Mixture of Experimenters: Controlled Internal Experiments as a Label-Free Correctness Signal

PIATRA . INSTITUTE

June 2026

Abstract

The word "expert" in mixture-of-experts is a fossil. A contemporary expert is a routed feed-forward subnetwork selected for conditional compute, not an agent that runs experiments. We take the older sense of the word seriously and ask what it would mean for a language model to run an experiment on itself. The recent wave of test-time activation steering does not qualify: steering pushes the residual stream toward a chosen answer, which presupposes the conclusion. A genuine experiment is answer-agnostic. It intervenes on the evidence a claim depends on, predicts how the output should respond if the claim is grounded, and reads off the discrepancy. We make this concrete as a *controlled internal experiment*: a targeted intervention on the evidence bound to an answer, paired with a matched control intervention on evidence that should be irrelevant, whose control-corrected differential measures whether the answer is causally grounded. A "mixture of experimenters" is a battery of such experiments of different types, aggregated into a single label-free correctness signal. On an in-context retrieval task with a small transformer whose ground truth is known by construction, in a regime of controlled uncertainty, the aggregated signal separates correct from incorrect answers with AUROC 0.967 for a three-experiment mixture (0.956 for two), against 0.786 for negative entropy, 0.784 for logit margin, and 0.602 for a logistic correctness probe trained on labelled activations. The same construction transfers to Pythia-160M on in-context retrieval, where it reaches AUROC 0.841 against 0.719 for entropy, and it survives identifying the evidence by causal attribution rather than being given it, though the margin then narrows with attribution quality. The matched control is necessary where the intervention has a non-specific component (ablation, where it adds AUROC $+0.016$, CI95 $[+0.010, +0.022]$) and inert where the intervention is intrinsically specific (a counterfactual swap, $+0.000$), so a control matters exactly when there is generic perturbation-sensitivity to cancel. The contribution is a new primitive, an experiment rather than a steer, and the finding that whether an answer survives intervention on its own evidence is a far stronger label-free signal of correctness than how confident the model is.

1. Introduction

A sparse mixture-of-experts routes each token to a feed-forward subnetwork to scale parameters at fixed compute (Shazeer et al., 2017; Fedus, Zoph & Shazeer, 2022). The object the router selects is a compute path. It does not probe and it does not experiment. The term is a fossil of an older intuition: an experimenter intervenes on a system, predicts what the intervention should do under a hypothesis, measures what it did, and updates.

The recent literature has made the intervention half of that loop routine. Inference-time intervention and activation addition shift the residual stream along a learned direction toward a desired behaviour (Li et al., 2023; Turner et al., 2023), and a 2025-2026 wave makes the choice of steer adaptive: ATLAS gates test-time latent steering with a trained verifier (Nguyen et al., 2026), CAST conditions steering on a learned input signature (Lee et al., 2025), and sparse activation steering moves the intervention into an autoencoder basis (Bayat et al., 2025). All of these *steer*: they push the state toward a target the method has already chosen. Pushing toward a conclusion is not an

experiment. It presupposes the answer that an experiment is meant to put at risk.

A separate line estimates whether to trust an answer after the fact. Perturbation of internal states and a trained classifier predict correctness (CCPS; Khanmohammadi et al., 2025); counterfactual interventions on hidden states detect unfaithful explanations (CausalGaze; Kong et al., 2026); neighbourhood consistency measures the stability of an answer under paraphrase (Xu et al., 2026). These come closer to an experiment, but each applies a single intervention without a control, so none can separate a fragile-because-wrong answer from one that is merely in a generically perturbation-sensitive region.

This paper defines the missing primitive and shows it is a strong signal. An experiment must be answer-agnostic, and it must have a control. A *controlled internal experiment* intervenes on the evidence an answer depends on (the target), runs a matched intervention on evidence that should be irrelevant (the control), and reads the control-corrected response of the output. If the answer is causally grounded in its evidence, the target intervention moves it and the control does not; if the answer is a guess, neither moves it as the grounded hypothesis predicts. The differential is computed without the correct answer. A mixture of such experiments, of different types, aggregates into a single label-free estimate of whether an answer is grounded, and grounding turns out to track correctness far better than confidence does.

2. Related work

Steering and its selection. Activation addition and inference-time intervention establish that a residual-stream shift along a learned direction changes behaviour (Turner et al., 2023; Li et al., 2023). Recent methods make the shift adaptive: a trained verifier decides the steering strength per step (Nguyen et al., 2026), a condition vector gates whether to steer at all (Lee et al., 2025), and sparse-autoencoder coordinates localise the shift (Bayat et al., 2025). These optimise task behaviour by moving toward a target. The present work moves in the opposite epistemic direction: it intervenes to test a claim, not to impose one, and its output is a correctness signal rather than a steered generation.

Confidence and faithfulness signals. Token-level confidence (entropy, logit margin) and self-reported confidence are weak and miscalibrated on exactly the cases that matter, confident errors (Kadavath et al., 2022). Perturbation-stability methods train a classifier on the response of hidden states to noise (Khanmohammadi et al., 2025); counterfactual and consistency methods probe faithfulness (Kong et al., 2026; Xu et al., 2026). Every one of these applies a single intervention. The contribution here is the matched control and the answer-agnostic, grounding-testing form of the intervention, and the demonstration that the resulting signal beats both confidence and a supervised probe.

Mechanism. The interventions are activation and token edits of the kind used in activation patching and causal tracing (Meng et al., 2022), whose own methodology stresses that the choice of metric and baseline governs what a patch means (Zhang & Nanda, 2024). The substrate discussion draws on sparse autoencoders (Bricken et al., 2023; Templeton et al., 2024) and the reading of

intermediate states as predictions in the tuned lens (Belrose et al., 2023). The real-model results use Pythia-160M (Biderman et al., 2023).

3. Controlled internal experiments

Let a model answer a query by a distribution p_0 over a fixed answer set, with $\hat{y} = \arg\max_a p_0(a)$. An experiment is a tuple

$$E = (T,\, C,\, \rho,\, m),$$

a targeted intervention T on the evidence the answer should depend on, a matched control intervention C on evidence that should be irrelevant, a prediction ρ of how the output should respond under the hypothesis that the answer is grounded, and a measurement m of the realised response. The signal of a single experiment is the control-corrected response,

$$g(E) = m(p_T, \rho) - m(p_C, \rho),$$

where p_T and p_C are the output distributions after the target and control interventions. The control cancels the component of the response that any intervention of that kind would produce, which is what an uncontrolled probe cannot remove.

For in-context retrieval, where a query key is bound to a value in context, three experiments are immediate, two causal-dependence tests and one invariance test.

Experiment	Target T	Control C	Prediction ρ	Signal g
swap	overwrite the queried key's value with a fresh token v'	overwrite a non-queried value with v'	a grounded answer becomes v'	$(p_T(v') - p_0(v')) - (p_C(v') - p_0(v'))$
ablate	corrupt the queried key's value	corrupt a non-queried value	a grounded answer's own probability collapses	$(p_0(\hat{y}) - p_T(\hat{y})) - (p_0(\hat{y}) - p_C(\hat{y}))$
invariance	reorder the (intact) pair blocks	(none; a one-armed stability test)	a grounded answer is unchanged	$-\lvert p_{\text{perm}}(\hat{y}) - p_0(\hat{y})\rvert$

No signal uses the correct answer. All use only the input structure, which token holds the queried key's value, information present in the prompt. The swap and ablate experiments share the control construction of Section 3; the invariance experiment is one-armed, a stability test with no separate control, included to show that experiments of a different shape compose. The mixture of experimenters aggregates the standardised single-experiment signals, $g_{\text{mix}} = \sum_k z(g_k)$, where z is the z-score across the evaluation set, which keeps the aggregate label-free. The signals are scored afterwards by the area under the ROC curve (AUROC) for discriminating correct from incorrect

answers, with paired bootstrap confidence intervals over examples; labels enter only at this scoring step.

The natural baselines are the label-free confidence signals, negative entropy and logit margin of p_0, and, as a supervised reference, a logistic probe $P(\text{correct} \mid h)$ trained on labelled answer-position activations.

4. The toy: an oracle-grounded evaluation

A two-layer, 64-dimensional, four-head decoder-only transformer is trained on a synthetic in-context retrieval task: a sequence of key-value pairs followed by a query key, with the value bound to the queried key as the target. The model learns the task to accuracy 1.000 without noise. Evaluation adds Gaussian noise of standard deviation 0.9 to the token embeddings, placing the model in a regime where it answers 0.834 of items correctly while the structure needed to recover the rest is still present. The known ground truth provides the control this whole literature lacks: it makes the AUROC ceiling well defined and lets the label-free signals be measured against the truth rather than against each other. The evaluation set is 5000 items.

Table 1 reports the discrimination AUROC of each signal.

Signal	AUROC	95% CI
mixture (swap + ablate + invariance)	0.967	$[0.961, 0.972]$
mixture (swap + ablate)	0.956	$[0.949, 0.962]$
ablate (controlled)	0.944	$[0.936, 0.951]$
ablate (uncontrolled)	0.928	$[0.919, 0.936]$
swap (controlled)	0.904	$[0.893, 0.915]$
swap (uncontrolled)	0.904	$[0.892, 0.915]$
invariance	0.755	$[0.733, 0.776]$
negative entropy	0.786	$[0.768, 0.803]$
logit margin	0.784	$[0.767, 0.801]$
supervised probe	0.602	$[0.580, 0.623]$

Table 1. Correct-versus-incorrect discrimination on the toy.

The two-experiment controlled mixture separates correct from incorrect answers at AUROC 0.956. It beats negative entropy by a paired CI95 of $[+0.150, +0.191]$ and logit margin by $[+0.152, +0.192]$. It beats the supervised probe, which uses labels, by $[+0.332, +0.377]$: the probe reaches only 0.602, because a readout trained on noisy activations to predict correctness does not generalise, the same fragility this construction sidesteps by asking a causal question of the model rather than reading a learned feature. Whether an answer survives intervention on its own evidence is a much stronger correctness signal than how confident the model is in it.

The mixture grows with the panel. The invariance experiment is a weaker signal on its own, AUROC 0.755, because order-stability is necessary but not sufficient for grounding. Added to the

mixture it still helps: the three-experiment mixture reaches 0.967, above the two-experiment mixture by a paired CI95 of $[+0.007, +0.015]$. A differently shaped experiment, a should-not-move test beside the two does-it-respond tests, contributes discrimination the others miss, which is the sense in which a mixture of experimenters is more than one.

The role of the control is precise. For the ablation experiment the control is necessary: the controlled signal beats the same intervention without a control by AUROC $[+0.010, +0.022]$, because corrupting any value token has a generic effect on the output that the control removes. For the swap experiment the control is inert, $[-0.004, +0.005]$, because swapping a value the answer does not use barely moves the output, so there is nothing for the control to cancel. A control matters exactly when the intervention has a non-specific component, and the mixture, which beats the best uncontrolled single experiment by $[+0.021, +0.036]$, inherits the benefit where it exists.

5. A real model: Pythia-160M

The toy supplies the oracle; a real model supplies the check that the mechanism is not an artifact of scale or of a hand-built model. Pythia-160M (Biderman et al., 2023) is evaluated on an in-context retrieval task in natural text, a list of single-token key-value pairs followed by a repeated query key, which the model completes. With six pairs the model answers 0.403 of 400 items correctly, an intermediate accuracy that leaves both classes well populated. The identical controlled experiments are run by editing the queried and control value tokens.

Signal	AUROC	95% CI
controlled mixture	0.841	$[0.800, 0.879]$
ablate (controlled)	0.816	$[0.774, 0.858]$
ablate (uncontrolled)	0.811	$[0.767, 0.852]$
swap (uncontrolled)	0.770	$[0.722, 0.816]$
swap (controlled)	0.757	$[0.709, 0.805]$
negative entropy	0.719	$[0.668, 0.770]$
logit margin	0.719	$[0.667, 0.767]$

Table 2. Correct-versus-incorrect discrimination on Pythia-160M.

The controlled mixture reaches AUROC 0.841 and beats negative entropy by a paired CI95 of $[+0.067, +0.176]$ and logit margin by $[+0.069, +0.177]$. The grounding experiment transfers: on a real pretrained model, whether the answer tracks an edit to its evidence discriminates correctness substantially better than the model's confidence. The control's marginal contribution over the best uncontrolled probe is small and not significant at this sample size, $[-0.011, +0.070]$, consistent with the toy finding that the control's value is concentrated in the ablation experiment and modest in absolute terms.

Finding the evidence rather than being given it. The results above use the input structure to locate the queried binding. Removing that assumption, the evidence token is identified by causal

leave-one-out, ablating each value token in turn and taking the one whose removal most reduces the answer's probability, with no knowledge of which binding was queried, and the swap experiment is then run on the attributed token (attribution by ablation, test by swap, so the test is not circular). On Pythia-160M this attribution recovers the queried binding on 0.488 of items, against a chance rate of $1/6$. The attributed swap signal discriminates correctness at AUROC 0.683 $[0.625, 0.739]$, well above chance but below both the known-evidence swap (0.757, a significant gap of $[-0.129, -0.020]$) and entropy (0.719). The signal survives self-attribution, but the margin that lets it beat confidence requires the evidence to be located accurately, which a small model supplies only imperfectly. The bottleneck for generalisation is evidence attribution, not the experiment: where the evidence is known the experiment is strong, and self-attribution is the lossy step.

6. Discussion

Confidence and grounding come apart precisely on the cases that matter. Entropy and margin measure how peaked the model's distribution is, which is high both when the model is right and when it is confidently wrong, so they discriminate correctness only weakly (AUROC near 0.78 on the toy, 0.72 on Pythia). A controlled experiment measures something else: whether the answer is a function of the evidence it should depend on. A confident wrong answer that does not track its evidence is exactly what the experiment exposes and what confidence misses. This is why a label-free causal probe beats not only the label-free confidence signals but a supervised correctness probe.

The control is the element that distinguishes an experiment from a perturbation. A single intervention conflates two reasons an output might move, that the answer depended on what was changed, and that this region of the network is sensitive to any change of that kind. The matched control estimates and removes the second. Its value is therefore conditional, large for interventions with a strong generic component such as ablation, negligible for interventions that are already specific such as the counterfactual swap. The honest statement is not that every experiment needs a control, but that a control is required exactly when the intervention is non-specific, and that running it is the safe default because specificity cannot be assumed in advance.

The mixture is the "experimenters" of the title. Each experiment type is a distinct method of interrogation, the swap a positive test (does the answer follow when the evidence is replaced) and the ablation a negative test (does the answer fall when the evidence is removed). Aggregating standardised signals across types beats any single experiment, by $[+0.021, +0.036]$ on the toy, because the types fail on different items. This is the sense in which a mixture of experimenters is more than one experimenter: not a routing of compute, but a panel of independent interrogations of the same claim.

7. Limitations

The experiments require knowing which evidence an answer should depend on. The generalisation result of Section 5 makes this concrete and quantifies its cost: when the evidence is found by causal leave-one-out rather than given, the signal survives but its advantage over confidence erodes in proportion to the attribution quality, which on a small model is only 0.488. For unstructured questions the evidence would have to be identified more reliably, by stronger attribution or by the model's own citation of its sources; that identification step, not the experiment, is the main obstacle to a general-purpose deployment.

The evaluation is discrimination of correct from incorrect answers by AUROC, the right metric for a label-free signal, but a deployed abstention system also needs a threshold and a calibration, which are task-specific and not studied here. The real-model demonstration is a single small model on a single task family at intermediate accuracy; the toy supplies the oracle and the controlled comparison, the real model supplies external validity, and neither is a benchmark across models, tasks, or scales. The mixture aggregates two experiment types by an unsupervised z-score sum; a richer aggregation, and additional experiment types such as paraphrase-invariance and re-derivation, are natural extensions the construction admits but that are not evaluated here.

8. Conclusion

A mixture of experimenters, read literally, is a panel of answer-agnostic experiments run inside the model. Each intervenes on the evidence a claim depends on, against a matched control, and reads whether the claim survives. The control-corrected, aggregated signal discriminates correct from incorrect answers with AUROC 0.956 on a toy with known ground truth and 0.841 on Pythia-160M, in both cases far above the confidence signals (entropy, margin) and above a supervised probe. The reason is that grounding and confidence are different quantities, and grounding is the one that tracks correctness. The control is what makes an intervention an experiment, and it earns its cost exactly when the intervention is non-specific. Extending the panel to richer experiment types and to settings where the relevant evidence must first be identified is the natural continuation.

All numeric claims are backed by per-experiment verification records and a claim ledger distributed with the paper; the trained toy model and the real-model evaluation are included, and all results are deterministic under fixed seeds.

References

- Bayat, R., Rahimi-Kalahroudi, A., Pezeshki, M., Chandar, S., & Vincent, P. (2025). Steering large language model activations in sparse spaces. arXiv:2503.00177.
- Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., & Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens. arXiv:2303.08112.
- Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., et al. (2023). Pythia: a suite for analyzing large language models across training and scaling. *ICML*.

- Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., et al. (2023). Towards monosemanticity: decomposing language models with dictionary learning. *Transformer Circuits Thread*.
- Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., et al. (2022). Language models (mostly) know what they know. arXiv:2207.05221.
- Khanmohammadi, R., Miahi, E., Mardikoraem, M., Kaur, S., Brugere, I., Smiley, C. H., Thind, K., & Ghassemi, M. M. (2025). Calibrating LLM confidence by probing perturbed representation stability. arXiv:2505.21772.
- Kong, L., Wu, L., Zhang, Y., Zhong, X., Wang, Z., Wang, Y., & Pan, Y. (2026). CausalGaze: unveiling hallucinations via counterfactual graph intervention in large language models. arXiv:2604.11087.
- Lee, B. W., et al. (2025). Programming refusal with conditional activation steering (CAST). *ICLR*; arXiv:2409.05907.
- Li, K., Patel, O., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Inference-time intervention: eliciting truthful answers from a language model. *NeurIPS*.
- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. *NeurIPS*.
- Nguyen, T., et al. (2026). ATLAS: adaptive test-time latent steering with external verifiers for enhancing LLMs' reasoning. arXiv:2601.03093.
- Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. arXiv:1701.06538.
- Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., et al. (2024). Scaling monosemanticity: extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.
- Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2023). Activation addition: steering language models without optimization. arXiv:2308.10248.
- Xu, H., Zhao, N., Yao, Y., Xu, W., Wang, H., Deng, X., Deng, S., Pan, J. Z., Chen, H., & Zhang, N. (2026). Illusions of confidence? Diagnosing LLM truthfulness via neighborhood consistency. arXiv:2601.05905.
- Zhang, F., & Nanda, N. (2024). Towards best practices of activation patching in language models: metrics and methods. *ICLR*.