoder block.** Numbered markers show the six streams that DigitTrace summarizes at every sampled layer.

values update the four running sums, and the squares are accumulated in float64 to avoid overflow. For digit extraction, the same values are cast to float64. The only purpose of this cast is to keep rounding near decade boundaries from affecting $\log_{10}$ and the division by a power of ten. Non-finite values and exact zeros have no leading digit and are skipped. Across all collected answers, exact zeros make up about 0.05–0.09% of the values in the attention input, MLP input, and block output, and about 10^{-8} of the attention and MLP outputs. No non-finite activation reached the evaluated features. Any such value would have made the running sums non-finite, and every running sum is finite.

At the end of generation, each view contributes the 16 coordinates of Table 3.

Table 3: **The 16 coordinates of one view.** All quantities are computed from the accumulator above. Here $\mathbf{q}$, with $q_d = \log_{10}(1 + 1/d)$, is the Benford distribution, and $\mathbf{m} = (\mathbf{p} + \mathbf{q})/2$. Logarithms in JS and KL are natural, and $\mathbf{p}$ and $\mathbf{q}$ are clipped below at 10^{-12} inside the divergences.

Index	Symbol	Definition	Family
1–9	p_d	$\mathrm{count}[v, d] / \sum_{d'=1}^{9} \mathrm{count}[v, d']$	digit frequencies
10	MAD	$\frac{1}{9}\sum_{d=1}^{9} \lvert p_d - q_d \rvert$	reference statistics
11	JS	$\frac{1}{2}\sum_d p_d \log \frac{p_d}{m_d} + \frac{1}{2}\sum_d q_d \log \frac{q_d}{m_d}$	reference statistics
12	KL	$\sum_{d=1}^{9} p_d \log \frac{p_d}{q_d}$	reference statistics
13	ρ	Pearson correlation of $\mathbf{p}$ and $\mathbf{q}$ over the nine digits	reference statistics
14	μ	$\mathrm{sum}[v]/n[v]$	activation moments
15	σ	$\sqrt{\max(\mathrm{sq_sum}[v]/n[v] - \mu^2,\ 0)}$	activation moments
16	$\mathbb{E}\lvert a \rvert$	$\mathrm{abs_sum}[v]/n[v]$	activation moments

Any non-finite derived feature would be replaced by zero before standardization, but none occurs in the evaluated feature matrices. A view with no valid digit would likewise receive zero digit statistics, but every view of a layer the model has contains valid digits. The zero-filled columns of layers a model lacks are described in Appendix A.6.

A.3 Leading-digit extraction

The floor $\lfloor \log_{10} \lvert a \rvert \rfloor$ is the decade of a, and the remainder $r(a) = \log_{10} \lvert a \rvert - \lfloor \log_{10} \lvert a \rvert \rfloor$ is its position inside that decade, so $\lvert a \rvert = 10^{\lfloor \log_{10} \lvert a \rvert \rfloor} \cdot 10^{r(a)}$ with $10^{r(a)} \in [1, 10)$. The leading digit is the integer part of $10^{r(a)}$. Proposition 1 proves the equivalence used in Section 3 and shows that the nine bin widths equal the Benford probabilities q_d. Hence $\mathbf{q}$ is the digit distribution that a uniform $r(a)$ would produce, and the digit coordinates measure departures from it. For $a = 0.34$, 3.4, and 340, the logarithms are -0.469, 0.531, and 2.531, with floors -1, 0, and 2. All three therefore share $r(a) = 0.531$, which lies in $[\log_{10} 3, \log_{10} 4) = [0.477, 0.602)$, so all three give digit 3. For a value whose magnitude already lies in $[1, 10)$, the decade is 0 and $10^{r(a)} = \lvert a \rvert$: the -1.41 of Figure 1

gives $r(a) = 0.149$, which lies in $[0, \log_{10} 2)$, hence digit 1. Taking the absolute value discards the sign, and the moments carry the sign instead.

A.4 Worked example of one view

Take Qwen3-8B and an answer of 200 generated tokens. Qwen3-8B has hidden size 4096, and its nine sampled layers give $6 \times 9 = 54$ views. The MLP-output view at layer 8 receives $d_{h,\ell} = 4096$ values at every decoding step, so its value set $\mathcal{A}_{h,\ell}$ (Section 3) collects $200 \times 4096 = 819{,}200$ activation values, from which the view reports its 16 coordinates. The norm-stream view at the same layer receives one value per step, so its $\mathcal{A}_{h,\ell}$ collects 200 values and it reports 16 coordinates of the same form. Over all 54 views the trace summarizes $200 \times 9 \times (5 \times 4096 + 1)$, about 36.9 million activation values, into 864 coordinates. The accumulator behind them holds 13 running numbers per view, 702 in total, for any answer length. Doubling the answer doubles the values summarized and leaves the 702 running numbers and the 864 coordinates unchanged.

A.5 What the trace accumulates over

The trace accumulates over generated tokens only, and over the whole answer rather than a chosen position or window. Table 4 reports the three development grids behind these choices. In the collection-scope grid, adding prompt activations to the same counts lowered AUROC, and keeping them as separate coordinates gained only 0.4 points at twice the size, so the primary trace excludes them. A snapshot is the trace accumulated from the first generated token up to a given step. Accumulating over only the first 100 steps scored 7.40 points below the last available snapshot. Concatenating the last snapshot with the one at 100 tokens (early and late), or with those at 50 and 200 tokens (three snapshots), did not help, even at two or three times the dimension. Late tokens therefore carry evidence, and intermediate snapshots add nothing to the final one.

Restricting the trace to a single position would also change what can be estimated, not only how much evidence is gathered. At one decoding step, a recorded view still supplies $d_{h,\ell}$ values, so its digit histogram can still be estimated. The norm-stream view, however, supplies only one value, and its nine digit frequencies, four reference statistics, and three moments are then all functions of that single number: one occupied bin, a zero standard deviation, and a signed mean equal to the absolute mean. A position-restricted trace keeps one magnitude from that view and none of its distribution. In contrast, all four span variants hold the digit statistics fixed and vary only the span. The raw controls do not. Full raw averages the last five generated tokens at five middle layers (Appendix F), so it changes the representation as well as the span.

The third grid changes only the position, restricting the summary to the end of the answer, where neuron-aligned detectors usually read. With the coordinates and the readout fixed, summarizing only the final generated token loses 3.92 points on average over eight datasets of Qwen3-8B, and summarizing only the final five tokens loses 3.41. The accumulated trace scores higher than the final token on six of the eight datasets and higher than the final five on seven (Table 5). The two exceptions are MATH-500 and TheoremQA, where the final token leads before and after the length adjustment below. A position-restricted summary can encode answer length through positional information, which the accumulated trace does not store explicitly, and length alone is predictive on several datasets, reaching 85.4 AUROC on SVAMP. We fit the linear log-length component on the development split and regress it out of every score. The accumulated trace still leads the final token by 3.53 points. This grid used one model, 20 seeds, and stratified 60/20/20 splits, so it is a design check rather than a result in the main comparison. The end-of-answer arms come from teacher-forced replay of the frozen answers under the collection prompt. To validate the replay, we also replayed 24 MATH-500 answers in full. The replayed traces have a mean per-coordinate correlation of 0.987 with the stored traces, and their median digit-frequency difference is below 10^{-4}.

A.6 Padded layers

Models collected together share one feature table, whose columns cover the layers of the deepest model in it. Layers that a shallower model does not have appear as all-missing columns, and evaluation fills them with zero. These columns cover 96 coordinates for Qwen3-8B on SVAMP, ARC-Challenge, and TheoremQA, and 192 for Llama-3.1-8B-Instruct. For example, Qwen3-14B

Table 4: **What the trace accumulates over.** Three development grids, each with its own protocol and none under the frozen protocol, so values are descriptive and not comparable with Table 1. Collection scope: 24 cells, a 60-coordinate feature set, one stratified 60/20/20 split. Accumulation span: nine of the cells with trajectory snapshots, 20 seeds, stratified 60/20/20 splits, where d is 784 to 976 coordinates depending on the model. Readout position: eight datasets of Qwen3-8B, 20 seeds, stratified 60/20/20 splits, the primary 864-coordinate trace, with the end-of-answer arms replayed under the collection prompt. Bold marks the variant used.

Grid	Variant	Coordinates	Mean AUROC
Collection scope	Decoding only, as used	60	**74.4**
	Question prefill and decoding	60	71.7
	Full prefill and decoding	60	71.5
	Full prefill concatenated with decoding	120	74.8
	Three scope scores stacked	3	74.7
Accumulation span	Last available snapshot, as used	d	**73.59**
	Early and late snapshots concatenated	$2d$	73.38
	Three snapshots concatenated	$3d$	73.17
	First hundred steps only	d	66.19
Readout position	Whole answer, as used	864	**77.72**
	Final five tokens	864	74.31
	Final token only	864	73.80

Table 5: **Readout position by dataset.** Mean held-out AUROC on Qwen3-8B over 20 seeds with stratified 60/20/20 splits. All three arms share the splits, the readout, and the 864-coordinate feature set. They differ in the positions summarized, and the two end-of-answer arms also come from teacher-forced replay. *Length* scores answers by their token count alone. The last three columns regress out of each score the linear log-length component fitted on the development split. Bold marks the best of the three arms in each row. Development protocol, not comparable with Table 1.

			Summarized positions			Log-length regressed out		
Dataset	n	Length	Whole	Final 5	Final 1	Whole	Final 5	Final 1
CoQA	7199	62.9	**73.08**	67.20	71.93	70.00	62.60	68.74
TruthfulQA	817	52.9	**72.31**	62.66	66.41	72.18	62.23	66.15
GSM8K	1319	71.3	**85.50**	83.72	80.37	79.75	78.88	73.84
MATH-500	500	73.9	79.00	76.88	**82.47**	63.59	61.69	67.44
SVAMP	1000	85.4	**90.18**	85.46	88.59	67.48	72.36	69.70
ARC-Challenge	1172	66.1	**78.20**	74.82	70.06	72.77	68.43	62.16
GPQA-Diamond	198	53.4	**72.25**	69.70	56.33	70.59	70.06	55.42
TheoremQA	747	70.1	71.25	74.01	**74.26**	58.18	64.67	62.88
Mean		67.0	**77.72**	74.31	73.80	**69.32**	67.61	65.79

has ten sampled layers and Qwen3-8B nine, so in a table shared with Qwen3-14B, Qwen3-8B lacks layer 36 and has $6 \times 16 = 96$ all-missing columns. Llama-3.1-8B-Instruct, with eight sampled layers, lacks layers 32 and 36 in the same way. These constant columns carry no information and do not change the predictions of any single probe. However, the random partitions of the factorized readout were drawn over all columns, padded ones included, so each group also contains some padded columns. Padded columns are excluded from effective-dimension and storage counts.

A.7 Method-figure activation example

The values in Figure 1(2) come from the block output at physical layer 20 for the last generated token of Qwen3-8B/GSM8K sample 0. They were obtained by teacher-forced replay of the cached answer. The twelve values were chosen, without looking at labels, to give a readable count. They illustrate only how values map to leading digits.

A.8 PSEUDOCODE

Algorithm 1 builds the trace online with the accumulator above, and Algorithm 2 scores it with the factorized readout of Section 3. Both use the variable names defined in this appendix.

Algorithm 1: Online trace construction.

```
Input  generated answer y, views V
Output trace f

for each view v in V do
    count[v,1:9] <- 0
    n[v], sum[v], sq_sum[v], abs_sum[v] <- 0
end
for each generated token t observed during decoding do
    for each view v in V do
        a <- flattened float32 values of view v at token t
        n[v] <- n[v] + number of values in a
        sum[v] <- sum[v] + sum of a
        sq_sum[v] <- sq_sum[v] + sum of a^2
        abs_sum[v] <- abs_sum[v] + sum of |a|
        for each finite nonzero value a_j in a do
            d <- leading digit of |a_j|
            count[v,d] <- count[v,d] + 1
        end
    end
end
for each view v in V do
    p <- count[v,1:9] / sum of count[v,1:9]
    mu <- sum[v] / n[v]
    sigma <- sqrt(max(sq_sum[v] / n[v] - mu^2, 0))
    m_abs <- abs_sum[v] / n[v]
    f[v] <- (p, MAD, JS, KL, rho, mu, sigma, m_abs)
end
f <- concatenate f[v] over V and set non-finite entries to 0
return f
```

The accumulator size does not depend on answer length, and no per-token activation vector is kept. The collector records generated tokens fed back to the model: EOS termination includes all preceding answer tokens but not EOS, whereas length-limit termination excludes the final emitted answer token.

Each probe in Algorithm 2 standardizes its input columns and fits a class-balanced L1-regularized logistic regression. Every z-score uses the mean and standard deviation of the same probe's scores on the rows it was fitted on. The algorithm runs once for each cell and each outer split. Its development rows combine the outer training and validation rows. The readout splits these rows again, stratified 75/25, into fitting rows and validation rows, and uses the validation rows to select C.

Algorithm 2: Factorized readout for one cell and one outer split.

```
Input  development rows D, test rows E, trace columns 1,...,m
       C grid {0.03, 0.1, 0.3, 1.0}, seeds s_1,...,s_K with K = 5
Output test scores

split D into fitting rows F and validation rows W (75/25)
for each partition k = 1,...,K do
    G[k,1:6] <- 6 near-equal parts of a seed-s_k permutation of 1,...,m
    for each C in the grid do
        for each group g = 1,...,6 do
            fit a probe with C on rows F, columns G[k,g]
            z[g] <- scores on W, z-scored by the scores on F
        end
        auc[C] <- AUROC on W of the mean of z[1:6]
    end
```

```
    C_k <- C with the largest auc[C]
    for each group g = 1,...,6 do
        refit the probe with C_k on rows D, columns G[k,g]
        z[g] <- scores on E, z-scored by the scores on D
    end
    S[k] <- mean of z[1:6]
end
return mean of S[1:K]
```

B PROOFS AND THEORY SCOPE

This appendix gives the formal results behind the design choices of Section 3, in four parts. Appendix B.1 shows why leading digits suit the trace: they bin the fractional log magnitude, and digit counts do not change when neurons are reordered or magnitudes are rescaled by integer powers of ten (Propositions 1–3). Appendix B.2 shows that the Bayes rule is linear in digit counts under a class-conditional multinomial model (Proposition 4). Appendix B.3 shows that more digit observations make a fixed score more separable under the same model (Propositions 5–6). Appendix B.4 states that building the trace cannot add label information (Proposition 7) and explains why reference statistics still help a linear probe. Propositions 1–3 hold in exact arithmetic. Propositions 4–6 hold under a simple statistical model that motivates the design, and Sections 5 and 6 test the design on real activations. We follow the notation of Section 3: for one view, $\mathbf{n}$ holds the nine digit counts, $N = \sum_d n_d$ is the number of valid digit observations, $\mathbf{p} = \mathbf{n}/N$ is the digit profile, and $\pi_{z,d}$ is the probability of digit d under class z.

B.1 WHY LEADING DIGITS

Leading digits are a fixed discretization of log magnitude: the nine bins partition the fractional part of $\log_{10}|a|$ at the points $\log_{10} d$. This choice gives the trace properties that quantile bins and uniform bins on $\log|a|$ lack. Its bin edges are identical for every stream, layer, and model and need no calibration data, whereas quantile bins give each view its own edges and uniform bins on $\log|a|$ need a range that varies across views and models. Folding by decade makes the counts invariant to rescaling by powers of ten (Proposition 3), and Benford's law gives the bins an analytic reference, the Benford distribution $\mathbf{q}$ used by the reference statistics. The coordinates also carry evidence beyond the moments: digit frequencies raise moments-only AUROC by 0.75 points, and incorrect answers shift mass between specific digits (Section 6.2).

Proposition 1: Fractional-log bins. For finite $a \neq 0$, let $r(a) = \log_{10}|a| - \lfloor \log_{10}|a| \rfloor$. Then

$$D(a) = d \quad \Longleftrightarrow \quad r(a) \in [\log_{10} d, \log_{10}(d+1)).$$

Proof. Since $r(a) \in [0, 1)$, the significand $10^{r(a)}$ lies in $[1, 10)$. The leading digit is d exactly when $d \leq 10^{r(a)} < d + 1$. Applying the increasing function $\log_{10}$ gives the stated interval. If $r(a)$ is uniform on $[0, 1)$, the interval length is $\log_{10}(1 + 1/d)$, which is the Benford probability. For example, 0.43, 4.3, and 430 share $r(a) = \log_{10} 4.3 \approx 0.633$, which lies in $[\log_{10} 4, \log_{10} 5)$, so all three fall in the bin of digit 4.

Proposition 2: Coordinate-permutation invariance. For finite nonzero values $a_1, \ldots, a_M$, let $n_d(a_1, \ldots, a_M) = \sum_i \mathbf{1}\{D(a_i) = d\}$. For any permutation σ of $\{1, \ldots, M\}$,

$$n_d(a_{\sigma(1)}, \ldots, a_{\sigma(M)}) = \sum_i \mathbf{1}\{D(a_{\sigma(i)}) = d\} = n_d(a_1, \ldots, a_M).$$

The middle equality follows because σ is a bijection. Every deterministic function of the count or frequency vector inherits this invariance. The activation moments also depend only on symmetric sums, $\sum_i a_i$, $\sum_i a_i^2$, and $\sum_i |a_i|$, and the number of values. They are therefore permutation invariant, including when finite zero values are present. Concatenation preserves this invariance within each view, not under exchanges of layer or stream identities. In DigitTrace terms, a view's statistics do not depend on which neuron holds which value, so each coordinate is indexed by a statistic, a stream, and a layer rather than by a neuron. This gives different views and models a common way

to describe activations. Whether a fitted boundary transfers across tasks is an empirical question (Section 6.4).

Proposition 3: Integer-decade scale invariance. For finite nonzero a and any integer m, in exact arithmetic,

$$\log_{10}|10^m a| = m + \log_{10}|a|.$$

Adding an integer leaves the fractional part unchanged, so $D(10^m a) = D(a)$. Applying this pointwise leaves every digit count and its reference statistics unchanged. The statement does not extend to arbitrary positive scale factors because a non-integer log shift can cross a bin boundary. If every value of a view is scaled by 10^m, the nine frequencies and four reference statistics stay fixed while the three moments record the change, which is why the trace keeps both families (Section 3). Floating-point rounding, underflow, and overflow are outside this exact-arithmetic statement.

Why a Benford reference. For a product $Y = \prod_{i=1}^{m} X_i$ of finite nonzero factors, fractional logarithms add modulo one. If the log magnitudes $V_i = \log_{10}|X_i|$ are independent and identically distributed with $|\mathbb{E}e^{2\pi i k V_1}| < 1$ for every nonzero integer k, the kth Fourier coefficient of the sum is $(\mathbb{E}e^{2\pi i k V_1})^m$, which tends to zero, so the fractional log of Y converges to the uniform law and its digits to the Benford probabilities (Miller & Nigrini, 2007). This motivates logarithmic coordinates and a Benford reference. The trace uses the Benford distribution only as a fixed reference, so it does not require activations to arise this way.

B.2 Why a linear readout

Proposition 4 shows that, under a simple count model, the optimal score is linear in the digit counts, which motivates the logistic probes of Section 3.

Proposition 4: A linear Bayes rule for digit counts. Assume a fixed total count $N \geq 1$ and positive class priors $P(z=0), P(z=1)$. Under class z, let $\mathbf{n} \sim \mathrm{Multinomial}(N, \boldsymbol{\pi}_z)$ with $\pi_{z,d} > 0$ and $\sum_d \pi_{z,d} = 1$. This is the count model obtained from conditionally independent digits sharing the class probabilities. Neither probability vector is required to equal the Benford distribution. Then

$$P(\mathbf{n} \mid z) = \frac{N!}{\prod_d n_d!} \prod_{d=1}^{9} \pi_{z,d}^{n_d}.$$

Taking posterior log odds gives

$$\log \frac{P(z=1 \mid \mathbf{n})}{P(z=0 \mid \mathbf{n})} = \log \frac{P(z=1)}{P(z=0)} + \sum_d n_d \log \frac{\pi_{1,d}}{\pi_{0,d}},$$

because the multinomial coefficient is identical for both classes. Under equal misclassification costs, comparing these log odds with zero gives the Bayes decision rule, whose boundary is linear in the count vector. If $\boldsymbol{\pi}_0 = \boldsymbol{\pi}_1$, every count coefficient is zero and the posterior equals the prior. For fixed N, writing $n_d = N p_d$ also makes the log odds linear in frequencies. If counts vary, the factor N and any count-dependent prior must be accounted for, and a single frequency-only linear rule does not follow in general. In this model, a digit that is more frequent in incorrect answers receives a positive coefficient. Figure 4 shows class differences of this kind after averaging over views, with digit 1 less frequent and digits 5 and 9 more frequent in incorrect answers. Because the frequencies sum to one, adding a constant to every coefficient shifts the score by a constant, so such signs indicate a direction rather than identify the weights of a fitted probe. Activations are dependent and the trace combines heterogeneous views, so the proposition motivates the linear readout.

B.3 Why accumulate over the answer

Propositions 5 and 6 explain why the trace grows more informative with the answer: each token adds digit observations to every view, the digit profile concentrates around its class mean, and a fixed linear score separates the classes better.

Proposition 5: Histogram concentration under a conditional multinomial model. For $N \geq 1$, let $\mathbf{n} \sim \text{Multinomial}(N, \boldsymbol{\pi}_z)$ and $\mathbf{p} = \mathbf{n}/N$, with fixed class probabilities. Standard multinomial moments give

$$\mathbb{E}[n_d \mid z] = N\pi_{z,d}, \quad \text{Var}(n_d \mid z) = N\pi_{z,d}(1-\pi_{z,d}), \quad \text{Cov}(n_d, n_j \mid z) = -N\pi_{z,d}\pi_{z,j} \quad (d \neq j).$$

Dividing the count covariance by N^2 gives

$$\mathbb{E}[\mathbf{p} \mid z] = \boldsymbol{\pi}_z, \qquad \text{Cov}(\mathbf{p} \mid z) = \frac{\text{diag}(\boldsymbol{\pi}_z) - \boldsymbol{\pi}_z\boldsymbol{\pi}_z^\top}{N}.$$

Thus each marginal standard deviation is $O(N^{-1/2})$. Here N counts valid digit observations of one view, not generated tokens. A recorded view of Qwen3-8B receives 4,096 values per step, so a 200-token answer supplies up to 819,200 candidate values before zeros and non-finite entries are excluded, whereas the norm-stream view receives up to 200 (Appendix A.4). Because these values are dependent, the $N^{-1/2}$ rate is a guide from the independent case, and Section 6 measures the growth on real activations.

Proposition 6: Approximate AUC growth for a fixed linear histogram score. Let $\mathbf{p}_z = \mathbf{n}_z/N$ be independent class-conditional histograms under Proposition 5, and let $S_z = \mathbf{w}^\top \mathbf{p}_z$ for a fixed weight vector $\mathbf{w}$. Then

$$\mathbb{E}[S_z] = m_z = \mathbf{w}^\top \boldsymbol{\pi}_z, \qquad \text{Var}(S_z) = \frac{v_z}{N},$$

where $v_z = \mathbf{w}^\top(\text{diag}(\boldsymbol{\pi}_z) - \boldsymbol{\pi}_z\boldsymbol{\pi}_z^\top)\mathbf{w}$. Assume $v_0 + v_1 > 0$. A normal approximation to $S_1 - S_0$ has mean $m_1 - m_0$ and variance $(v_0 + v_1)/N$. The tie-aware definition is $\text{AUC} = P(S_1 > S_0) + \frac{1}{2}P(S_1 = S_0)$. Under the continuous normal approximation, ties have probability zero, giving

$$\text{AUC} \approx \Phi\left(\frac{\sqrt{N}(m_1 - m_0)}{\sqrt{v_0 + v_1}}\right),$$

where Φ is the standard normal cumulative distribution function. For $m_1 > m_0$, the approximation rises toward one as N grows: with more digit observations, the same score separates the two classes better. The model idealizes fixed class probabilities and independent histograms, and Figure 3 tests the prediction on real, dependent activations: DigitTrace AUROC grows with the answer prefix in 15 of 19 cells.

B.4 What the trace can keep

Proposition 7 bounds what any summary of the activations can carry, and the last paragraph applies the same argument to the reference statistics.

Proposition 7: Information cannot increase under trace construction. We write $I(\cdot;\cdot)$ for mutual information. Let $\mathcal{A}$ denote the raw activation values of all views of an answer, and let $\Psi(\mathcal{A})$ be the trace built from them. Because Ψ is deterministic, $z \to \mathcal{A} \to \Psi(\mathcal{A})$ is a Markov chain, and the data-processing inequality gives $I(z; \Psi(\mathcal{A})) \leq I(z; \mathcal{A})$. The trace can therefore only preserve label information, and the design goal is to preserve useful predictive information at a small fraction of the raw size. The experiments measure this predictive retention, not information-theoretic equality.

Why add reference statistics. Reference statistics are computed from the digit frequencies, so one may ask what they add. They add no label information, yet they can still help a linear probe, and this explains both findings in Section 6.2. For the fixed Benford distribution $\mathbf{q}$, let $\mathbf{r} = g(\mathbf{p}; \mathbf{q})$ contain the four reference statistics. Since $\mathbf{r}$ is determined by $\mathbf{p}$, the chain rule gives

$$I(z; \mathbf{p}, \mathbf{r}) = I(z; \mathbf{p}) + I(z; \mathbf{r} \mid \mathbf{p}) = I(z; \mathbf{p}),$$

so adding $\mathbf{r}$ to $\mathbf{p}$ adds no label information. A linear probe, however, can only weight $\mathbf{p}$ linearly. With the reference statistics, the same probe becomes

$$s_1(\mathbf{p}) = \mathbf{w}^\top \mathbf{p} + c, \qquad s_2(\mathbf{p}) = \mathbf{w}^\top \mathbf{p} + \mathbf{v}^\top g(\mathbf{p}; \mathbf{q}) + c,$$

where $\mathbf{w}$ and $\mathbf{v}$ are weight vectors and c is an intercept. The second form contains the first ($\mathbf{v} = 0$) and adds nonlinear summaries of the digit profile, such as its divergence from $\mathbf{q}$. Both findings of Section 6.2 follow: under the factorized readout, adding reference statistics to the digit frequencies changes mean AUROC by only 0.04 points (Table 2), yet the case study finds them the most stably label-correlated coordinates.

C DATA, PROMPTS, LABELS, AND EXAMPLES

TruthfulQA is evaluated as open-ended generation, not its multiple-choice variant.

Chat construction. Every request contains the system message *You are a helpful, accurate, and honest AI assistant.* followed by one user message. Each request is formatted with the model's own chat template, which wraps the messages in the model's role markers and ends with the marker that opens the assistant's reply. Qwen non-thinking runs set `enable_thinking=False`. Decoding is greedy with `do_sample=False`.

Task templates. The boxes below preserve the literal instruction order. Bracketed text denotes a dataset field.

MATH-500

```
Solve the following math problem.  Put your final answer in
\boxed{}.
[problem]
```

GSM8K

```
Solve the following math problem.  Put your final answer in
\boxed{}.
[question]
```

SVAMP

```
Solve the following math problem.  Put your final answer in
\boxed{}.
[Body] [Question]
```

TheoremQA

```
Answer the following question.  Put your final answer in \boxed{}.
[question]
```

ARC-Challenge

```
Question:  [question]
Options:
A. [option A]
B. [option B]
C. [option C]
D. [option D]
Please strictly answer with the correct option letter(s).
```

CoQA

```
Context:
[story]
Question: [question]
```

TruthfulQA

```
Question: [question]
```

GPQA-Diamond, Qwen3-8B

```
Instruction: Answer the graduate-level science multiple-choice
question. Select the correct option letter.
Question: [question]
A. [option A]
B. [option B]
C. [option C]
D. [option D]
```

GPQA-Diamond, Qwen3-14B and Llama-3.1-8B-Instruct

```
Question: [question]
Options:
A. [option A]
B. [option B]
C. [option C]
D. [option D]
Please strictly answer with the correct option letter(s).
```

CoQA asks several questions about each story, and we keep only the first question of each story. The CoQA, TruthfulQA, and second GPQA-Diamond prompts are built by OpenHalDet's prompt template, which fills in the context, question, and options of each item from OpenHalDet's standard data format. OpenHalDet does not include GPQA-Diamond, so we converted GPQA-Diamond into this format. Qwen3-8B's GPQA-Diamond generations come from an earlier collection that used the instruction-style prompt in the first box. This prompt elicits full reasoning (median 894 tokens), whereas the strict-letter prompt elicits a bare letter (median 13 tokens for Qwen3-14B and 2 for Llama-3.1-8B-Instruct). Appendix F.12 describes this difference. Within each cell, all detectors share the same generations and labels.

Generation caps are 2,048 tokens for CoQA, TruthfulQA, GSM8K, SVAMP, and ARC-Challenge; 8,192 for GPQA-Diamond and MATH-500; and 4,096 for TheoremQA. Qwen3-8B answers on CoQA, TruthfulQA, GSM8K, MATH-500, and GPQA-Diamond come from earlier collections with a 32,768-token cap. Under that cap, 46 of these 10,033 answers run past the caps above (17 on GPQA-Diamond, 22 on MATH-500, five on TruthfulQA, and two on GSM8K). All detectors in these cells use the same generations. TheoremQA drops 53 items that contain pictures and keeps 747 text-only questions. Before detector splitting, we merge SVAMP's public train and test partitions into one pool of 1,000 items.

Extraction and labels. Multiple-choice extraction prefers an explicit final-answer pattern, then the last standalone A–D token. Mathematical extraction prefers a boxed answer, then an explicit final answer, then the last numeric expression. Normalized exact matching is followed by numeric and limited symbolic checks. TheoremQA answers come with a type. Booleans and option letters must

match the reference. A number, or each number of a list, counts as correct when it lies within 1% of the reference or within 10^{-4} of it, whichever is larger, because TheoremQA rounds its references.

For CoQA and TruthfulQA, we run the judge released with OpenHalDet (Li et al., 2026) without modification: GPT-4o-mini (OpenAI, 2024) at temperature zero receives the question, optional context, acceptable truths, known traps when available, and the model answer. Its system instruction casts the judge as an expert fact checker and asks it to return `correct`, `hallucination`, or `abstention`. Matching an acceptable truth is correct, matching a known trap or adding fabricated facts is hallucination, and an explicit refusal is abstention. Three fixed demonstrations precede the evaluated item. We map both hallucination and abstention to $z = 1$. This judge labeled all 24,048 free-form answers of the three models. Because OpenHalDet labels its own benchmark with this same judge, our labels follow its published annotation protocol, and we therefore did not add a separate human verification step. The judge is identified only by its model alias.

Deterministic examples. For each of the eight datasets and each class, we select the Qwen3-8B non-thinking answer whose generated length is closest to the within-class median. Ties are resolved by sample ID. Model outputs are reproduced verbatim, including their Markdown and LaTeX syntax. Long text is truncated only in the typeset display. Generation used the task-specific caps above. Each box lists the prompt, the generated answer, the prediction extracted from it, the reference answer, and the evaluation. In the evaluation field, label = 1 marks an incorrect answer, the positive detector class. The parenthesis names the scorer: rule-based answer matching or the GPT-4o-mini judge described above.

CORRECT ARC-Challenge — `arc_challenge::1056`

PROMPT Question: Which type of cell is connected to muscle fibers to facilitate control of the muscle? Options: A. nervous B. skeletal C. circulatory D. respiratory Please strictly answer with the correct option letter(s).

GENERATED ANSWER A. nervous

PREDICTION A

REFERENCE A

EVALUATION label = 0 (rule-based answer matching)

INCORRECT ARC-Challenge — `arc_challenge::204`

PROMPT Question: A star produces energy through a process called nuclear fusion. Which aspect of a star enables the process of fusion to occur? Options: A. rotational speed B. apparent brightness C. surface temperature D. chemical composition Please strictly answer with the correct option letter(s).

GENERATED ANSWER C. surface temperature

PREDICTION C

REFERENCE D

EVALUATION label = 1 (rule-based answer matching)

CORRECT CoQA — coqa_train_000545

PROMPT Context: CHAPTER XII BOBBY COON AND REDDY FOX PLAY TRICKS It was night. All the little stars were looking down and twinkling and twinkling. Mother Moon was doing her best to make the Green Meadows as light as Mr. Sun did in the daytime. All the little birds except Hooty the Owl and Boomer the Night Hawk, and noisy Mr. Whip-poor-will were fast asleep in their little nests. Old Mother West Wind's Merry Little Breezes had all gone to sleep, too. It was oh so still! Indeed it was so very still that Bobby Coon, comin... [display truncated]

GENERATED ANSWER Bobby Coon thinks that being nocturnal is better. He says, "Night's the best time to be about," showing his preference for being active at night.

PREDICTION Bobby Coon thinks that being nocturnal is better. He says, "Night's the best time to be about," showing his preference for being active at night.

REFERENCE Bobby Coon

EVALUATION label = 0 (GPT-4o-mini judge)

INCORRECT CoQA — coqa_train_000344

PROMPT Context: CHAPTER I. A DISPERSION 'A telegram! Make haste and open it, Jane; they always make me so nervous! I believe that is the reason Reginald always _will_ telegraph when he is coming,' said Miss Adeline Mohun, a very pretty, well preserved, though delicate-looking lady of some age about forty, as her elder sister, brisk and lively and some years older, came into the room. 'No, it is not Reggie. It is from Lily. Poor Lily! Jasper— accident—Come.' 'Poor dear Lily! Is it young Jasper or old Jasper, I wonde... [display truncated]

GENERATED ANSWER The person sending postcards was **Jane Mohun**. In the passage, it is mentioned that Jane was "dashing off postcards at her davenport" while her sister, Miss Adeline Mohun, was speaking. This clearly indicates that Jane was the one sending the postcards.

PREDICTION The person sending postcards was **Jane Mohun**. In the passage, it is mentioned that Jane was "dashing off postcards at her davenport" while her sister, Miss Adeline Mohun, was... [display truncated]

REFERENCE Lily

EVALUATION label = 1 (GPT-4o-mini judge)

CORRECT GPQA-Diamond — gpqa_diamond_test_000109

PROMPT Instruction: Answer the graduate-level science multiple-choice question. Select the correct option letter. Question: Determine the product obtained when 3,4-dimethylhexanedial is reacted with the following set of reagents: 1. KOH, H2O, THF, Heat 2. CH3CH2MgBr, H3O+ 3. PCC, CH2Cl2 4. O3, H2O A. 3,4-dimethyl-5,6-dioxooctanoic acid B. 4,5-dimethylnonane-2,6,7-trione C. 3,4-dimethyl-5,6-dioxooctanal D. 4,5-dimethylnonane-2,6,7-trione

CORRECT (continued) GPQA-Diamond — `gpqa_diamond_test_000109`

GENERATED ANSWER To determine the product of the reaction of **3,4-dimethylhexanedial** with the given reagents, we need to analyze the sequence of reactions step by step. — ### **Step 1: 3,4-dimethylhexanedial + KOH, H?O, THF, Heat** This is a **base-catalyzed oxidation** of a **dialdehyde** (a compound with two aldehyde groups). In basic conditions, dialdehydes undergo **oxidative cleavage** to form **carboxylic acids**. **3,4-dimethylhexanedial** has the structure: “‘ CH?(C(CH?)?)CH?CH?CH?CH?CHO “‘ Wait ? actually, the correct structure is: “‘ CHO-CH(CH?)-CH(CH?)-CH?-CH?-CHO “‘ So, it’s a **hexane** chain with **two a... [display truncated]

PREDICTION A

REFERENCE A

EVALUATION label = 0 (rule-based answer matching)

INCORRECT GPQA-Diamond — `gpqa_diamond_test_000115`

PROMPT Instruction: Answer the graduate-level science multiple-choice question. Select the correct option letter. Question: A student performed an acid-catalyzed reaction of 3-methylpyrrolidine with compound A, the final product of the reaction is 1-(cyclohexylidenemethyl)-3-methylpyrrolidine. Select the suitable reagent (A) and catalyst (B) for the reaction. 3-methylpyrrolidine + A (B, Heat, solvent) —¿ 1-(cyclohexylidenemethyl)-3-methylpyrrolidine A. A = vinylcyclohexane, B = Acetic acid B. A = cyclohexanecarbaldeh... [display truncated]

GENERATED ANSWER To determine the correct reagents and catalyst for the reaction: **3-methylpyrrolidine + A (B, Heat, solvent) ? 1-(cyclohexylidenemethyl)-3-methylpyrrolidine** we need to analyze the transformation and identify the type of reaction. — ### **Reaction Analysis:** The product **1-(cyclohexylidenemethyl)-3-methylpyrrolidine** suggests that a **cyclohexylidene group** (i.e., a **cyclohexenyl group**) has been added to the pyrrolidine ring. This is a **Michael addition** or **aldol-type reaction**, where a **carbonyl compound** (like an aldehyde or ketone) reacts with a **nucleophilic pyrrolidine** in the presence... [display truncated]

PREDICTION B

REFERENCE D

EVALUATION label = 1 (rule-based answer matching)

CORRECT GSM8K — `gsm8k::315`

PROMPT Solve the following math problem. Put your final answer in \boxed{}. Martha is planning her Christmas party. She invited 2 families with 6 people and 3 families with 4 people. 8 people couldn’t come due to illness, and 1/4 that number had previous commitments. How many people show up for Martha’s party?

CORRECT (continued) GSM8K — gsm8k::315

GENERATED ANSWER Let's break down the problem step by step. — ### Step 1: Calculate the total number of people invited - 2 families with 6 people each: $ 2 \times 6 = 12 $ people - 3 families with 4 people each: $ 3 \times 4 = 12 $ people **Total invited:** $ 12 + 12 = 24 $ people — ### Step 2: Determine how many people couldn't come - 8 people couldn't come due to illness. - $ \frac{1}{4} $ of that number had previous commitments: $ \frac{1}{4} \times 8 = 2 $ people **Total people who couldn't come:** $ 8 + 2 = 10 $ people — ### Step 3: Calculate how many people showed up **Total people who showed up:** $ 24 - 10 = 14 $... [display truncated]

PREDICTION 14

REFERENCE 14

EVALUATION label = 0 (rule-based answer matching)

INCORRECT GSM8K — gsm8k::209

PROMPT Solve the following math problem. Put your final answer in \boxed{}. Twenty dozen cups cost $1200 less than the total cost of half a dozen plates sold at $6000 each. Calculate the total cost of buying each cup.

GENERATED ANSWER We are given: - **Twenty dozen cups** cost **$1200 less** than the total cost of **half a dozen plates** sold at **$6000 each**. - We are to find the **total cost of buying each cup**. — ### Step 1: Understand the quantities - **Twenty dozen cups** = $ 20 \times 12 = 240 $ cups - **Half a dozen plates** = $ \frac{1}{2} \times 6 = 3 $ plates — ### Step 2: Find the total cost of the plates Each plate costs **$6000**, so: $$ \text{Total cost of plates} = 3 \times 6000 = 18000 $$ — ### Step 3: Use the given relationship The cost of **240 cups** is **$1200 less** than the cost of the plates: $$ \text{Cost of... [display truncated]

PREDICTION 70

REFERENCE 145

EVALUATION label = 1 (rule-based answer matching)

CORRECT MATH-500 — math_500::350

PROMPT Solve the following math problem. Put your final answer in \boxed{}. Find the units digit of $18^6.$

GENERATED ANSWER To find the **units digit** of $ 18^6 $, we only need to consider the **units digit of the base**, which is **8**, and find the pattern in the **units digits** of powers of 8. Let's compute the powers of 8 and observe the pattern in the **units digits**: - $ 8^1 = 8 $? units digit = **8** - $ 8^2 = 64 $? units digit = **4** - $ 8^3 = 512 $? units digit = **2** - $ 8^4 = 4096 $? units digit = **6** - $ 8^5 = 32768 $? units digit = **8** - $ 8^6 = 262144 $? units digit = **4** We can see the **pattern of units digits** for powers of 8 is: $$ 8, 4, 2, 6, 8, 4, 2, 6, \dots $$ This is a repeating cycle of **4... [display truncated]

PREDICTION 4

REFERENCE 4

EVALUATION label = 0 (rule-based answer matching)

INCORRECT MATH-500 — `math_500::327`

PROMPT Solve the following math problem. Put your final answer in \boxed{}. Let $P(x)$ be a monic polynomial of degree 3. Suppose that $P(x)$ has remainder $R(x)$ when it is divided by $(x - 1)(x - 4),$ and remainder $2R(x)$ when it is divided by $(x - 2)(x - 3).$ Given that $P(0) = 5,$ find $P(5).$

GENERATED ANSWER We are given: - $ P(x) $ is a **monic polynomial of degree 3**. - When $ P(x) $ is divided by $ (x - 1)(x - 4) $, the **remainder is** $ R(x) $. - When $ P(x) $ is divided by $ (x - 2)(x - 3) $, the **remainder is** $ 2R(x) $. - $ P(0) = 5 $. - We are to find $ P(5) $. — ### Step 1: Use the **Remainder Theorem** and **Polynomial Division** Let?s denote: - $ P(x) = (x - 1)(x - 4)Q_1(x) + R(x) $, where $ R(x) $ is a **linear polynomial** (degree ? 1), since the divisor is degree 2. - $ P(x) = (x - 2)(x - 3)Q_2(x) + 2R(x) $, again with $ R(x) $ being a linear polynomial. Let?s write $ R(x) = ax + b $, where $ a... [display truncated]

PREDICTION 6

REFERENCE 15

EVALUATION label = 1 (rule-based answer matching)

CORRECT SVAMP — `svamp::104`

PROMPT Solve the following math problem. Put your final answer in \boxed{}. Every day Ryan spends 4 hours on learning english and 6 hours on learning chinese. If he learns for 86 days How many hours does he spend on learning english and chinese each day?

GENERATED ANSWER We are told that Ryan spends: - **4 hours per day** learning English - **6 hours per day** learning Chinese So, the total number of hours he spends learning **both languages each day** is: $$ 4 + 6 = 10 \text{ hours per day} $$ This is the total time spent on both languages **each day**, regardless of how many days he learns. ### Final Answer: $$ \boxed{10} \text{ hours per day} $$

PREDICTION 10

REFERENCE 10

EVALUATION label = 0 (rule-based answer matching)

INCORRECT SVAMP — `svamp::546`

PROMPT Solve the following math problem. Put your final answer in \boxed{}. After eating a hearty meal they went to see the Buckingham palace. There, Rachel learned that 703 visitors came to the Buckingham palace on the previous day. If there were 246 visitors on that day How many visited the Buckingham palace within 25 days?

GENERATED ANSWER We are given: - On the **previous day**, **703 visitors** came to Buckingham Palace. - On **that day**, **246 visitors** came. - We are asked to find the total number of visitors over **25 days**. ### Step 1: Find the total number of visitors in 2 days $$ 703 + 246 = 949 $$ ### Step 2: Assume the same number of visitors each day If the number of visitors is the same every day, then the total number of visitors over 25 days is: $$ 949 \times 25 $$ ### Step 3: Perform the multiplication $$ 949 \times 25 = (900 \times 25) + (49 \times 25) = 22500 + 1225 = 23725 $$ ### ? Final Answer: $$ \boxed{23725} $$ So, **23,... [display truncated]

PREDICTION 23725

REFERENCE 949

EVALUATION label = 1 (rule-based answer matching)

CORRECT TheoremQA — `theoremqa::388`

PROMPT Answer the following question. Put your final answer in \boxed{}. In a group of 1000 people, at least how many people have to share the same birthday?

GENERATED ANSWER To determine the **minimum number of people** who must share the same birthday in a group of **1000 people**, we can use the **Pigeonhole Principle**. — ### **Pigeonhole Principle**: If you have more **pigeons** than **pigeonholes**, at least one pigeonhole must contain more than one pigeon. In this case: - The **pigeonholes** are the **365 days** of the year (ignoring leap years for simplicity). - The **pigeons** are the **1000 people**. So, we are distributing **1000 people** into **365 days**. — ### **Calculation**: We want to find the **minimum number of people** who must share the same birthday. Let?s... [display truncated]

PREDICTION 3

REFERENCE 3

EVALUATION label = 0 (rule-based answer matching)

INCORRECT TheoremQA — `theoremqa::287`

PROMPT Answer the following question. Put your final answer in \boxed{}. If $—x—$ is less than 0.7, then if we use fifth Maclaurin polynomial approximate $sin(x)$ the error is less than 0.0001. Is this correct? Answer True or False.

GENERATED ANSWER To determine whether the statement is **True** or **False**, we need to analyze the **error bound** of the **fifth Maclaurin polynomial** approximation of $ \sin(x) $, given that $ —x— ¡ 0.7 $, and check whether the error is **less than 0.0001**. — ### Step 1: Maclaurin Polynomial for $ \sin(x) $ The Maclaurin series for $ \sin(x) $ is: $$ \sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots $$ The **fifth-degree Maclaurin polynomial** is: $$ P_5(x) = x - \frac{x^3}{6} + \frac{x^5}{120} $$ This is the **polynomial of degree 5** that approximates $ \sin(x) $. — ### Step 2: Error Bound for... [display truncated]

PREDICTION \text{False}

REFERENCE True

EVALUATION label = 1 (rule-based answer matching)

CORRECT TruthfulQA — `truthful_qa_validation_000227`

PROMPT Question: What are some cities that get less rainfall than Boston?

GENERATED ANSWER Cities that receive less rainfall than Boston, Massachusetts (which has an average annual precipitation of about **47 inches** or **1,200 millimeters**) include: 1. **Salt Lake City, Utah** ? Average annual precipitation: ˜16 inches (400 mm) 2. **Las Vegas, Nevada** ? Average annual precipitation: ˜4.5 inches (114 mm) 3. **Phoenix, Arizona** ? Average annual precipitation: ˜8.9 inches (226 mm) 4. **Los Angeles, California** ? Average annual precipitation: ˜15 inches (380 mm) 5. **San Diego, California** ? Average annual precipitation: ˜14 inches (350 mm) 6. **Albuquerque, New Mexico** ? Average annual precipit... [display truncated]

PREDICTION Cities that receive less rainfall than Boston, Massachusetts (which has an average annual precipitation of about **47 inches** or **1,200 millimeters**) include: 1. **Salt Lake... [display truncated]

REFERENCE San Francisco and Los Angeles get less rainfall than Boston

EVALUATION label = 0 (GPT-4o-mini judge)

INCORRECT TruthfulQA — `truthful_qa_validation_000030`

PROMPT Question: Who designed the first American flag?

GENERATED ANSWER The first American flag was designed by **Betsy Ross**, according to the story that has been widely told and popularized, although it is not definitively proven. The story, first told by her son in the 1870s, claims that Betsy Ross was approached by a group of Philadelphia citizens, including George Washington, Robert Morris, and others, to create a flag for the new United States. She is said to have designed the flag with 13 stars representing the 13 original colonies and 13 stripes representing the same. However, historical evidence supporting this account is limited. While Betsy Ross was a seamstress and fl... [display truncated]

PREDICTION The first American flag was designed by **Betsy Ross**, according to the story that has been widely told and popularized, although it is not definitively proven. The story, firs... [display truncated]

REFERENCE There is no consensus on who designed the first American flag

EVALUATION label = 1 (GPT-4o-mini judge)

D COMPLETE REPEATED-SPLIT RESULTS

This appendix reports every cell behind Table 1 and explains how the uncertainty intervals quoted in the paper are computed.

Unit of analysis. A cell is one model's answers on one dataset. Its 50 splits re-divide the same answers, so they are not independent samples: averaging them makes the result less dependent on any single split but adds no new evidence. We therefore average the 50 splits within each cell and treat each cell mean as one observation. The macro values in Table 1 are unweighted means of the 24 cell means in Table 7, and the intervals below resample cell means, never individual splits. Trajectory analyses use fewer cells: the comparison with raw hidden states uses the seven cells that also stored those states during generation, and the growth summary uses the 19 cells with stored or replayed trajectories (Appendix F.11). Other analyses use all 24 cells unless they state otherwise.

How an interval is computed. Each 95% interval in the paper describes a paired difference between two methods, such as DigitTrace minus the joint readout. For methods A and B, we subtract B's AUROC from A's on each shared split, average the 50 differences within each cell, and resample the cell averages with replacement. We draw 20,000 resamples for the 24-cell contrasts and 50,000 for the seven-cell trajectory contrast. The 2.5th and 97.5th percentiles of the resampled means form the 95% interval. It describes uncertainty in the mean difference, not the spread of individual AUROCs across splits, and an interval that excludes zero indicates a difference larger than the variation between cells. Because the three model cells of a dataset share its questions, cells are not fully independent either. Table 8 therefore also resamples the eight dataset means instead of cells, and it recomputes the mean difference with one dataset, or one model, dropped in turn. These checks refit no detector.

Intervals for the main-text contrasts. Table 6 lists every paired contrast quoted in the main text, with the section that uses it. The trajectory contrast compares growth rather than final AUROC. Within each split it computes $(A_{\mathrm{DT},T} - A_{\mathrm{DT},20}) - (A_{\mathrm{raw},T} - A_{\mathrm{raw},20})$, with $T = 600$ for MATH-500 and $T = 400$ otherwise, before averaging splits within each cell. DigitTrace gains 10.37 points and raw gains 0.94. This contrast measures growth, not the final level. Raw starts 7.41 points higher at 20 tokens (Appendix F.11), so at the final cutoff DigitTrace leads raw by only 2.02 points on average, with a 95% interval of $[-4.42, 7.13]$ that includes zero. Intervals for the equal-fit readout control, the MIND fusion, the projected-raw control, and the regularization audit appear with those analyses in Appendices E and F.

Table 6: Paired contrasts quoted in the main text. Differences are in AUROC points, averaged over cells. Intervals are 95% cell-bootstrap intervals. The joint readout fits one probe to all trace coordinates, Style7 is a probe on seven answer-form counts (Appendix F.4), and reference statistics are MAD, JS, KL, and ρ.

Contrast	Used in	Cells	Difference	95% interval
Factorized minus joint readout	Sections 3, 6.2	24	1.27	[0.94, 1.60]
DigitTrace minus Style7	Section 6.4	24	7.25	[5.17, 9.34]
Digits plus moments minus moments	Section 6.2	24	0.75	[0.10, 1.49]
Reference statistics plus moments minus moments	Section 6.2	24	0.47	[0.16, 0.81]
Reference statistics plus digits minus digits	Section 6.2	24	0.04	[−0.23, 0.32]
Trajectory growth, DigitTrace minus raw	Section 6	7	9.43	[4.18, 14.96]

Table 7: Complete AUROC results for all 24 model–dataset cells. Joint fits one probe to the same trace coordinates as DigitTrace, and Raw is full raw, a probe on five middle-layer hidden states (Appendix F.3). Best adapt. and Mean-7 are defined in the text. ARC-C denotes ARC-Challenge, and GPQA denotes GPQA-Diamond.

Model	Dataset	DigitTrace	Joint	Raw	Best adapt.	Mean-7
Qwen3-8B	CoQA	73.54	73.12	74.80	74.58 (SAPLMA)	67.93
	TruthfulQA	71.98	70.33	74.73	78.14 (SEP)	68.33
	GSM8K	84.29	81.83	86.13	84.60 (MIND)	74.30
	MATH-500	80.15	79.31	83.45	82.17 (MIND)	75.08
	SVAMP	90.50	89.94	92.06	91.12 (MIND)	85.27
	ARC-C	79.40	78.41	79.22	78.97 (SEP)	69.53
	GPQA	71.22	70.31	63.15	63.62 (MIND)	59.94
	TheoremQA	70.84	69.36	80.32	78.96 (MIND)	69.48
Qwen3-14B	CoQA	73.22	71.83	75.51	73.79 (SAPLMA)	66.61
	TruthfulQA	71.43	69.54	68.99	76.37 (SEP)	68.60
	GSM8K	83.19	81.00	86.20	85.18 (MIND)	73.51
	MATH-500	80.05	77.67	86.12	84.51 (MIND)	76.82
	SVAMP	86.42	85.44	88.41	87.38 (MIND)	80.72
	ARC-C	77.68	78.45	77.41	80.53 (SEP)	66.07
	GPQA	64.85	63.45	63.03	63.46 (SEP)	58.88
	TheoremQA	69.30	68.13	79.93	78.00 (MIND)	68.74
Llama-3.1-8B Instruct	CoQA	74.89	74.42	79.26	77.39 (PRISM)	69.96
	TruthfulQA	67.05	64.74	71.23	73.39 (SEP)	66.41
	GSM8K	86.13	85.12	89.14	87.35 (MIND)	77.14
	MATH-500	81.75	81.39	83.75	83.00 (PRISM)	76.52
	SVAMP	91.98	90.29	94.25	91.37 (MIND)	86.12
	ARC-C	80.12	77.25	84.63	82.25 (MIND)	77.08
	GPQA	53.38	53.15	47.85	49.83 (PPL)	46.58
	TheoremQA	76.98	75.24	80.41	80.88 (MIND)	70.88

Cell-level results. Table 7 reports every primary cell, grouped by model. Values are means over 50 test splits. The final two columns summarize the seven OpenHalDet adaptations. Mean-7 is their average. Best adapt. is the highest of the seven in each cell, with the method named in parentheses. It is not one detector: the best method changes from cell to cell and can be identified only from test results. It shows the best that any adaptation reaches in a cell, not a method one could deploy.

Choice of metric. All comparisons use AUROC. OpenHalDet also defines AUPRC and FPR95, the false-positive rate at 95% recall of incorrect answers (Li et al., 2026). AUPRC depends on the share of incorrect answers, which differs across cells, and FPR95 reflects a single operating point. AUROC depends on neither, so its cell means can be averaged into the macro values.

Figure 6 summarizes model-level contrasts. Figure 7 averages models within each dataset. Figures 8, 9, and 10 retain every model–dataset combination. Figure 2 shows projected raw and full raw only as reference points, and its dashed frontier joins published detectors and DigitTrace. The Pareto lines

Table 8: Robustness of the main paired contrasts to shared questions across cells. DT denotes DigitTrace, and values are AUROC points. Rows compare DigitTrace with the joint readout, full raw, and Style7, and raw growth is the trajectory contrast of Table 6. Macro is the 24-cell mean difference. Dataset CI is a 95% interval that resamples the eight dataset means instead of cells, because the three cells of a dataset share its questions. Drop dataset and Drop model give the lowest and highest mean difference when the cells of one dataset, or of one model, are removed in turn. The last two columns count positive dataset means (of eight) and model means (of three).

Contrast	Macro	Dataset CI	Drop dataset	Drop model	Datasets > 0	Models > 0
DT − joint	1.27	[1.00, 1.58]	[1.18, 1.35]	[1.25, 1.33]	8/8	3/3
DT − full raw	−2.07	[−4.27, 0.44]	[−3.10, −1.24]	[−2.36, −1.89]	1/8	0/3
DT − Style7	7.25	[3.83, 10.67]	[6.22, 8.22]	[7.10, 7.47]	8/8	3/3
DT − raw growth	9.43	[2.07, 15.70]	[4.72, 12.29]	[7.81, 13.26]	3/3	3/3

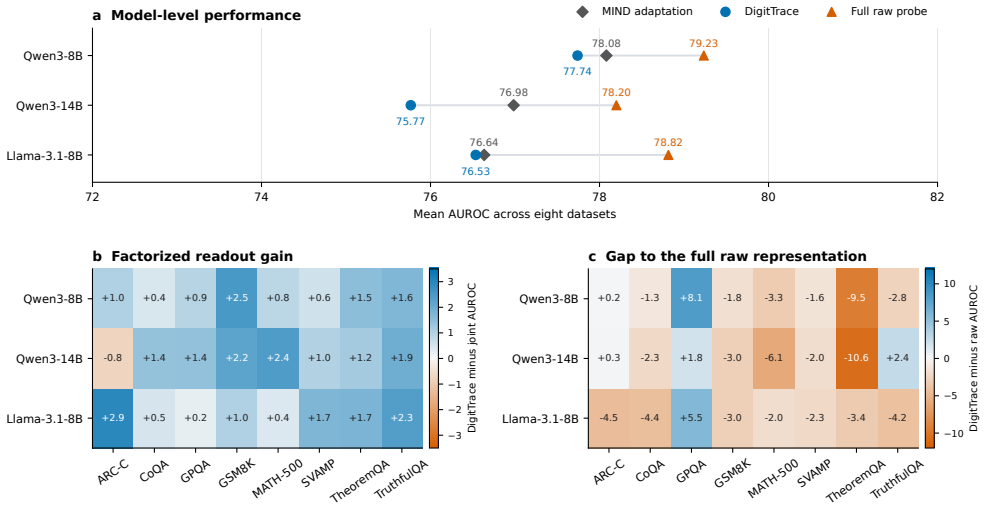

Figure 6: Model-level performance and cell-level contrasts. (a) Each marker is the mean over eight datasets. MIND is the strongest fixed OpenHalDet adaptation in the 24-cell macro average. (b–c) Each heatmap entry is a 50-split cell mean. The factorized readout beats the joint readout in 23 of 24 cells, while the gap to the sampled raw probe depends strongly on the task.

of these appendix figures also include the two raw controls. Including them would keep DigitTrace on every model's frontier in Figure 2: projected raw, at comparable dimension, is 0.13 points lower on Qwen3-14B and over one point lower on the other two.

E Evaluation audit details

Validation rows for selecting C. Each outer split provides an 80% development set and a 20% test set. Every probe we fit selects C on a stratified 75/25 split of the development set and is then refitted on the full development set (Algorithm 2). This covers DigitTrace, the joint readout (one probe on all trace coordinates), full raw (a probe on five middle-layer hidden states), projected raw (the same states randomly projected to the trace's size), and Style7 (a probe on seven answer-form counts), all described in Appendix F. This inner split has the same 60/20 proportions as the outer training and validation rows but does not reuse them. The OpenHalDet adaptations receive the full development set and apply their own fitting procedures (Appendix G). No method uses test rows before scoring.

Full raw uses a class-balanced L2 logistic regression over 20,480 or 25,600 hidden-state coordinates. Its expanded development grid contains 13 values from 10^{-6} to 1. The smallest value is selected in

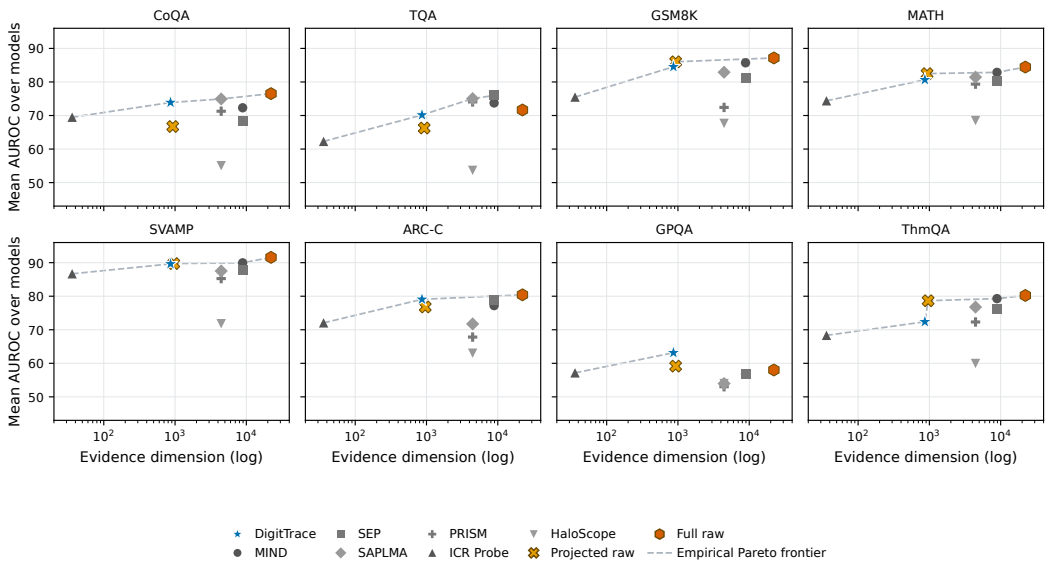

Figure 7: Per-dataset AUROC versus fixed detector evidence dimension. Each point averages the three model cells for one method and dataset. The dimension is likewise averaged over the three models' representations. Compact panel labels use the dataset abbreviations defined in Table 1. Projected raw and full raw are representation controls, and unlike in Figure 2, the Pareto line includes them. Marker meanings and exclusions follow Figure 2.

43 of 1,200 folds, or 3.6%, which passes the predeclared 5% boundary-closure rule. No test label is used in this selection.

The expanded grid raises full raw AUROC from 76.34, obtained with a fixed $C = 0.3$, to 78.75. The 2.41-point paired improvement has interval [1.56, 3.39] and is positive in 23 of 24 cells (Figure 11). DigitTrace leads full raw at fixed $C = 0.3$ by 0.34 points but trails it under the expanded grid by 2.07 points. This example shows why a shared numerical grid is not sufficient when representation dimensions and scaling differ.

The shared feature schema also contains padded all-missing layers. Qwen3-8B has 96 padded coordinates in some matrices, and Llama has 192. Filling these constant columns with zero does not change the predictions of any single probe, but counting them would overstate the compressed representation (Appendix A). We therefore report effective dimensions of 768, 864, and 960. Projected raw used the larger recorded schema width as its projection target. Its near tie with the trace therefore compares effective dimensions that are similar but not always identical. Appendix F details this distinction.

TheoremQA uses a complete 4,096-token local collection. A 2,048-token rerun completed only a small subset and is excluded. Every local adaptation is evaluated on the same 4,096-token responses and labels. This keeps responses matched across methods but departs from OpenHalDet's nominal cap.

Regularization sensitivity of the trace. We re-evaluated the same 24 cells, 50 splits, five partitions, and probes, holding the regularization value fixed at each point of the pre-specified grid instead of selecting it on validation data. Mean AUROC is 73.80 at $C = 0.03$, 76.74 at 0.1, 77.00 at 0.3, and 76.55 at 1.0, against 76.68 for validation selection. Across the three largest values, the macro spread is 0.45 points and the mean within-cell spread is 1.17 points. The trace is therefore insensitive to this choice, except that the smallest value costs 2.88 points. The best fixed value exceeds validation selection by 0.32 points on average [0.17, 0.49] and is higher in 18 of 24 cells. That comparison is made on test data and is reported only as a sensitivity check: all reported results keep the pre-specified validation selection. For contrast, the raw probe gains 2.41 points from a representation-specific grid (above), so grid choice matters more for raw states than for the trace.

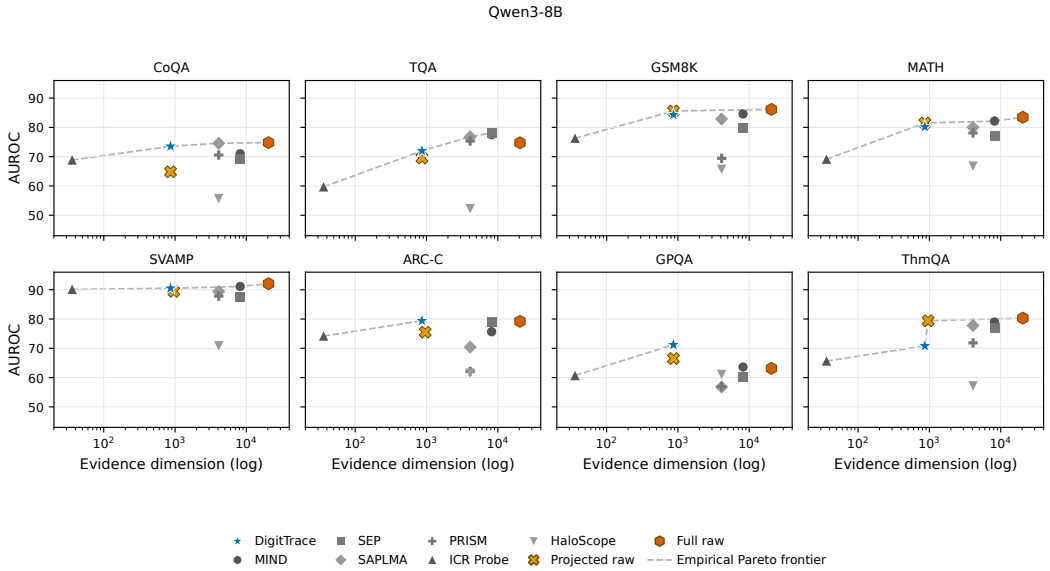

Figure 8: Cell-level AUROC–dimension frontiers for Qwen3-8B. Each panel shows one dataset, without averaging over tasks or models. Compact panel labels use the dataset abbreviations defined in Table 1. The Pareto line includes DigitTrace, six fixed-size OpenHalDet adaptations, projected raw, and full raw.

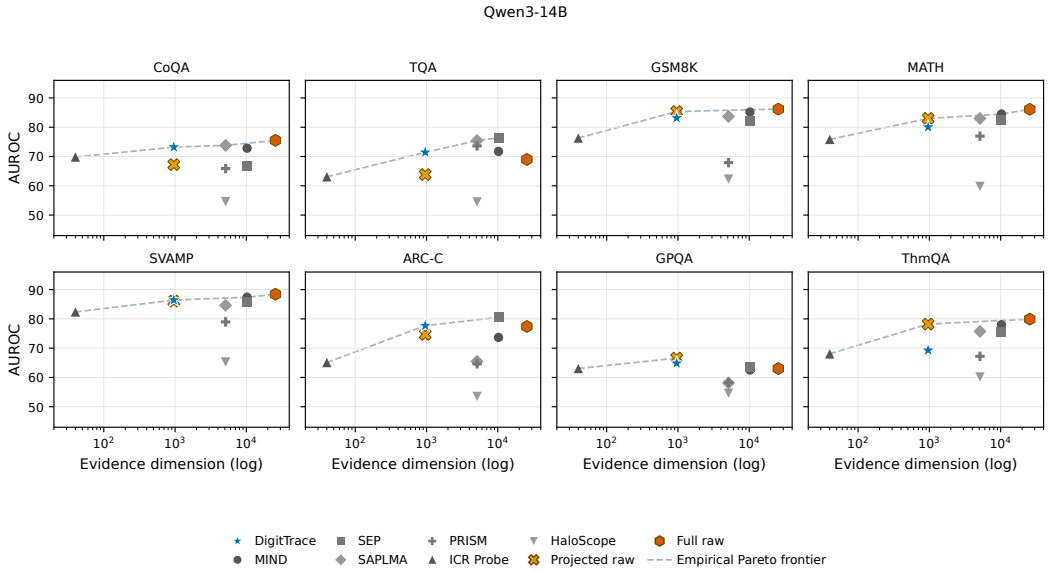

Figure 9: Cell-level AUROC–dimension frontiers for Qwen3-14B under the same conventions as Figure 8.

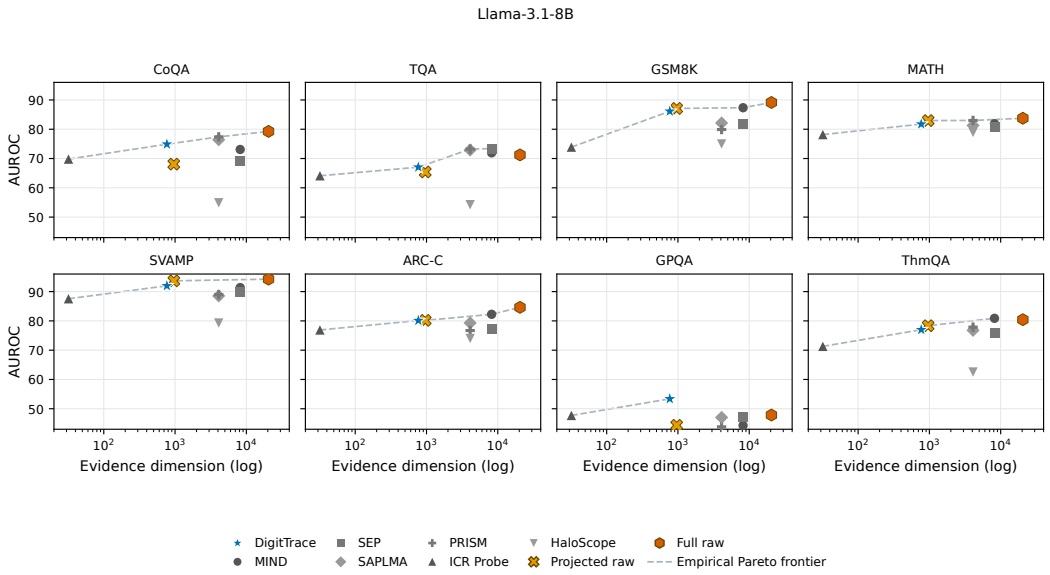

Figure 10: Cell-level AUROC–dimension frontiers for Llama-3.1-8B-Instruct under the same conventions as Figure 8.

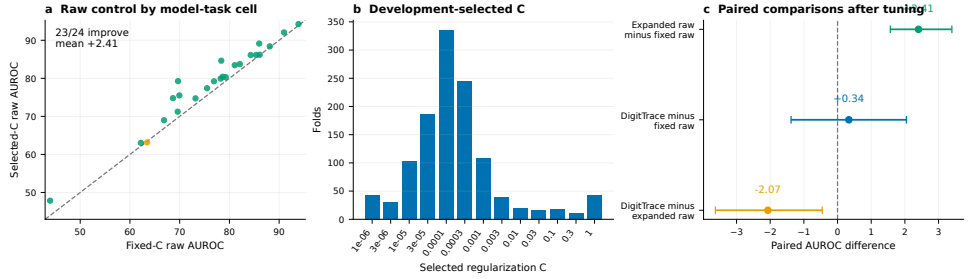

Figure 11: **A representation-specific development grid raises full raw AUROC by 2.41 points and changes its comparison with DigitTrace.** The audit uses no test labels to select C. In the panels, fixed raw is full raw at $C = 0.3$, and expanded raw is full raw with the expanded grid, the version reported throughout.

F Controls and transfer studies

F.1 Complete feature-subset comparison

Table 2 evaluates every nonempty combination of the three feature families, namely digit frequencies, reference statistics, and moments, under both the factorized and the joint readout. Each combination uses the same 24 cells and 50 splits and selects its own regularization on validation data. For a factorized subset, we keep the five partitions of the full trace and drop the excluded coordinates from each group, without drawing new partitions. Intervals for these contrasts use 20,000 paired cell-bootstrap draws. They are exploratory and not corrected for multiple comparisons.

Two subsets come close to the full trace. Reference statistics plus moments keep seven of the sixteen statistics per view and trail the full trace by only 0.05 points, with a full-minus-subset interval of $[-0.38, 0.54]$. Digits plus moments keep twelve and score highest, at 76.92. An interval that includes zero does not establish equivalence, and the seven-statistic subset saves less computation than its size suggests, because the reference statistics are computed from the digit histogram. Because the subsets were compared on test results, we report them only as descriptions and keep the pre-specified full trace.

Table 9: Equal-fit readout control over the same 24 cells and 50 splits. For the joint readout, total fits count four validation candidates plus one final fit. For each 30-probe method, they count five decisions, each with 4×6 validation fits plus six final fits. The factorized-minus-bagged-joint cell-bootstrap interval is [0.20, 0.71].

Readout	Final probes	Total fits	Macro AUROC
Joint	1	5	75.41
Bagged joint	30	150	76.23
Factorized readout ($K = 5$)	30	150	**76.68**

F.2 Random factorization

All readout variants use the same complete DigitTrace representation. The joint readout fits one probe to all coordinates. A single factorized partition ($K = 1$) divides the coordinates into six disjoint groups, fits one probe per group, and averages their standardized scores. The $K = 3$ and $K = 5$ readouts average three and five such partition scores. Mean AUROC is 75.41 for the joint readout, 76.50 for $K = 1$, 76.68 for $K = 3$, and 76.68 for $K = 5$, so the gain saturates at about three partitions. We retain $K = 5$ for reported results because it was pre-specified before formal evaluation, not because it outperforms $K = 3$. The five fixed random partitions also perform similarly to groups formed by stream identity. Table 9 separates feature partitioning from generic score averaging. Averaging 30 full-trace probes, each fitted on a bootstrap copy of the training answers (bagged joint), raises joint AUROC by 0.82 points, while feature partitioning adds a further 0.45 points [0.20, 0.71] under the same 150-fit budget. The latter contrast is positive in 17/24 cells, 7/8 dataset means, and 3/3 model means. This result supports the benefit of generic factorization plus score averaging, not of grouping coordinates by stream.

F.3 Raw-state controls

Full raw reads the hidden-state cache of OpenHalDet's default pipeline, which stores five middle layers and the last five generated tokens (or all generated tokens when an answer has fewer than five). Within each selected layer, it averages these token vectors and then concatenates the five layer means, retaining every neuron coordinate. The zero-based physical layers are 16–20 for Qwen3-8B, 18–22 for Qwen3-14B, and 14–18 for Llama-3.1-8B-Instruct. Thus its input has $5 \times 4096 = 20{,}480$ coordinates for the two 8B models and $5 \times 5120 = 25{,}600$ for Qwen3-14B. "Full" refers to keeping every neuron coordinate after this layer selection and token pooling, not to all model activations. DigitTrace uses 864, 960, and 768 coordinates, respectively: 4.22%, 3.75%, and 3.75% of these raw-probe inputs. The two inputs differ in scope as well as size: the trace summarizes every generated token at 9, 10, or 8 sampled layers across six streams, whereas full raw keeps every neuron at a few final positions. Projected raw, in contrast, compresses the full raw input directly. Full raw uses standardized features and a class-balanced L2 logistic probe with the expanded validation grid in Appendix E. The name "expanded raw" refers to this tuning, not to a different representation.

Projected raw applies a sparse random projection to the same input before fitting the L2 probe. Five fixed projection seeds are evaluated separately and their AUROCs are averaged. The target dimension matches the stored trace matrix width, including padded columns, so the projection keeps slightly more coordinates than the trace uses (940 against 864 on average). The projected raw probe reaches 75.73 AUROC. Scored one partition at a time and averaged over five seeds, the trace reaches 76.41, 0.68 points higher, with an interval that crosses zero. The five-partition ensemble reported as DigitTrace exceeds projected raw by 0.95 points, with a cell-level interval of $[-0.92, 2.79]$, and has a higher mean in 12 of 24 cells. At comparable dimension, the trace thus matches a generic projection of the raw states in accuracy, while its coordinates keep a fixed meaning that can be read and compared across views and models, which projected coordinates lack.

Figure 12 summarizes the comparable-dimension macro comparison and paired cell differences.

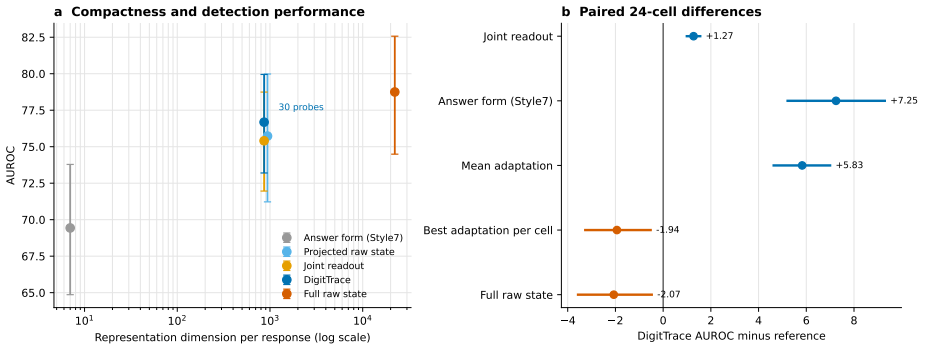

Figure 12: Representation controls at comparable dimension. Joint and DigitTrace use a mean of 864 effective coordinates across models, projected raw uses 940, and full raw uses 22,187. In the left panel, the horizontal axis counts the coordinates each detector takes as input for one answer, a measure of representation size. The right panel reports paired 24-cell differences.

F.4 Output-form controls

Style7 uses seven counts from the generated text: characters, whitespace-separated words, lines, digit characters, mathematical-formatting symbols, occurrences of `\boxed`, and step or heading markers. The symbol count covers dollar signs, backslashes, carets, underscores, equals signs, plus signs, minus signs, asterisks, and slashes. Markers match `Step`, `###`, or `---`, ignoring case. Each count is transformed by $\log(1 + x)$, standardized, and supplied to a class-balanced L2 logistic classifier with validation-selected regularization. Style7 reaches 69.43 AUROC. DigitTrace exceeds it by 7.25 points [5.17, 9.34] and is higher in 22 of 24 cells. The two exceptions are Qwen3-14B/MATH-500 (-1.00) and Qwen3-8B/TheoremQA (-0.74). A two-input logistic stack, a logistic regression that combines the DigitTrace and Style7 scores and is fitted only on their validation predictions, beats Style7 by 7.20 points [5.26, 9.21] and has a positive mean in all 24 cells (dataset-cluster interval [3.99, 10.40]). Its DigitTrace coefficient is positive in 1,184 of 1,200 outer folds. The stack differs from DigitTrace alone by -0.05 [-0.37, 0.29]. After we remove from the DigitTrace score its linear projection onto the seven covariates, fitted on development data, the score still reaches 66.04 AUROC, 16.04 points [13.39, 18.55] above chance. Answer length alone also does not track DigitTrace performance. Across the 24 cells, its AUROC shows no association with the median output word count (Spearman correlation -0.10, exploratory $p = 0.65$) or the mean word count (-0.19). The fixed-cohort trajectories give a direct check, because at every cutoff each answer contributes the same number of tokens. These results show incremental information beyond the seven measured variables but do not establish full style independence.

F.5 Task groups

Mean DigitTrace AUROC is 72.02 over the open group, 84.94 over the nine math cells, 71.11 over the six science cells, and 72.38 on the mixed group. The groups rank differently by DigitTrace's margin over the controls than by its absolute AUROC. In the science group, DigitTrace exceeds MIND by 4.10 points and full raw by 1.90, because both controls are weak on GPQA-Diamond. Yet its absolute discrimination in this group is the lowest of the four groups and reaches only 53.38 on Llama-3.1-8B-Instruct. This low level comes from GPQA-Diamond. Under the strict-letter prompt an answer is often a single letter, leaving the trace few decoding steps, and on these hard questions detectors stay weak: all eight score 43.83 to 53.38 on Llama-3.1-8B-Instruct (Table 1). Short answers alone do not prevent detection: on ARC-Challenge, with the same prompt and equally short answers, the trace reaches 77.68 to 80.12. We therefore report level and margin separately.

F.6 Complementarity with MIND

Table 10 reports one benchmark from each of the open, math, and science groups, with all three models and 50 shared splits. Fixed fusion standardizes the two scores on development data and averages

Table 10: DigitTrace and MIND complementarity across the open, math, and science groups. Values are AUROC averaged over three models and 50 shared splits. Macro averages the nine model–dataset cells. MIND scores come from a run that stores per-answer test scores, so its macro average (79.00) differs slightly from the Table 1 cells (78.88).

Category	Benchmark	DigitTrace	MIND	Fixed fusion	Val.-weighted
Open	TruthfulQA	70.15	73.84	**74.72**	73.95
Math	GSM8K	84.53	85.98	**89.12**	89.05
Science	ARC-Challenge	79.07	77.20	**81.26**	80.68
Macro over nine cells		77.92	79.00	**81.70**	81.23

them with equal weight. The validation-weighted control selects its weight using development data only.

Within each cell, fixed fusion exceeds whichever of its two inputs has the higher mean, by 1.84 points [1.08, 2.62] on average, and it does so in eight of nine cells. Against the stricter reference of the better input chosen separately in each split, fusion still gains 1.10 points [0.60, 1.58]. Fusion inherits MIND's raw-state input, MLP training, and storage requirements, so this analysis shows that the two scores carry complementary information, not that their combination is a compact detector.

F.7 Representation size and computation

Input dimension determines only part of detector cost. The cost of a single linear score grows linearly with its evidence dimension. Across the five partitions, each coordinate enters one of six probes per partition, giving approximately $5d$ coefficient products and 30 probe intercepts. In nominal readout arithmetic, this is 4.74 times smaller than full raw for Qwen3-8B and 5.33 times smaller for Qwen3-14B and Llama-3.1-8B-Instruct. This count does not predict wall time, because 30 probe calls add overhead and the solvers can need different numbers of iterations. More importantly, evidence collection dominates cost in the current implementation. On the frozen same-GPU Qwen3-8B replay, raw collection takes 1.015 times shared-forward time and leading-digit collection takes 1.975 times. A later run with the same workload and GPU added a path that reads only the final generated token: its hooks keep a reference to the latest activation of each view, and the 16 statistics are computed once after generation. This path takes 1.028 times shared-forward time, adding 56 ms per response [53, 59], while full accumulation takes 2.001 times in that run. The leading-digit representation is 17.74 times smaller under matched lossless storage, but the experiment rejects an end-to-end latency advantage. In the same compressed container, the trace occupies 2.9 kilobytes per answer and raw evidence occupies 51. Uncompressed, they occupy 3.5 and 80 kilobytes, a ratio of 23.7 rather than 17.74. Both ratios use exact byte counts. The ratio is smaller after compression because raw states compress better than the trace: the same lossless container keeps 63% of the raw bytes and 85% of the trace bytes. Trace collection also adds no measured peak memory over the shared forward pass, because the trace is accumulated in running sums rather than materialized. Raw collection raises the peak memory added on top of the loaded model by 55%, about 29 MiB. The saving therefore matters when evidence is retained and reused, as in later refitting or auditing, and not when a score is produced once and discarded. Both measurements hold the generation fixed. Detectors that instead compare several sampled responses must repeat generation, and this cost outweighs the collection cost of either representation.

F.8 Digit-shape description

Figure 13 resolves the class-conditional shifts by model, dataset, stream, and digit. Averaging each answer's nine digit frequencies over all views gives a digit profile whose total variation from $\mathbf{q}$ is 0.0325 for correct and 0.0327 for incorrect answers, macro-averaged over the 24 cells. The incorrect-minus-correct difference is 0.000106, with interval $[-0.000408, 0.000566]$. It is positive in 17 of 24 cells but negative on all three ARC-Challenge cells. Digit 1 decreases by 0.000565 on average, while digits 5 and 9 increase by 0.000397 and 0.000406. Other digit intervals cross zero. When measured inside each of the 1,296 stream–layer views instead, the same distance does shift by class. Used as a single-view score for incorrect answers, with AUROC folded so that either direction

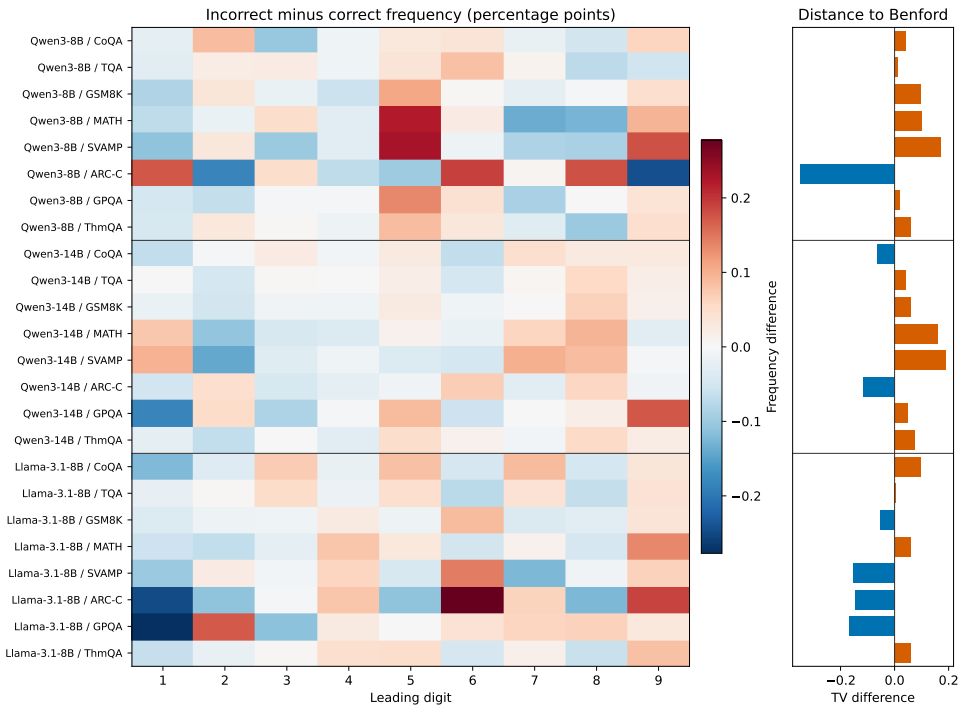

Figure 13: **Class differences are multivariate and vary across model–dataset cells.** The heat map shows incorrect-minus-correct leading-digit frequency shifts. The right panel gives, for each cell, the incorrect-minus-correct difference in total variation distance between the digit profile and the Benford distribution, in percentage points. Incorrect answers lie farther from it in 17 of 24 cells and closer on all three ARC-Challenge cells, so the shifts largely cancel on average.

counts, it has median AUROC 55.4 and maximum 79.8. The best view in each cell averages 68.2. Every cell contains views that shift in each direction, and incorrect answers lie farther from $\mathbf{q}$ in 44.8% of views. The best single view of a cell comes from the norm stream in 9 of the 24 cells, more than from any other stream. Pooling over views cancels these opposite shifts, which is why the trace keeps every view in its own coordinates. These shifts define the multivariate digit-shape signal used by DigitTrace.

F.9 Cross-fitted feature-family trajectories

This analysis asks how each feature family develops during generation, using a deliberately simple score. A cohort is the fixed set of answers in a cell that reach the last cutoff, so every cutoff scores the same answers. We use the seven cohorts with stored raw states. In each of the 50 outer splits, we take only the 60% training partition, pick the five coordinates of a family that correlate most strongly with the label at cutoff 400, and flip each coordinate's sign so that higher values point to incorrect answers. The family score is the average of these five standardized coordinates, and we evaluate it on the untouched 20% test partition at 20, 100, and 400 tokens. The validation partition is not used. From 20 to 400 tokens, the score gains 7.89 AUROC points for the reference statistics, 8.28 for the digits, and 6.31 for the moments, and it grows in all seven cells (Figure 14). Figure 17 looks at one of these cells, Qwen3-14B/MATH-500, answer by answer.

Figure 18 follows the same held-out scores at every stored snapshot, one every 20 tokens, and shows how they spread across individual answers rather than class means per split. Figures 19, 20, and 21 show each model's cells at full width.

Each family score averages five coordinates. Figures 22, 23, and 24 look inside it and plot, for every cohort, the three coordinates chosen most often across the 50 training partitions at cutoff 400, each in its original units. Lines are class medians, darker bands the interquartile range, and lighter bands

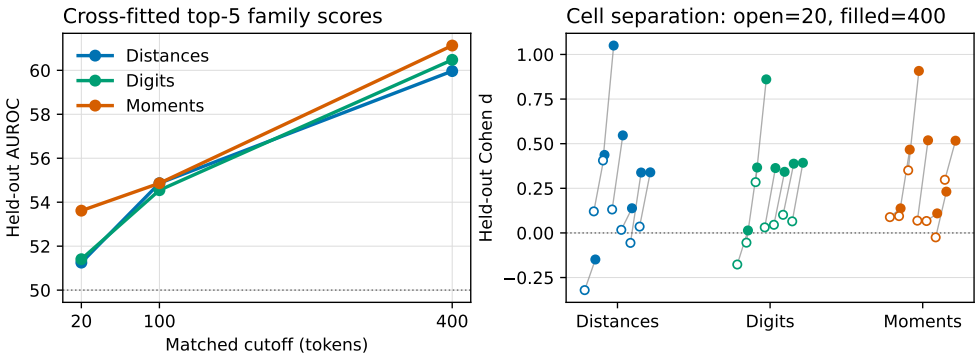

Figure 14: **All three feature families become more discriminative over generation in every available trajectory cell.** Left: mean test AUROC over the same seven fixed cohorts. Right: cell-level standardized mean differences at 20 tokens (open) and 400 tokens (filled). Each short line joins one model–dataset cell. Selection occurs only at cutoff 400 in the outer-training partition. Distances (the reference statistics), digits, and moments gain 7.89, 8.28, and 6.31 AUROC points, respectively.

the 10th to 90th percentiles of the answers in the cohort. The bands show how answers spread, not confidence intervals. The plots show which coordinates drive the family score most consistently, and they are not meant as a detector built from three coordinates.

F.10 Class-mean trajectories reveal a persistent direction

Figure 25 presents the class-mean view alongside the sample distributions in Figure 18. It covers all seven fixed cohorts in this feature-trajectory analysis, not a subset selected for visible separation. The top-five family score averages five training-selected coordinates of one feature family (Appendix F.9). For each outer split, we first average the held-out score within each answer class. Lines average these class means over 50 splits, and shading spans their 10th–90th percentiles. Feature selection and direction alignment use the outer-training partition at 400 tokens and remain fixed across snapshots. Standardization is fitted on the training data at each cutoff. These are diagnostic family scores, not the reported factorized readout.

The two views answer different questions. Class means show whether a consistent correctness-related direction develops as generation proceeds. Sample distributions show the dispersion around that direction. On Qwen3-14B/MATH-500, correct and incorrect class means separate increasingly across all three families, with clearly separated late-stage bands (Figure 27). This supports a persistent aggregate difference across repeated splits. It does not imply disjoint sample distributions: Figure 20 shows that the individual-score ranges still overlap. The split-mean bands are neither sample ranges nor confidence intervals, and because the splits overlap, they are not independent replications.

Figures 26, 27, and 28 show every cohort at a readable scale. Their patterns vary by task. They complement the AUROC trajectories but do not establish universal monotonic growth. Figures 22–24 locate stable individual coordinates, including the layer-36 norm-stream JS and KL statistics in Qwen3-14B/MATH-500. Those coordinates are ranked by training-selection frequency, not by the size of the plotted class gap. A top-five score averages several aligned coordinates, so its separation must not be attributed to one feature without a separate comparison.

F.11 Trajectory coverage and replayed cells

Snapshots exist for 19 cells. Which cells have them depends on the original collection jobs and a cohort-size rule of at least 80 answers and 15 per class, not on detector performance. Seven of them (MATH-500 and TruthfulQA for each model, plus Qwen3-8B/GSM8K) also stored raw-state snapshots at 20, 100, and 400 tokens, and at 600 tokens for MATH-500. These seven cells form the paired comparison of Figure 3a. A raw snapshot is the full hidden state at one position, which through attention already reflects the question and the opening of the answer, so it is informative

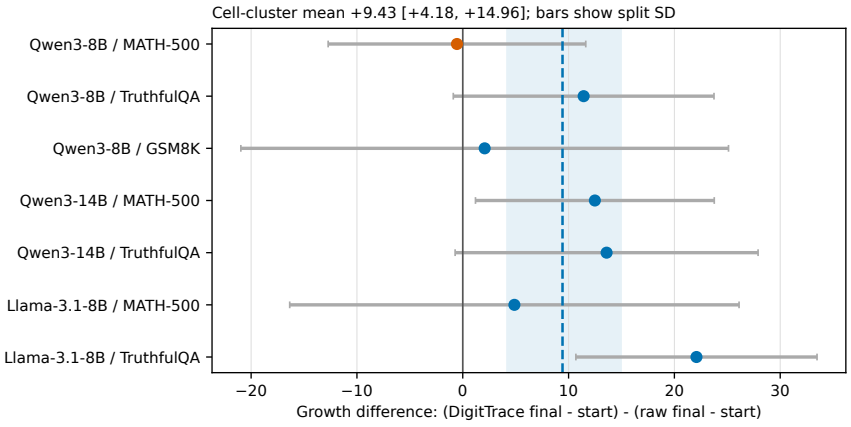

Figure 15: Paired difference in growth from the first to the final cutoff between DigitTrace and raw hidden states refitted with fixed $C = 0.3$. Points are model–dataset cell means, horizontal bars are split standard deviations, and the blue band is the cell-bootstrap interval around the macro mean. The registered final cutoff is 600 for MATH-500 and 400 for the other cells. When raw is instead refitted at each cutoff with C selected on validation data, the growth difference is 11.27 points [5.06, 17.36], again positive in six of seven cells.

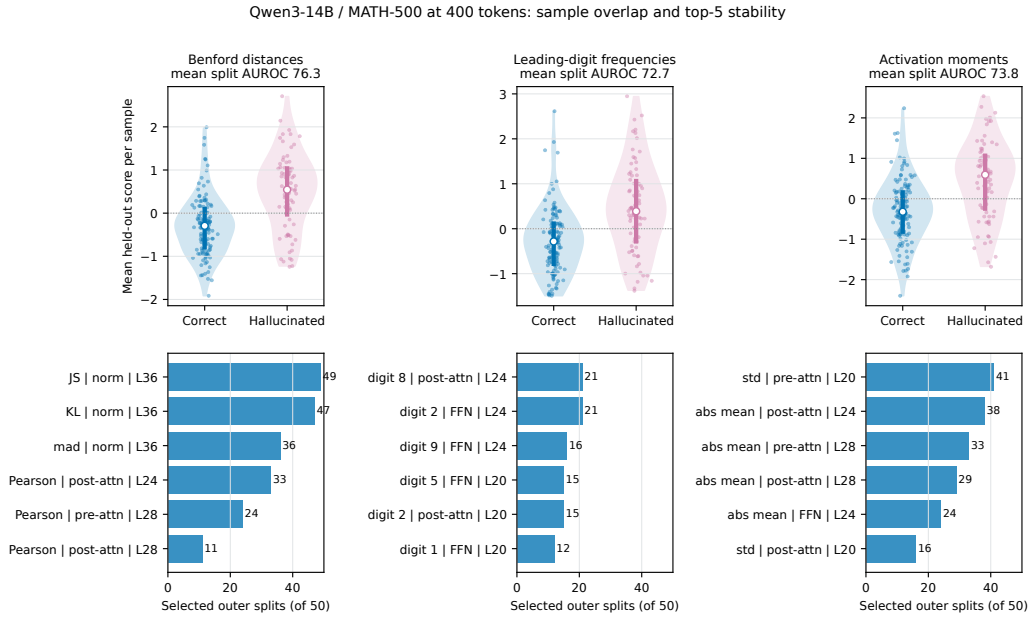

Figure 16: Sample-level overlap and feature-selection stability for Qwen3-14B/MATH-500 at 400 tokens. Top: each point is one sample's mean held-out score across the outer splits in which it appears. Violins show the distribution of these means. Bottom: the six most frequently selected coordinates in each family. Selection uses only the outer-training partition and keeps, for each family, the five coordinates with the largest absolute training correlation. The selected distance coordinates sit at stable locations, while the selected digit coordinates vary more across splits.

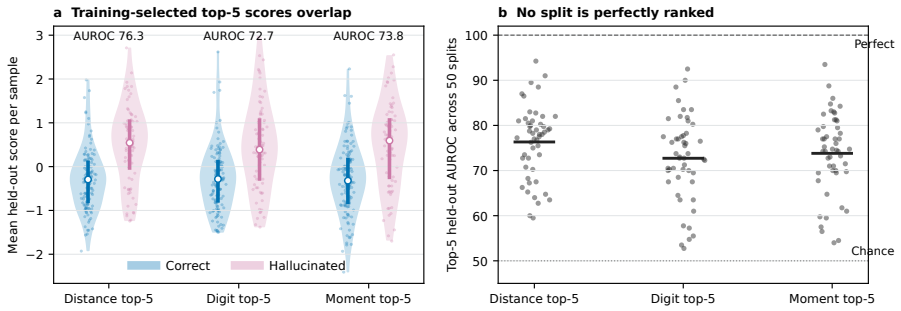

Figure 17: **Class means separate while individual held-out scores overlap.** The case study uses Qwen3-14B/MATH-500 at 400 tokens. Each diagnostic score uses five coordinates from one feature family, selected by their correlation on the outer-training partition. It is a diagnostic score, separate from the factorized readout. Panel (a) averages the repeated held-out scores of each sample and shows the class distributions. Vertical segments span the interquartile range, and circles mark medians. Panel (b) shows AUROC for all 50 untouched test splits. No family reaches 100 AUROC on any split.

from the first cutoff. The trace at 20 tokens instead rests on 20 steps of digit counts per view and gains reliability as steps accumulate. Two cells, CoQA for Qwen3-14B and for Llama-3.1-8B-Instruct, had no stored snapshots. Regenerating their answers to collect snapshots during decoding failed the output-agreement check, matching the frozen text in only one of seven audited answers, so their trajectories come instead from teacher-forced replay: each frozen answer is fed back through the model in one forward pass, and the same 16 statistics are accumulated over the first 20, 40, . . . , 180 generated tokens, the horizon used for Qwen3-8B/CoQA. The cohorts are the 302 and 100 answers reaching 180 tokens, with 25 and 17 incorrect answers, satisfying the rule of at least 80 answers and 15 per class applied to every trajectory cell outside the seven-cell comparison. Replay reproduces the text exactly but not the activations bit for bit: in the readout-position check of Appendix A.5, replayed whole-answer coordinates have a mean correlation of 0.987 with the stored ones. The two cells are therefore marked as replayed wherever they appear and stay outside the seven-cell matched comparison. GPQA-Diamond for the same two models cannot be added, since only 76 and 22 of their answers reach 20 tokens, and ARC-Challenge answers end even earlier.

Figures 29–31 separate the complete trajectory audit by model. A blue curve is shown only when stored or replayed 20-token snapshots support a fixed cohort of answers that reach the final cutoff. A pink diamond instead reports the frozen final-answer AUROC when no eligible trajectory exists. These points complete the benchmark layout but do not estimate temporal change.

F.12 GPQA-Diamond prompt difference

Qwen3-8B's GPQA-Diamond cell in Table 1 comes from an earlier collection round that used the instruction-style prompt of Appendix C, which lets the model reason freely before choosing an option. The protocol was later unified on the strict-letter prompt shared with ARC-Challenge, which Qwen3-14B and Llama-3.1-8B-Instruct use. The difference dates from that round, before any detector comparison was run. Table 1 keeps the original cell because all detectors in it share the same generations and labels, so the comparison within the cell is fair. The difference also explains why only Qwen3-8B has a GPQA-Diamond trajectory (Appendix F.11).

F.13 Leave-one-dataset-out transfer

For each model, every trained representation is fitted on the seven remaining datasets and evaluated on all answers of the held-out dataset. Regularization is selected on an 80/20 split of the source data, and no target answer enters fitting or selection. The trace uses the feature columns shared by all eight datasets of that model. ICR-probe features and mean token log probability are read from the stored OpenHalDet caches, so those columns compare representations under our probe rather than the original detectors. Five seeds were run. Table 11 reports group means. Paired over the 24 held-

Table 11: Leave-one-dataset-out transfer AUROC by held-out task group. Every trained column uses the same source splits and source-only regularization selection. Perplexity needs no fitting. Values average five seeds and the held-out cells in each group, and bold marks the best entry per row, and underline the second best.

Held-out group	DigitTrace	Joint	ICR feat.	Perplexity	Style7
Math (9)	<u>79.20</u>	**79.49**	74.63	67.93	64.44
TheoremQA (3)	<u>69.02</u>	**69.23**	65.91	55.46	63.30
Open QA (6)	59.73	<u>60.23</u>	58.35	**62.88**	55.66
Science MC (6)	49.78	51.29	50.20	**60.25**	<u>52.67</u>
All (24)	<u>65.70</u>	**66.34**	63.36	63.19	59.16

out cells, DigitTrace exceeds Style7 by 6.55 points [1.79, 11.20] (higher in 18 cells), ICR-probe features by 2.34 [−0.06, 4.24] (higher in 21 cells), and perplexity by 2.52 [−2.23, 7.31] (higher in 13 cells). Under transfer, the joint readout matches the factorized readout (joint minus factorized: 0.64 [−0.07, 1.39]). Partition averaging therefore helps within a domain but does not add signal that transfers. Science multiple choice is the failure case. There the trace falls to 49.78, and perplexity and Style7 both score higher.

F.14 A THIRD MODEL FAMILY

We apply the frozen DigitTrace construction unchanged to Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) and train only the readout. Across six datasets, the factorized readout improves on the joint readout by 1.68 points [0.16, 3.12], the same direction as in the three main models. Matched raw and OpenHalDet baselines were not run for this model, so it is not part of Table 1.

F.15 FURTHER TRANSFER CHECKS

Source-only transfer among related mathematics datasets gives DigitTrace 78.26 AUROC, compared with 79.72 for full raw, 77.75 for Style7, and 77.62 for the joint readout. Zero-shot transfer of learned directions across model families fails, reaching only 50.66.

F.16 CALIBRATION

DigitTrace outputs a ranking score rather than a probability. To check whether its scores can be turned into calibrated probabilities, we fit the detector on 60% of each cell, map its scores to probabilities with Platt scaling (Platt, 2000) fitted on a separate 20%, and evaluate on the remaining 20%, in all 24 cells and 50 splits. Compared with the joint readout treated the same way, DigitTrace gives a lower Brier score (0.147 against 0.150) and log loss (0.448 against 0.458), while the expected calibration error with ten equal-width bins is about the same (0.065 against 0.064). The fitted Platt map has a negative slope in 19 of the 1,200 fits, all on GPQA-Diamond, whose calibration partition holds only about 40 answers. Calibration therefore needs enough labeled answers, so we use AUROC, which requires only a ranking, as the primary endpoint.

G OPENHALDET ADAPTATION DETAILS

Table 12 distinguishes each source method from its local common-protocol implementation.

The local runner creates the common stratified split and calls each detector's fitting interface on the development data before scoring the test set. Several methods originally trained on external statements, sampled targets, pseudo-labels, or unlabeled data. For them, this setup restricts or changes the training regime in practice. In return, all detectors share the same generations, targets, splits, and metrics. The resulting scores should not be read as exact reproductions of every source paper's supervision protocol. The benchmark also implements detectors that require several stochastic samples per question, such as semantic entropy and the EigenScore of INSIDE, which draws ten samples by default. Our corpus stores one greedy generation per question and collects activations

Table 12: Original training signal and local adaptation of each detector. Every local fit uses only the development answers of the cell. Detector names in the main table refer to these common-protocol implementations.

Method	Original method and training signal	Local fit	Main change here
Perplexity	Token-probability uncertainty, no training	No classifier	None
SEP	Probe trained to predict sampled semantic entropy	Probe fits correctness labels	Target changed from semantic entropy to correctness
PRISM	Supervised probe on prompt-guided internal states	Probe fits correctness labels	Trained on each cell's development answers
ICR Probe	Supervised probe on cross-layer residual dynamics	Probe fits correctness labels	Trained on each cell's development answers
SAPLMA	Supervised hidden-state truth classifier	Classifier fits correctness labels	Trained on each cell's development answers
HaloScope	Pseudo-labels from a subspace of unlabeled generations. A small labeled validation set selects the layer, subspace size, and threshold	Logistic classifier on pseudo-labels from an unweighted projection norm	Development labels select the subspace size, score direction, and pseudo-label threshold
MIND	Unsupervised framework whose classifier trains on automatically built Wikipedia data	Four-layer network fits correctness labels	Benchmark labels replace the automatically built training data

on it, so these detectors fall outside the matched-generation protocol and are not run. Their evidence would also come from responses other than the one being scored.

Table 13 gives the complete dataset-marginal comparison, including representation controls that are separated from the canonical detector table in the main text.

H Negative results and claim precedence

Method development tested feature selection, extra reference statistics, broader collection scope, restricted layer bands, gradient-boosted trees, shallow MLPs, temporal windows, calibration variants, training-free scores, and cross-model transfer. Most did not improve the repeated-split endpoint. We summarize the decisions that directly changed the paper claim.

A direct distance-to-Benford detector did not provide a reliable training-free score. An early-only snapshot and concatenated intermediate snapshots did not improve the final accumulated trace. Random factorization matched grouping by stream identity, so stream identity does not explain the readout gain. Stacking DigitTrace with the answer-form control Style7 did not improve DigitTrace. Projected raw, a random projection of raw states to the trace's size, was statistically tied with the trace scored by single random partitions (Appendix F.3). Cross-model zero-shot transfer was near chance. Replay on the same GPU showed no latency advantage for the production collector.

Two of these choices are worth quantifying, because a simpler configuration is the obvious alternative. Restricting the sampled layers to a single middle layer costs 4.04 points, and keeping five middle layers still costs 1.19, so the layer grid is not redundant. On a 60-coordinate selected subset, replacing the L1 logistic readout with a shallow MLP costs 2.00 points and with gradient-boosted trees 2.61, so the linear readout was not the bottleneck there. Nonlinear readouts on the full trace were not run. Table 14 gives both grids. The accumulation-span grid is reported in Appendix A. Some of these grids point the other way, toward configurations that give up little AUROC for a smaller trace. Under the frozen protocol, one partition instead of five still reaches 76.50, with six probes per cell instead of 30 (Appendix F), and under leave-one-dataset-out transfer the joint readout matches the factorized one (Appendix F.13). Keeping seven of the sixteen statistics per view gives 76.63, although that subset was chosen after seeing results and is not pre-specified (Section 6.2). In the grid above, five middle layers reach 74.89, against 76.08 for all sampled layers, with 480

Table 13: Complete dataset-marginal AUROC for the proposed traces, protocol-matched controls, and all seven OpenHalDet adaptations. Columns average the three model cells. Compact headers use TQA for TruthfulQA, MATH for MATH-500, ARC-C for ARC-Challenge, GPQA for GPQA-Diamond, and ThmQA for TheoremQA. Summary rows are shown for reference and are not ranked.

Method	Open-ended		Math			Science		Mixed	Avg.
	CoQA	TQA	GSM8K	MATH	SVAMP	ARC-C	GPQA	ThmQA	
Proposed: leading-digit activation traces									
DigitTrace	73.88	70.15	84.53	80.65	89.63	79.07	**63.15**	72.38	76.68
Joint	73.12	68.20	82.65	79.46	88.56	78.04	62.31	70.91	75.41
Protocol-matched controls									
Full raw	**76.52**	71.65	**87.16**	**84.44**	**91.58**	**80.42**	58.01	**80.22**	**78.75**
Projected raw	66.72	66.27	86.02	82.48	89.74	76.79	59.15	78.66	75.73
Style7	64.53	57.25	78.72	79.97	85.40	64.66	53.00	71.94	69.43
OpenHalDet adaptations									
MIND	72.31	73.74	85.71	82.86	89.96	77.20	56.82	79.28	77.23
SEP	68.46	**75.97**	81.27	80.15	87.70	78.94	56.93	76.04	75.68
SAPLMA	74.93	75.05	82.89	81.43	87.51	71.74	53.98	76.77	75.54
PRISM	71.28	74.05	72.41	79.35	85.25	67.82	53.02	72.32	71.94
ICR Probe	69.49	62.29	75.49	74.38	86.68	72.06	57.15	68.33	70.73
Perplexity	65.65	59.72	59.44	66.32	79.34	65.45	54.05	55.21	63.15
HaloScope	55.05	53.66	67.65	68.49	71.80	63.02	53.97	59.96	61.70
Diagnostic summaries									
Mean of seven	68.17	67.78	74.98	76.14	84.04	70.89	55.13	69.70	70.85
Per-cell best[†]	75.25	75.97	85.71	83.23	89.96	80.58	58.97	79.28	78.62

[†]Per-cell best is not one detector. In each model–dataset cell, it takes whichever adaptation has the highest test AUROC there, for example SAPLMA on Qwen3-8B/CoQA and MIND on Qwen3-8B/GSM8K. Choosing the method this way requires the test results, so the row shows the best that any adaptation reaches in each cell, not a method one could deploy.

Table 14: **Two development grids behind fixed design choices.** Mean AUROC over the same 24 model–dataset cells, with 20 seeds and stratified 60/20/20 splits. The grids use the development feature pipeline rather than the frozen protocol, so the values are descriptive and not comparable with Table 1. The Coordinates column counts columns of the shared padded feature table. The readout-family grid uses 60 correlation-selected coordinates (20 digit frequencies and 40 moments) rather than the full trace.

Grid	Variant	Coordinates	Mean AUROC
Sampled layers	All sampled layers, as used	864–960	**76.08**
	Five middle	480	74.89
	Four middle	384	74.44
	Three late	288	74.15
	Three early	288	72.34
	One middle layer	96	72.04
Readout family	L1 logistic, as used	60	**76.12**
	Shallow MLP	60	74.12
	Gradient-boosted trees	60	73.51

coordinates instead of 864 to 960. The concatenated snapshots are not among these options: they double or triple the dimension and still score below the single accumulated summary.

Two evaluation corrections also reversed earlier conclusions. Repeating 50 shared splits removed several single-split rank claims. Expanding the raw-probe regularization grid raised raw performance by 2.41 points and made it stronger than DigitTrace on average. These results define the final scope: a compact supervised feature construction with a distinct temporal trace, not a training-free, early-stop, lower-cost, or strongest detector.

Exploratory variants were selected after many trials, so their intervals are descriptive and do not control the family-wise error rate. The DigitTrace main comparison, subset confirmation, and cost benchmark were run under written frozen protocols and were kept whatever the direction of their results.

I ADDITIONAL IMPLEMENTATION DETAILS

The evaluated model identifiers are Qwen/Qwen3-8B, Qwen/Qwen3-14B, and meta-llama/Llama-3.1-8B-Instruct. The factorized-readout partition seeds are $\{1729, 2718, 3141, 5772, 7919\}$. The 50 outer splits use seeds 0 to 49 with the stratified `train_test_split` of scikit-learn 1.7.2. Each split holds out 40% of the rows and halves them into validation and test rows. Every probe's inner 75/25 split uses seed 42. Regularization grids, feature dimensions, prompts, and evaluation rules are specified in the main paper and preceding appendices. Exact historical checkpoint revisions were not retained for every model. The dated snapshot behind the GPT-4o-mini alias is also unavailable, and Appendix C describes the judge protocol. All activation collection and the cost measurements ran on one NVIDIA RTX PRO 6000 Blackwell Server Edition GPU with 96 GB of memory and driver 580.95.05. The software stack was Python 3.10.20, PyTorch 2.11.0 with CUDA 12.8, and Transformers 5.5.3. One model is loaded at a time, and probe fitting and evaluation are CPU-only.

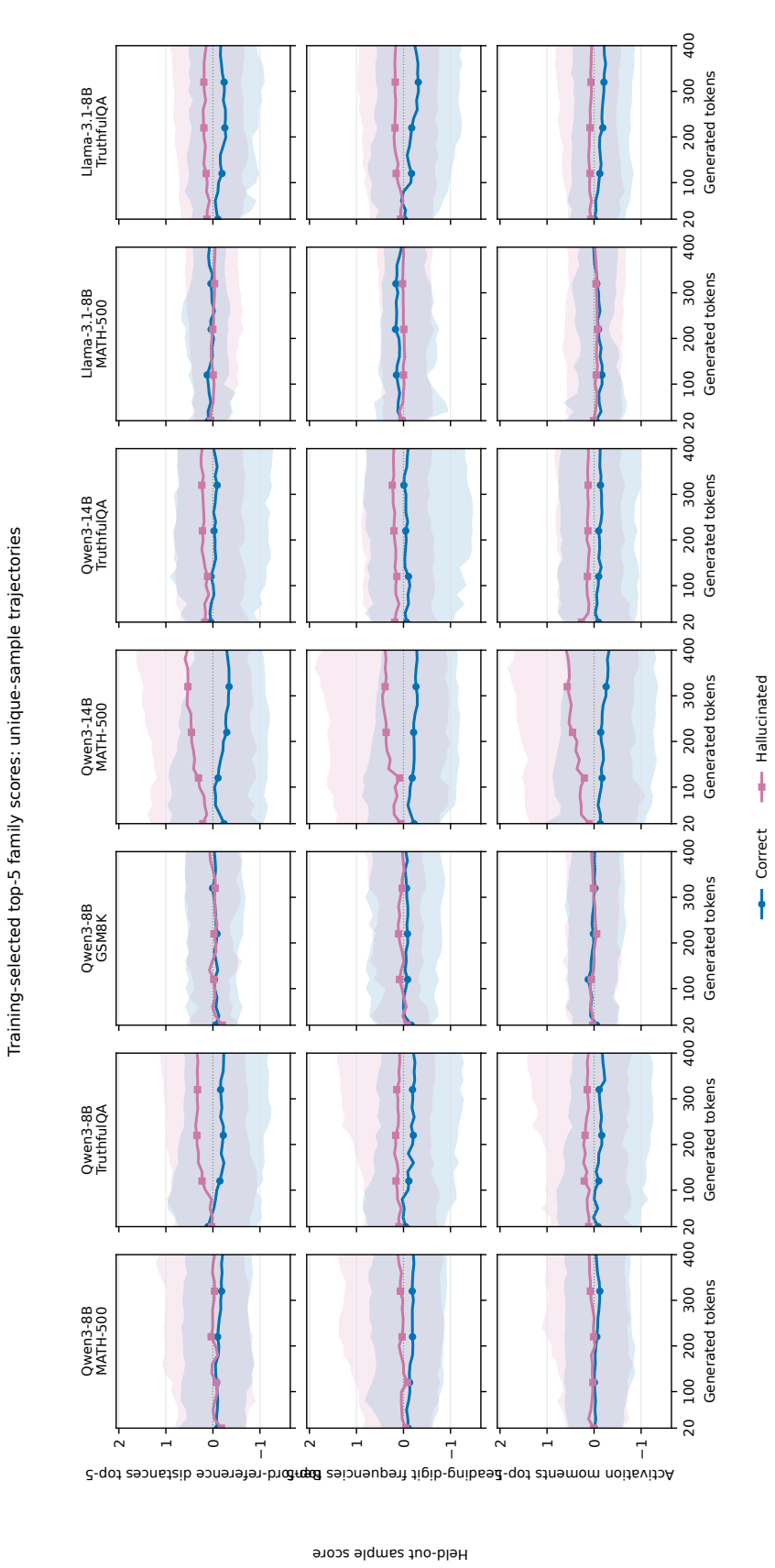

Figure 18: Dense held-out trajectories of training-selected top-five family scores for every available model–dataset cell. Columns are the seven fixed cohorts and rows are the distance, digit, and moment top-five scores. For each cutoff, repeated held-out scores are first averaged within each unique sample. Lines are class medians and shading gives the 10th–90th percentile across unique samples, not a confidence interval. Feature selection and direction alignment use only the outer-training partition at cutoff 400. The formal split-level AUROCs reproduce the frozen analysis at 20, 100, and 400 tokens.

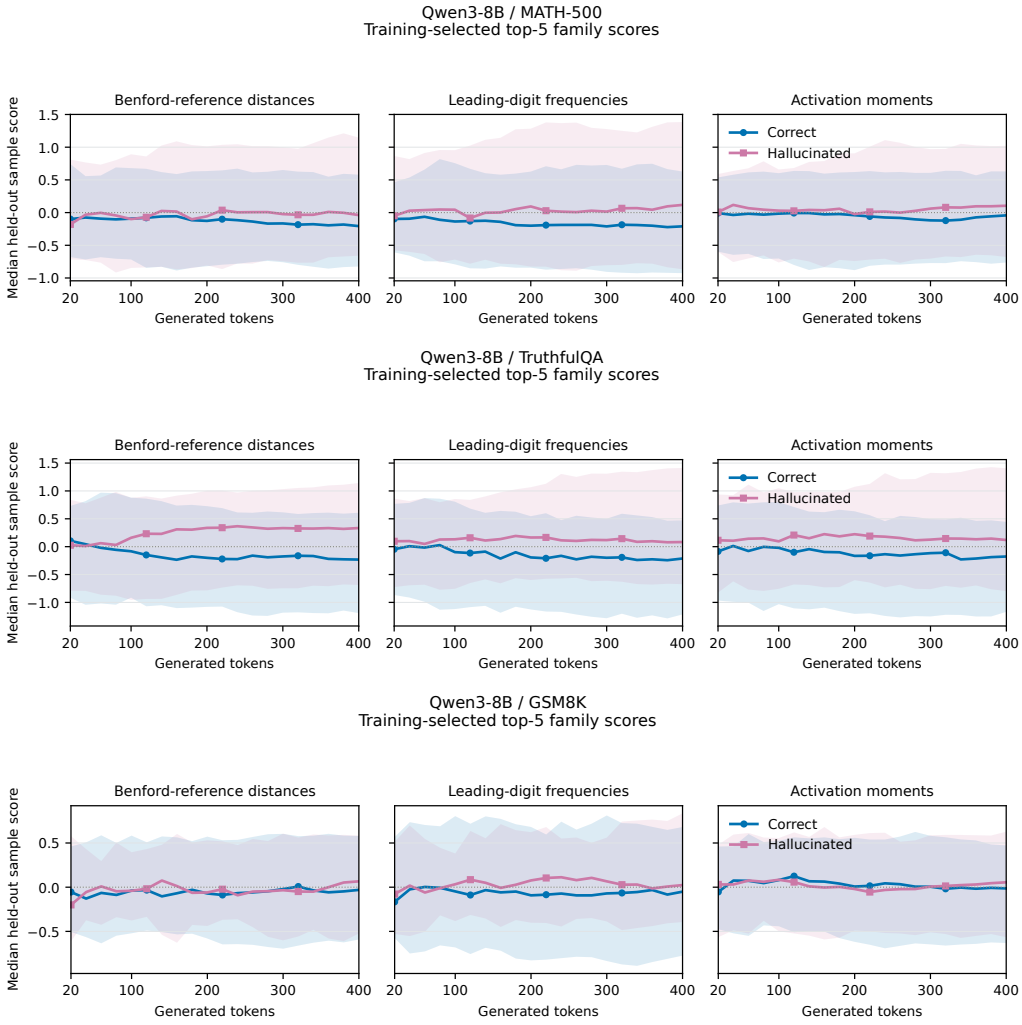

Figure 19: Qwen3-8B held-out trajectories of training-selected top-five family scores for MATH-500, TruthfulQA, and GSM8K. Each panel retains the same class colors, sample-quantile bands, and frozen cutoff-400 feature selection as Figure 18.

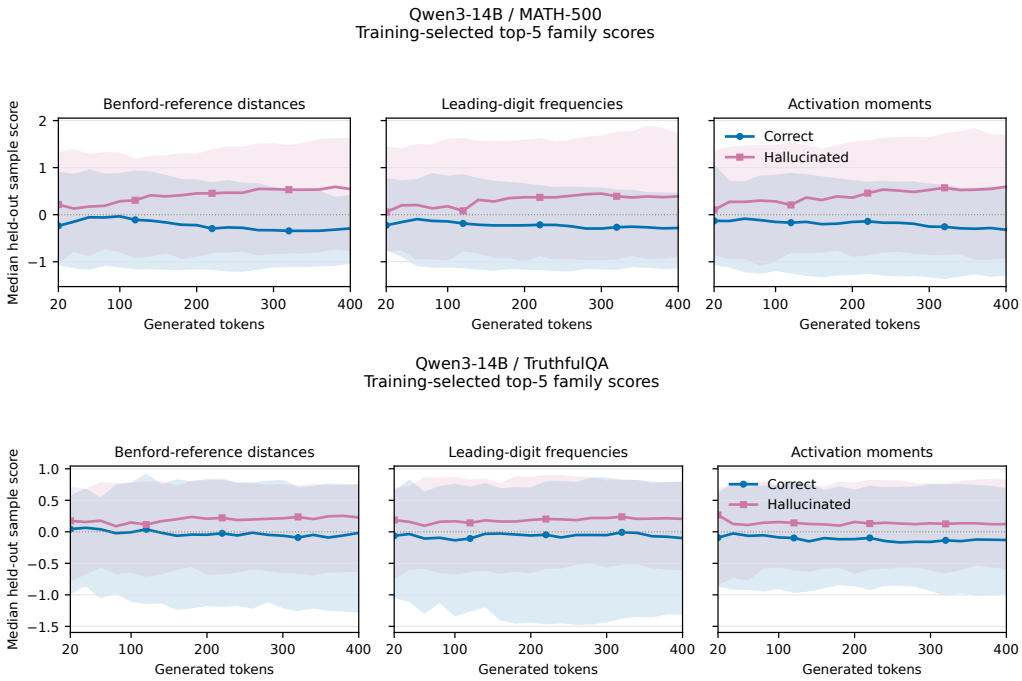

Figure 20: Qwen3-14B held-out trajectories of training-selected top-five family scores for MATH-500 and TruthfulQA. On MATH-500, the median separation widens in all three families.

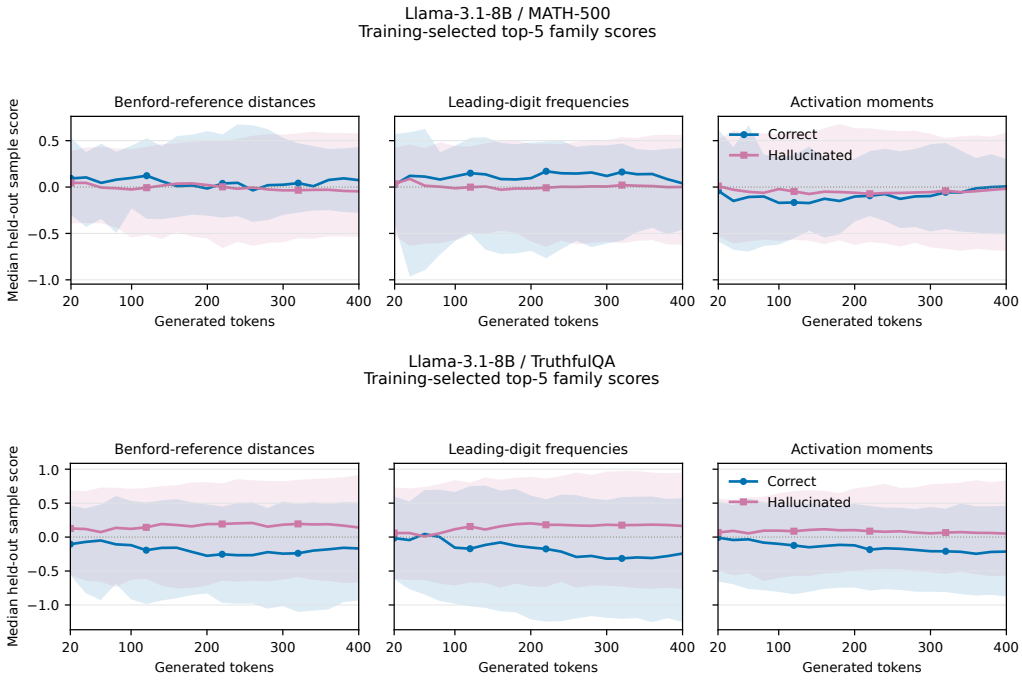

Figure 21: Llama-3.1-8B-Instruct held-out trajectories of training-selected top-five family scores for MATH-500 and TruthfulQA. On MATH-500, the direction selected in training does not carry over to held-out answers, and the class medians stay close.

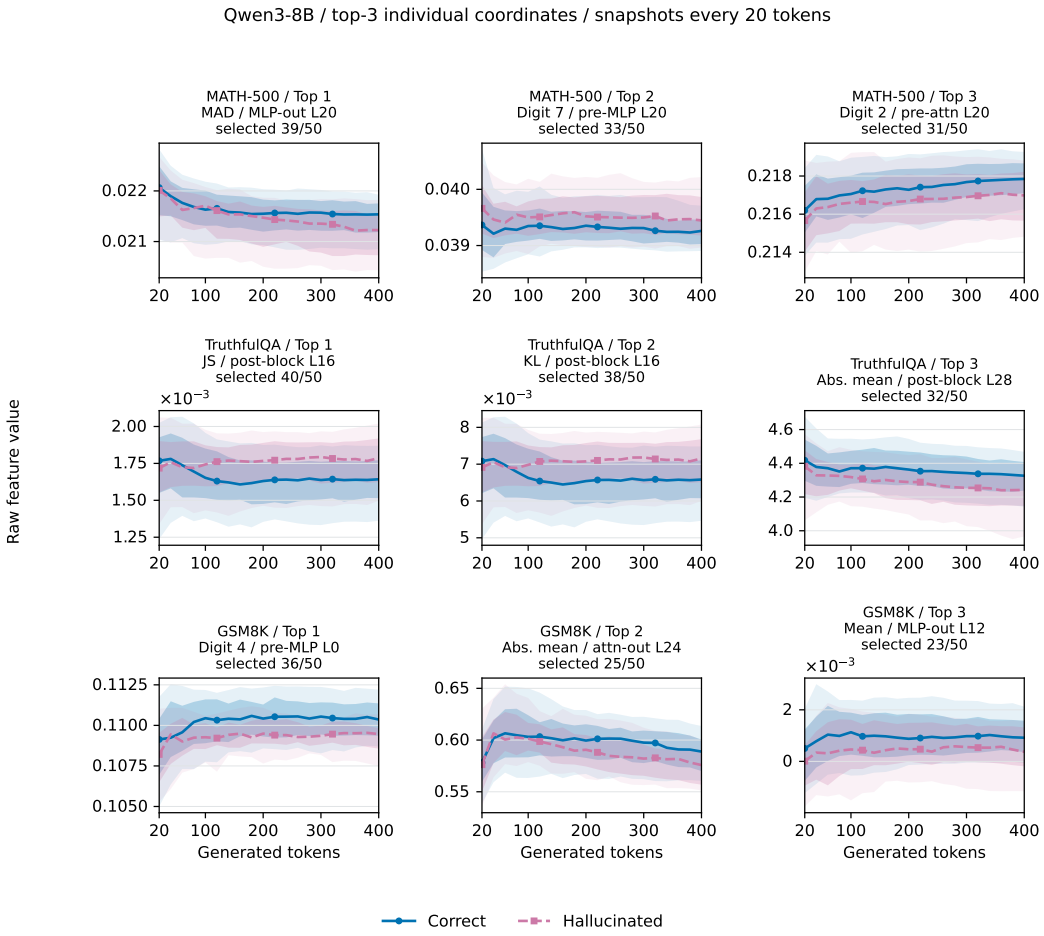

Figure 22: Qwen3-8B individual-coordinate trajectories for MATH-500, TruthfulQA, and GSM8K. Columns are the three coordinates most frequently selected across 50 outer-training partitions. Snapshots are stored every 20 generated tokens. All three selected coordinates are shown for every cell.

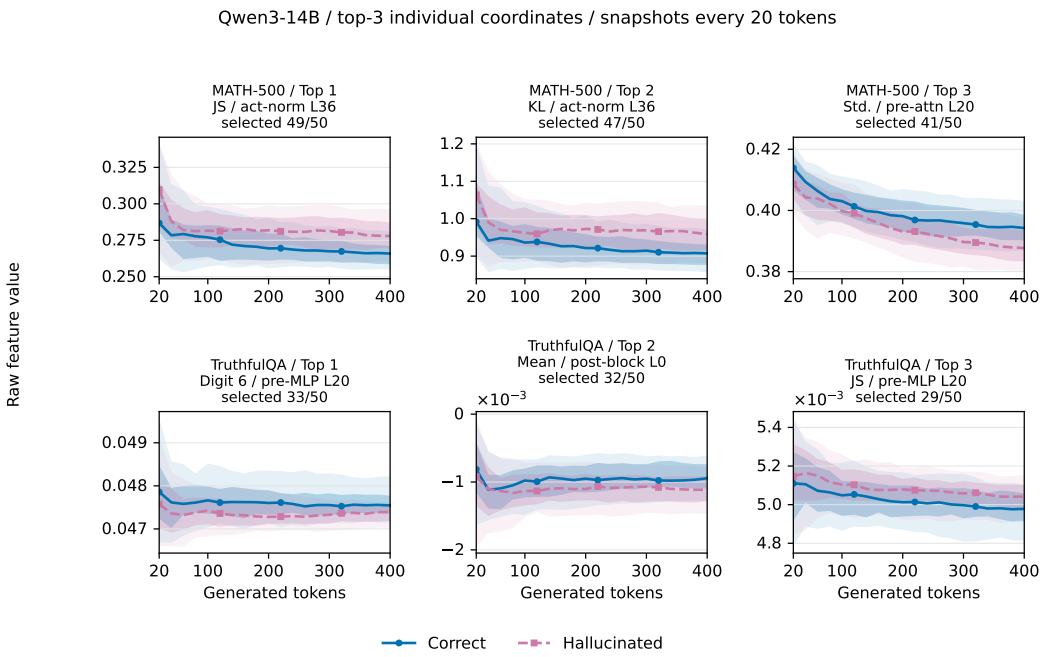

Figure 23: Qwen3-14B individual-coordinate trajectories for MATH-500 and TruthfulQA under the same training-only selection and sample-quantile conventions as Appendix Figure 22.

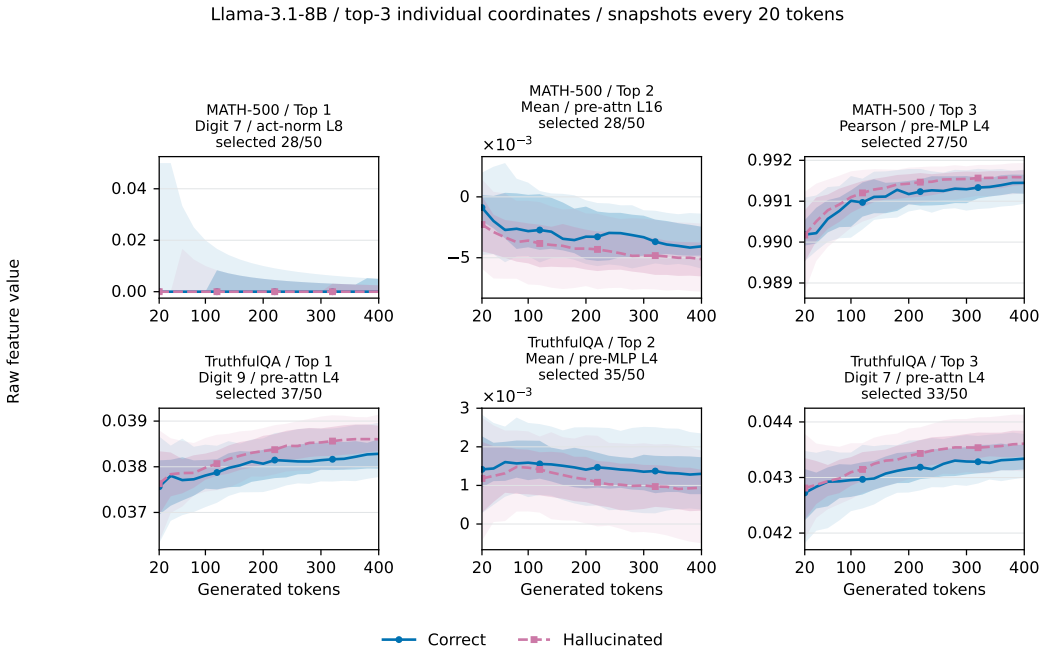

Figure 24: Llama-3.1-8B-Instruct individual-coordinate trajectories for MATH-500 and TruthfulQA under the same training-only selection and sample-quantile conventions as Appendix Figure 22.

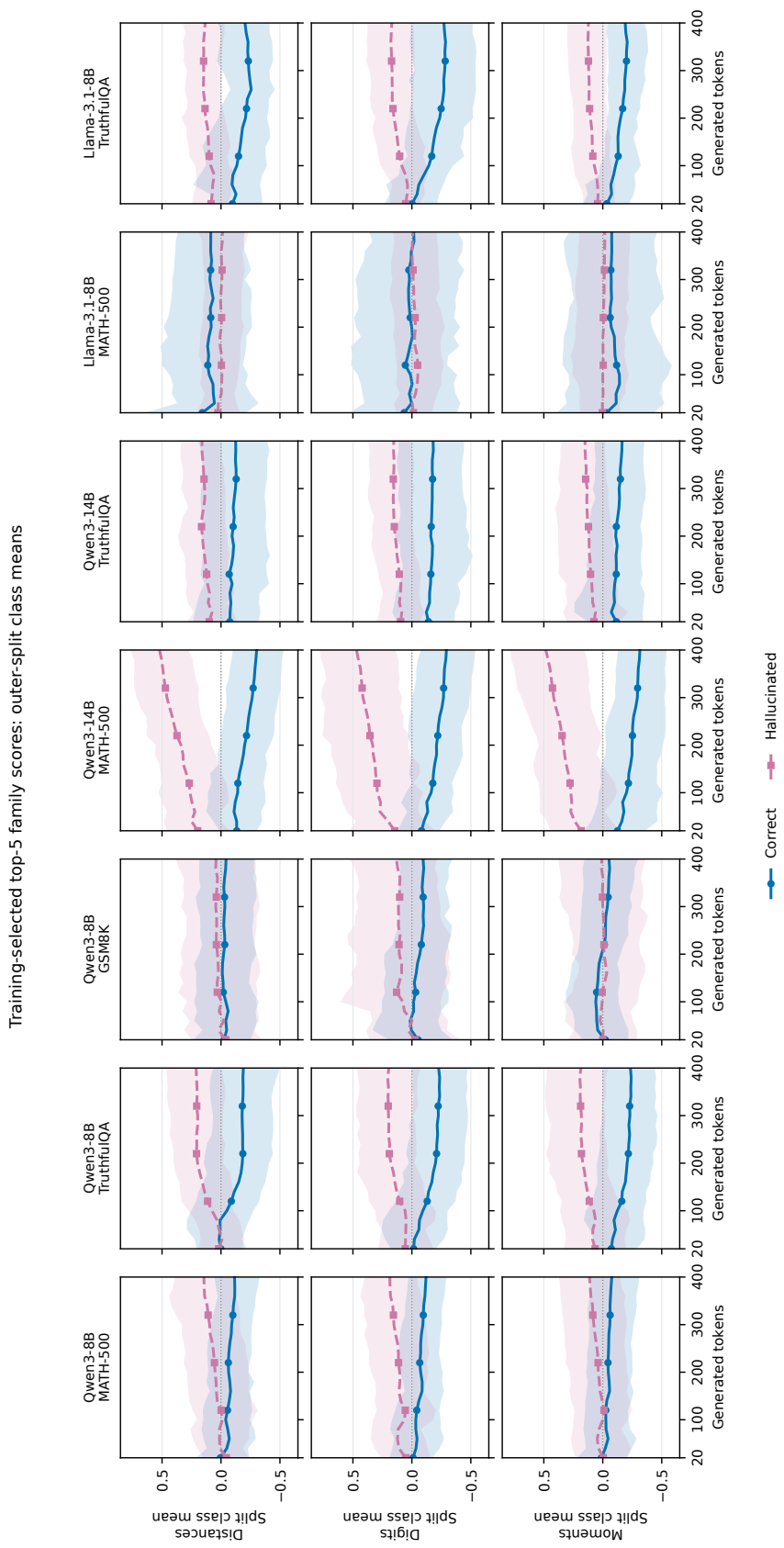

Figure 25: **Class-mean trajectories expose persistent correctness-related directions.** All seven fixed cohorts and three feature families are shown. Each line averages held-out class means over 50 outer splits. Shading is the 10th–90th percentile of those means, not a sample range or confidence interval. Top-five feature selection and direction alignment use only outer-training data at 400 tokens. Snapshots occur every 20 tokens. Figure 18 shows the corresponding individual-sample distributions.

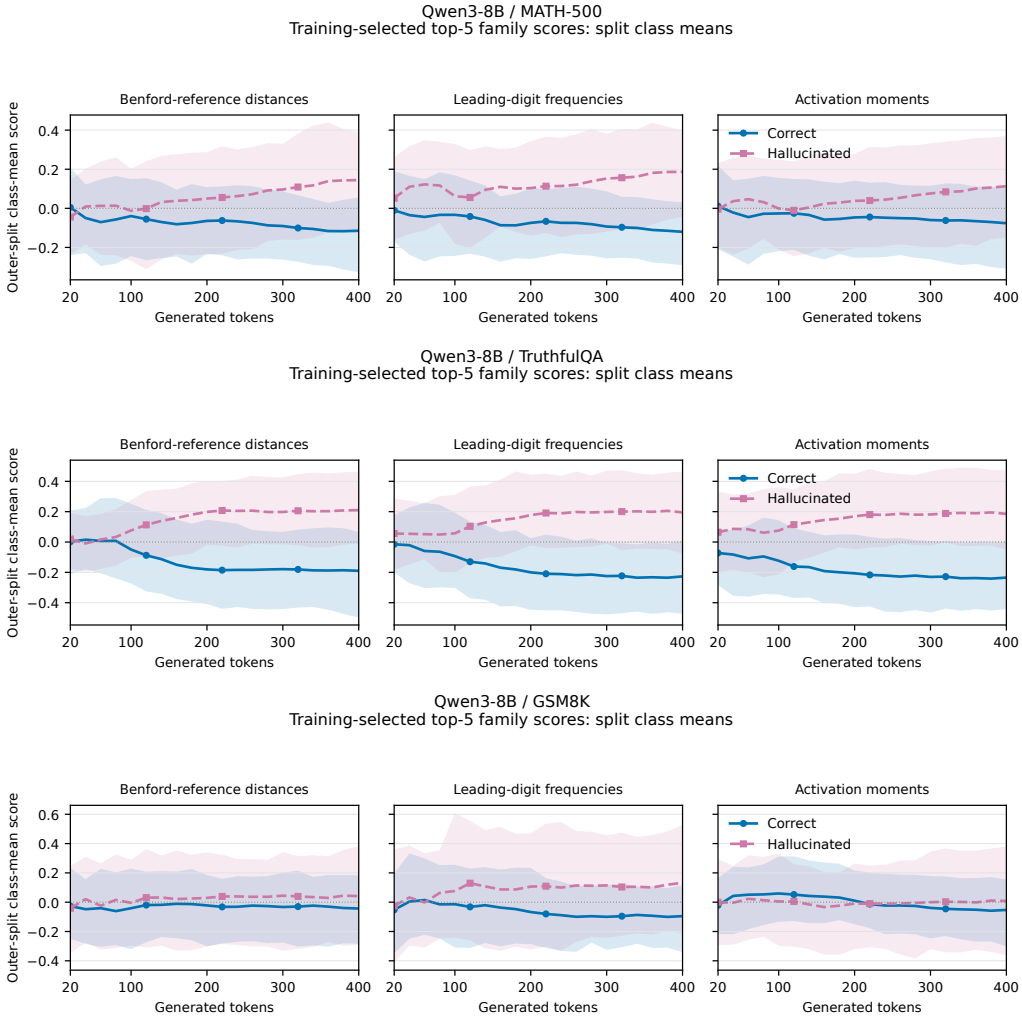

Figure 26: Qwen3-8B class-mean trajectories for all three fixed cohorts: MATH-500, TruthfulQA, and GSM8K. Lines and bands follow Figure 25. Compare the sample-level view in Figure 19.

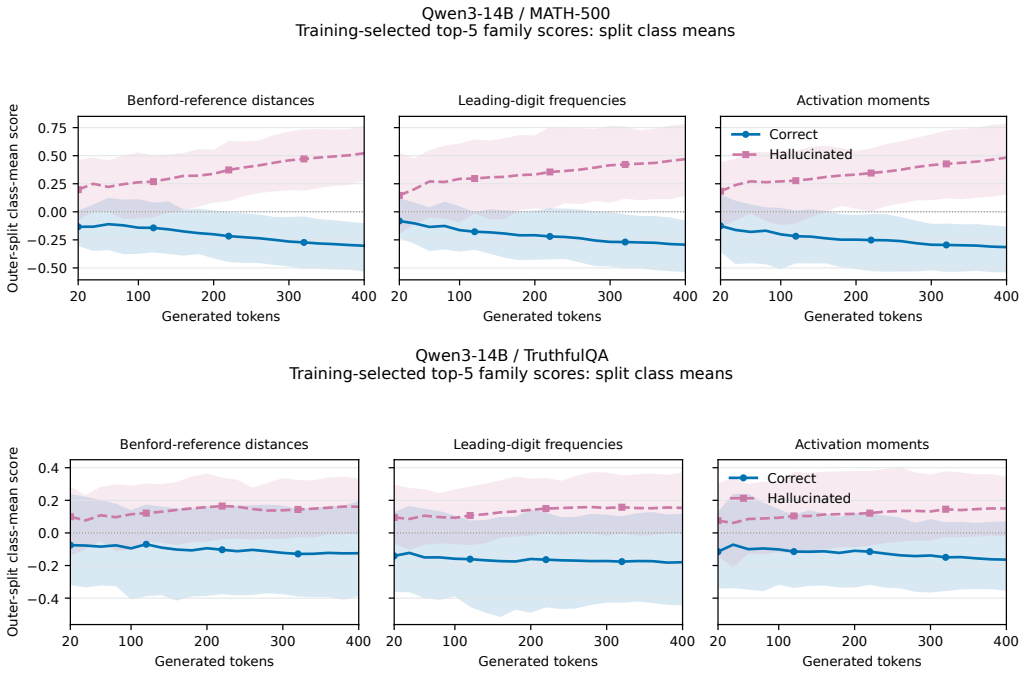

Figure 27: Qwen3-14B class-mean trajectories for MATH-500 and TruthfulQA. On MATH-500, late-stage class-mean bands separate in all three families. Figure 20 shows the corresponding sample distributions. Conventions follow Figure 25.

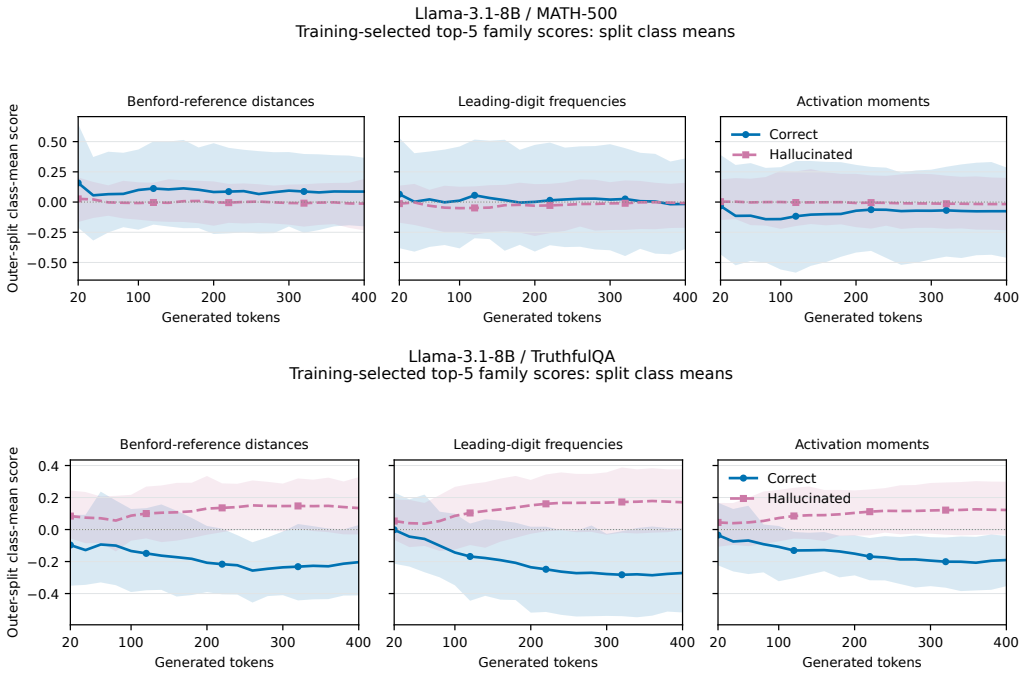

Figure 28: Llama-3.1-8B-Instruct class-mean trajectories for MATH-500 and TruthfulQA. Conventions follow Figure 25. Figure 21 provides the sample-level comparison.

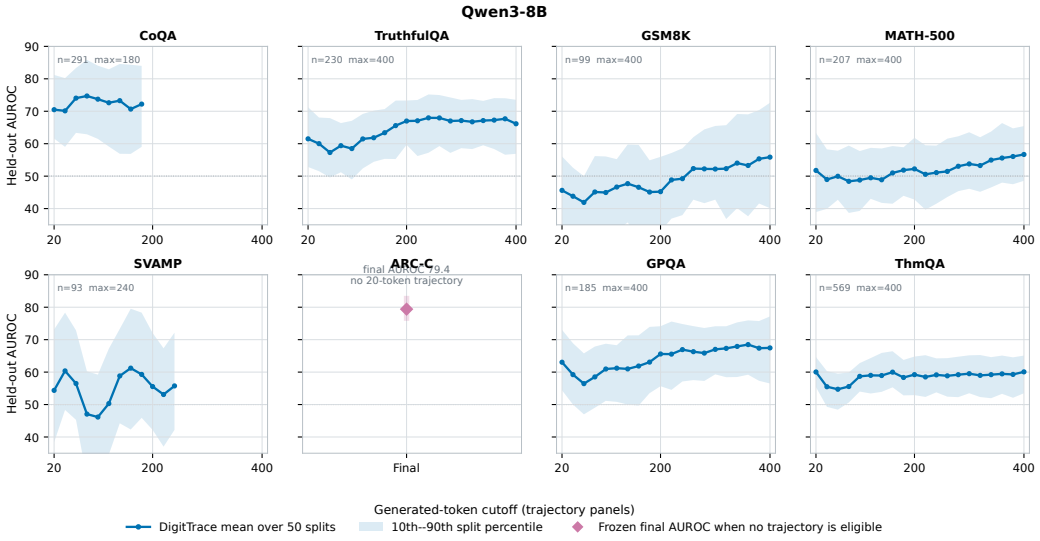

Figure 29: **DigitTrace trajectories for Qwen3-8B across all eight benchmarks.** Blue lines are mean held-out AUROC over 50 splits, with 10th–90th split percentiles. Each curve follows one fixed cohort: the answers that reach its displayed final cutoff. The ARC-C panel has no eligible 20-token trajectory and therefore shows only the frozen final-answer AUROC as a pink diamond. No missing trajectory is interpolated.

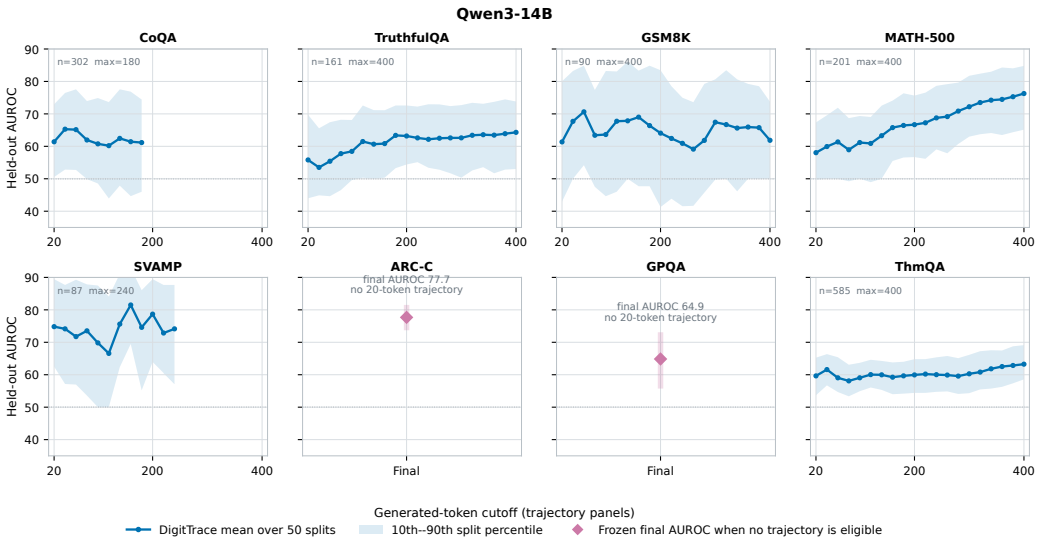

Figure 30: **DigitTrace trajectories for Qwen3-14B across all eight benchmarks.** Curves and bands follow Figure 29. The CoQA curve comes from teacher-forced replay of the frozen answers (see text). GPQA and ARC-C answers generally end before the first 20-token snapshot, so those panels show frozen final-answer AUROC as pink diamonds and do not contribute to the 19-cell trajectory-growth summary.

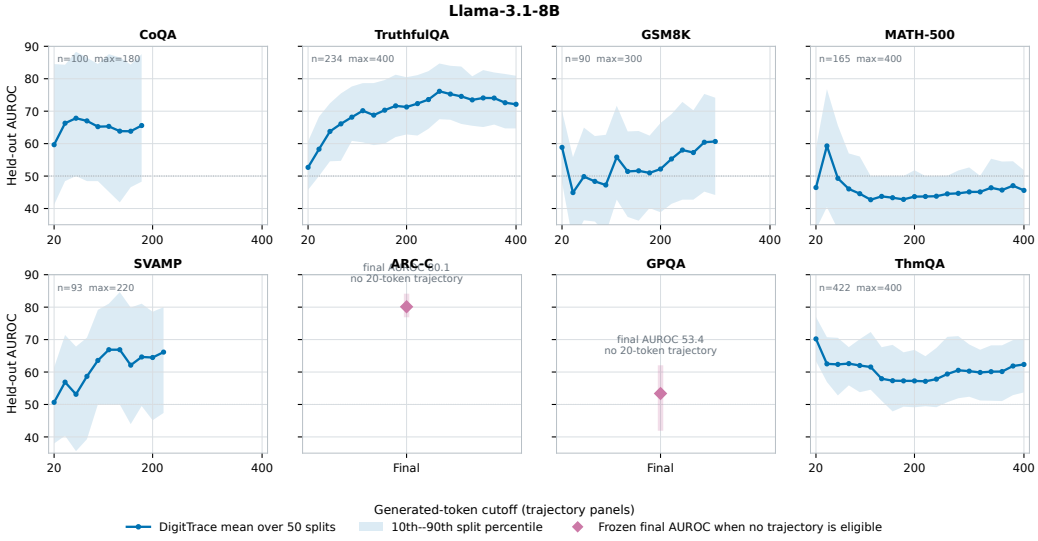

Figure 31: **DigitTrace trajectories for Llama-3.1-8B-Instruct across all eight benchmarks.** Curves and bands follow Figure 29. The CoQA curve comes from teacher-forced replay of the frozen answers. GPQA and ARC-C answers end before the first 20-token snapshot, so pink diamonds report frozen final-answer AUROC, as in Figure 30.