J-FDCS: JACOBIAN FAULT DIAGNOSIS CONCEPT SPACES FOR INTERPRETING LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

When a language model diagnoses a machine from vibration signals, it remains unclear whether it represents the underlying fault or merely learns to produce the expected answer. Diagnostic accuracy alone cannot resolve this distinction. We introduce Jacobian Fault Diagnosis Concept Spaces (J-FDCS), a framework for examining fault-related representations in signal-grounded language models. Models receive vibration signals with equipment and task context and learn to answer with abstract codes. We analyze their hidden states using natural fault expressions. J-FDCS constructs expression-associated directions through the average Jacobian Lens or a full-sequence gradient formulation. It then removes their linear overlap with specified task-control and answer-code directions. The resulting low-dimensional space supports complementary tests of readability, semantic correspondence, signal dependence, cross-state stability, and local output effects. We systematically evaluate Qwen, Llama, and Mistral on five public rotating-machinery benchmarks and an in-house BW250 drilling-pump vibration dataset. The results distinguish strong linear class decodability from the variable readout of natural fault expressions. Fixed-space perturbations reveal input-responsive readouts, while cross-state transfer and retrospective BW250 interventions show variable semantic retention and local changes in diagnostic answer scores. Beyond predictive accuracy, J-FDCS provides a way to examine what fault-related information a model represents, which signals it depends on, and how changing it affects diagnosis.

1 INTRODUCTION

Large language models increasingly perceive and reason about the physical world as components of multimodal systems. Sensor signals, images, and other nonlinguistic inputs support tasks ranging from classification and prediction to decision support. However, a basic question remains unresolved: does a correct answer about an external state reflect a corresponding internal concept representation or a statistical input–output mapping? This distinction matters for understanding knowledge formation in multimodal systems and assessing their reliability.

Signal-grounded fault diagnosis provides a useful setting for studying how physical concepts enter language models. A model must infer an unobserved mechanical state from continuous sensor signals and express it through human-readable semantic labels. This task therefore connects physical states, signal representations, and linguistic concepts. Signal tokenization, feature alignment, and instruction tuning connect vibration signals to pretrained language models for fault classification, question answering, and report generation. However, an effective signal-to-answer mapping does not establish that the model internally represents the corresponding physical state as a concept.

Studying these representations requires distinguishing fault-name readout from fault-concept identification. Probes and lenses reveal class information in hidden states or map intermediate representations to readable words (Alain & Bengio, 2016; Belrose et al., 2023). The average-Jacobian method of Gurnee et al. (2026), the Jacobian Lens, identifies verbalizable directions through their influence on downstream computation and tests their effects through interventions. Fault diagnosis introduces an ambiguity: a fault name can describe a mechanical state and serve as the supervised answer.

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.

If a model learns to answer piston for a piston fault, reading piston before generation may reflect fault information, answer preparation, or both. We therefore test whether representations retain fault-related correspondence after controlling for answer vocabulary, respond to input signals, and affect diagnostic outputs.

To reduce this ambiguity, we separate the model's training answers from the fault expressions used for analysis. Models receive vibration signals and equipment/task context and learn to output abstract codes, while natural fault expressions are used to examine their hidden states. This separates the supervision target from the analysis vocabulary without excluding fault descriptions from the input context. We call the framework Jacobian Fault Diagnosis Concept Spaces (J-FDCS). Expression-associated directions are centered and adjusted for specified task-control and answer-code directions, yielding a low-dimensional space for readout and intervention. Within this space, we examine whether fault information is readable and whether the intended expression assignment is distinctive. We also test whether readouts depend on the signal, persist across model states, and affect diagnostic answers when changed.

Our main contributions are as follows.

1. We introduce a framework that separates training answers from analysis vocabulary. J-FDCS supervises diagnosis with abstract answer codes under specified equipment and task context and analyzes hidden states using natural fault expressions. This design separates the semantic expressions under examination from the supervised output targets, enabling analysis of fault-related representations beyond diagnostic accuracy.

2. We construct a jointly controlled fault-expression space for readout and intervention. The average Jacobian Lens and a separately defined full-sequence formulation produce expression-associated directions. We jointly remove their linear overlap with specified task-control and answer-code directions and construct geometry-matched readout controls. Readout and interventions within the same analysis space connect fault-expression correspondence to the local output effects of selected coordinates.

3. We provide a systematic evaluation on public benchmarks and an in-house pump dataset. We evaluate Qwen, Llama, and Mistral on public rotating-machinery benchmarks and our in-house BW250 drilling-pump vibration dataset. Semantic controls, input perturbations, and cross-state comparisons examine fault-related representations across tasks and training conditions. BW250 provides a focused setting for same-space interventions. The results distinguish class decodability, correspondence with natural fault expressions, and local output effects, showing why diagnostic performance alone cannot characterize these properties.

2 Related work

2.1 Signal-grounded language models for physical systems

Time-series patching and reprogramming connect numerical sequences to pretrained Transformers for forecasting (Jin et al., 2024). Vibration–language systems support signal-parameter identification and diagnostic question answering (Li et al., 2026). These tasks establish feasibility, but predictive accuracy alone does not identify the semantic content of internal representations. We examine representation evidence beyond task performance.

2.2 Readout of internal representations

Linear probes assess class decodability (Alain & Bengio, 2016), and vocabulary lenses reveal intermediate predictions (Belrose et al., 2023). Representation-similarity methods compare hidden-space geometry without requiring neuron-by-neuron correspondence (Kornblith et al., 2019). The average Jacobian Lens maps first-order effects on downstream computation to verbalizable directions and supports coordinate interventions (Gurnee et al., 2026). We distinguish our full-sequence implementation from the original operator. Our focus is the ambiguity introduced when the fault names under analysis also serve as supervised answers.

2.3 From readout to causal evidence

Causal abstraction and distributed alignment compare interventions on high-level variables and internal states (Geiger et al., 2025; 2024). Activation-patching conclusions depend on the metric and perturbation procedure (Zhang & Nanda, 2024). A subspace can change outputs through an alternative path without faithfully identifying normal computation (Makelov et al., 2024). J-FDCS therefore combines readout, semantic references, signal changes, transfer, and matched interventions, rather than treating any single test as sufficient evidence of a concept mechanism.

3 J-FDCS: A framework for analyzing fault-related representations

Figure 1 summarizes the framework. We first define code-supervised diagnostic models, then construct fault-expression directions and the jointly controlled space used for readout and intervention.

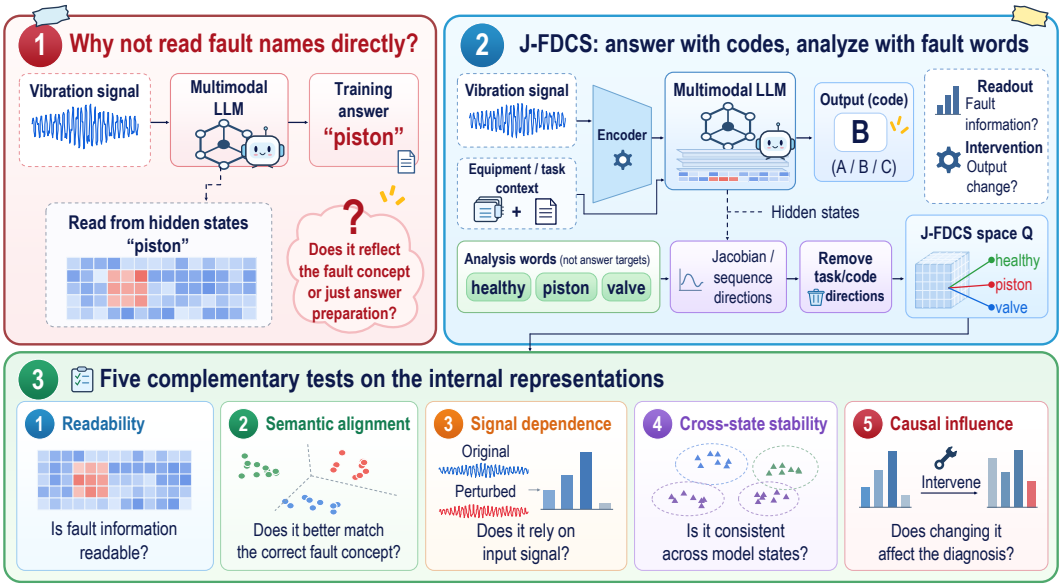


Figure 1: J-FDCS overview: code-supervised diagnosis, fault-expression space construction, and representation analysis.

3.1 Task design and model states

All diagnostic tasks use abstract codes as supervised answers and natural fault expressions as analysis targets. Three-class tasks use A/B/C, and four-class tasks use A/B/C/D. The number of class-to-code mappings is specified separately for each dataset. Within each configuration, equipment/task context remains fixed during training, direction fitting, and evaluation. CWRU and MCC5 include explicit fault descriptions, whereas the other prompts vary in detail. Separating answer codes from analysis expressions therefore does not remove all semantic background from the input (Appendix A).

A ResNet1D-18 encoder and model-specific projectors map a waveform window to 32 continuous embeddings, which precede the model's native chat prompt. The interface is pretrained on unlabeled waveforms from the parameter-training partition and then frozen. We use Qwen2.5-3B-Instruct (Yang et al., 2024), Llama-3.2-3B-Instruct (Meta, 2024), and Mistral-7B-Instruct-v0.3 (Mistral AI, 2024). The task-unadapted reference (S0) has the trained interface but no diagnostic task adapter; code-supervised states (S1) receive correct-code QLoRA supervision. Each model/data-split combination has three S1 seeds (42–44) and one shared S0. Nine existing BW250 S2 controls permute acquisition-session labels within operating-frequency strata; these do not constitute a new S2 evaluation grid across all datasets. Optimization settings vary by configuration. Seeds and folds add computational model states, not independent machines.

3.2 FAULT-EXPRESSION DIRECTIONS AND JOINT CONTROL

Let $h_{\ell,p}(x) \in \mathbb{R}^d$ denote a residual state, and let p^* be the last prompt position before the first answer token. The two implemented direction constructors share the subsequent control adjustment, but we do not assume that they are mathematically equivalent.

Single-token cross-position lens. The BW250 Qwen and Llama configurations retain the average Jacobian operator (Gurnee et al., 2026). Let T be the final target layer, $\mathcal{D}_{\text{fit}}$ the fitting inputs, and $P_x = \{16, \dots, n_x - 2\}$. Then

$$J_\ell = \frac{1}{|\mathcal{D}_{\text{fit}}|} \sum_{x \in \mathcal{D}_{\text{fit}}} \frac{1}{|P_x|} \sum_{p \in P_x} \sum_{\substack{q \in P_x \\ q \geq p}} \frac{\partial h_{T,q}(x)}{\partial h_{\ell,p}(x)}, \qquad v_{\ell k} = (W_U J_\ell)_{\tau(k),:}. \tag{1}$$

Here, W_U is the raw unembedding matrix, and $\tau(k)$ indexes the single-token expression. This exported form excludes the final RMSNorm. Fitting aggregates positions, whereas quantitative readout uses p^*.

Full-sequence directions. The BW250 Mistral configuration and all public-dataset configurations use complete expressions, including multitoken terms such as inner race. For an expression $w_k = (w_{k1}, \dots, w_{km_k})$, we evaluate each token using the preceding tokens of that expression as context. The direction is

$$v_{\ell k}^{\text{seq}} = \frac{1}{|\mathcal{D}_{\text{fit}}|} \sum_{x \in \mathcal{D}_{\text{fit}}} \nabla_{h_{\ell,p^*}} \left[\frac{1}{m_k} \sum_{t=1}^{m_k} z_{w_{kt}}(x, w_{k,<t}) \right]^\top, \tag{2}$$

where z denotes a raw model logit rather than a log probability. Both constructors use candidate expressions and fixed prompts, without the class labels of fitting examples; each codebook is fitted separately. Neither construction substitutes the first subtoken for the complete expression.

Control-adjusted space. Stack the K expression directions into V_ℓ. Let $C_K = I_K - \mathbf{1}\mathbf{1}^\top/K$, and let R_ℓ span the joint row space of specified task directions and centered answer-code directions. We compute

$$\widetilde{V}_\ell = C_K V_\ell (I - R_\ell R_\ell^\top), \qquad Q_\ell = \operatorname{orth}(\widetilde{V}_\ell^\top), \qquad r_\ell = \operatorname{rank}(\widetilde{V}_\ell) \leq K - 1. \tag{3}$$

The implementation centers again after projection and applies deterministic SVD truncation with a relative tolerance of 10^{-8}. The construction imposes low rank and removes only the specified linear overlap. Writing $a_\ell(x) = h_{\ell,p^*}(x)$, the class readout is $s_{\ell k}(x) = a_\ell(x)^\top \widetilde{v}_{\ell k}$. Equal-rank random frames, unrelated-expression frames, and available orthogonal-complement frames reuse $\widetilde{V}_\ell Q_\ell$ as class coordinates, preserving direction lengths and pairwise inner products. The independent answer-code reference need not share this class geometry (Appendix B).

3.3 EVALUATION DIMENSIONS AND COMPARISON PRINCIPLES

Readability and correspondence. We average window scores within each evaluation file before computing one-versus-rest macro-AUROC and PR-AUC. We normalize ordered native layers to $[0, 1]$ and integrate by the trapezoidal rule to obtain the full-depth area, without selecting the best evaluation layer. Codebook metrics are averaged only after separate computation. The supervised ridge probe uses all fitting windows and their labels; the answer-code logit lens and fault-expression readout have different targets. We compare the correct assignment with all $K! - 1$ incorrect assignments. Random and unrelated-expression controls use 128 and 120 fixed instances, respectively.

Signal and model-state dependence. With the prompt, directions, and decision position fixed, we zero the signal embeddings, rotate the 32 embeddings left by 11 positions, or replace them. Replacement uses embeddings from an input of the next class under the same operating condition. It retains recipient labels and is compared with the baseline on the same pairable subset. Within each dataset/model/fold, all six directed transfers among three seeds apply source frames to target activations. This evaluation grid does not include cross-fold comparisons.

Local output effects. We select one layer using causal-fit files that do not overlap causal-eval files. At that layer, we translate within Q_ℓ along the difference between source- and target-class centroids:

$$h'_{\ell,p^*} = h_{\ell,p^*} + \alpha Q_\ell(\mu_{\ell,t} - \mu_{\ell,s}), \qquad \Delta(\alpha) = [L_t - L_s]_\alpha - [L_t - L_s]_0. \tag{4}$$

Here, L_y is the length-normalized log score of the complete answer. Positive Δ favors the target class, whereas negative Δ favors the source class. Doses are $0, 0.5, 1, 2$. Five controls with direction-wise matched norms and a next-layer reverse translation characterize specificity and residual output effects, rather than exact hidden-state recovery.

4 Experiments and analysis

4.1 Experimental setup

Datasets. Experiments cover five public benchmarks and the in-house BW250 dataset (Table 1). CWRU (Smith & Randall, 2015), Ottawa (Huang & Baddour, 2019), and Paderborn (Lessmeier et al., 2016) support bearing fault diagnosis. MCC5-THU (Chen et al., 2024) and SEU (Shao, 2018; Shao et al., 2019) support gearbox fault diagnosis. CWRU uses 12 kHz drive-end records with 0.007-inch defects, and MCC5-THU uses 48 records at the fixed M severity. BW250 contains single-channel vibration records for healthy, piston-fault, and valve-fault conditions, sampled at 25.6 kHz for 10 s per record. It covers four operating conditions between 15 and 30 Hz. Appendix A provides data partitions, window settings, background prompts, and the BW250 apparatus.

Models and training. We use Qwen2.5-3B-Instruct, Llama-3.2-3B-Instruct, and Mistral-7B-Instruct-v0.3 with the signal interface and abstract-code supervision specified in Section 3.1. Each dataset/model/split combination uses training seeds 42, 43, and 44, yielding 153 trained model states, all included in the overall readout evaluation. Supplementary training comparisons use 51 task-unadapted reference states and nine BW250 states trained with permuted labels (Appendix C).

Evaluation. Table 1 characterizes diagnostic performance using macro-F1; its note specifies the evaluation role for each dataset. Representation readout uses the file-level AUROC and full-depth area defined in Section 3.3. Overall results macro-average the class expressions within each task. Subsequent case analyses separately evaluate how the valve direction distinguishes valve faults from the remaining classes.

Table 1: Dataset coverage and diagnostic macro-F1. Scores are medians across the corresponding training seeds and data splits; state counts include all training seeds across the three models and their splits. BW250 is an in-house dataset; the others are public.

Dataset	Equipment	Classes	Splits	States	Qwen	Llama	Mistral
BW250	Drilling pump	3	1	9	0.991	0.991	0.991
CWRU	Bearing	4	4	36	1.000	1.000	0.113
MCC5-THU	Gearbox	4	4	36	0.182	0.263	0.212
Ottawa	Bearing	3	1	9	1.000	1.000	1.000
Paderborn	Bearing	3	6	54	0.425	0.362	0.349
SEU	Gearbox	3	1	9	1.000	1.000	1.000

Diagnostic scores use BW250 readout-evaluation files, validation data for CWRU, MCC5-THU, Ottawa, and Paderborn, and fitting time blocks for SEU. These scores describe the diagnostic context of each configuration, rather than rank configurations under a common independent-test protocol. Appendix A lists file counts for each stage.

4.2 Fault-expression readout across models and datasets

Figure 2 combines a specific case with an overview across configurations to illustrate fault-expression readout under abstract-code supervision. For BW250 Llama seed 42, Figure 2(a) shows layer-by-position projections of a valve-fault input onto the jointly adjusted valve direction. S/P/D mark signal, prompt, and decision positions; colors use a symmetric-log scale. Figure 2(b) compares the same direction's responses to healthy, piston-fault, and valve-fault inputs at the decision position before answering. For the three illustrated 20 Hz windows, the valve-fault input has a higher projection than the other two classes at all 27 source layers. This illustrates discrimination of the target fault by a natural fault-expression direction in these examples.

Figure 2(c) summarizes all 153 trained states across three models and six datasets. Dots denote states and diamonds denote group medians. Of these, 75 states have a full-depth macro-AUROC area above 0.5, with such states present in all three model families and all six datasets. Thus, above-chance fault-expression readout is observed in public bearing and gearbox tasks and the in-house drilling-pump task. Its level nevertheless varies with the model, task, and training state.

The two displays have different analytical scopes. Figures 2(a,b) use a single valve direction and three selected windows; Figure 2(c) summarizes all task class expressions over complete evaluation files. Clear separation in the single-expression case can therefore coexist with state-dependent variation in overall readout. We next conduct controlled comparisons of this valve-fault expression to examine its correspondence and effects.

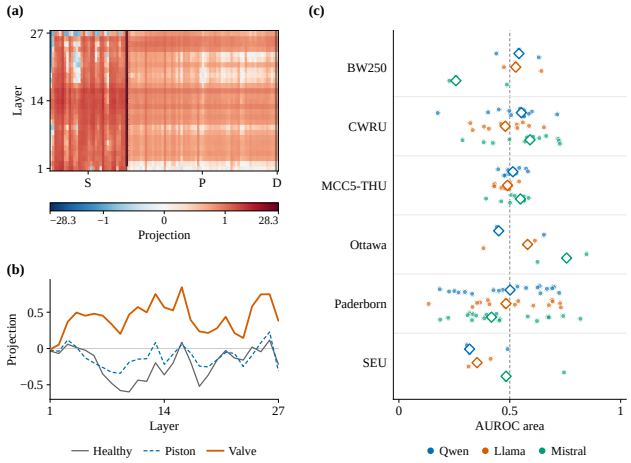


Figure 2: Fault-expression readout. (a) Projections across layers and positions. (b) Decision-position scores. (c) Macro-AUROC areas.

4.3 Correspondence between classes and expressions

Using BW250 Llama seed 42, we compare projections of healthy, piston-fault, and valve-fault records onto the same valve direction to examine its correspondence with the valve-fault class. The analysis uses the jointly adjusted direction and retains the previously selected layer 2. Evaluation uses 57 records excluded from layer selection, with 19 per class; each record's readout score averages projections from three windows. Figure 3(a) shows that valve-fault records generally have higher projections than healthy and piston-fault records.

We then fix the valve direction, score sign, and file set, and treat each of the three task classes in turn as the positive class. We compute one-versus-rest AUROC at all 27 source layers and its full-depth area (Figure 3(b)). The areas are 0.077, 0.481, and 0.943 when healthy, piston-fault, and valve-fault records are positive, respectively. Higher valve projections therefore correspond more clearly to valve faults in this file set.

4.4 Signal dependence

We test the signal dependence of the valve readout by changing the input signal. For BW250 Llama, we fix model parameters, background prompts, analysis directions, and the readout position. We separately zero the 32 signal embeddings, rotate them left by 11 positions, or replace them with embeddings from another class under the same operating condition. All conditions use the same 108 pairable records and retain the original record labels when computing valve-versus-rest AUROC.

Figure 4(a) shows a full-depth AUROC area of 0.934 for seed 42 with the original signal, falling to 0.074 after cross-class replacement. Zeroing gives an AUROC of 0.5 at every layer. In contrast, the area changes by only $+0.0012$ after the cyclic shift; the lower panel of Figure 4(a) shows layer-wise differences. For seeds 43 and 44, the areas also fall, from 0.774 and 0.896 to 0.025 and 0.138, respectively (Figure 4(b)). The replacement effect therefore occurs in more than one training seed.

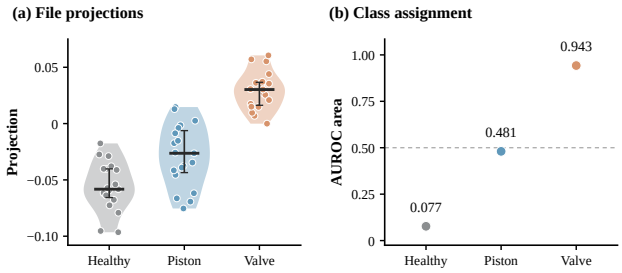


Figure 3: Valve-class correspondence (BW250, Llama seed 42). (a) Layer-2 projections: dots denote files, horizontal lines medians, and vertical bars interquartile ranges. (b) Full-depth AUROC with each class treated as positive.

These results support the signal dependence of the selected valve readout while showing relative insensitivity to this particular cyclic shift. Because replacement retains the original labels, the area decrease reflects weaker correspondence with the original class. It does not measure diagnostic accuracy for the replacement signal.

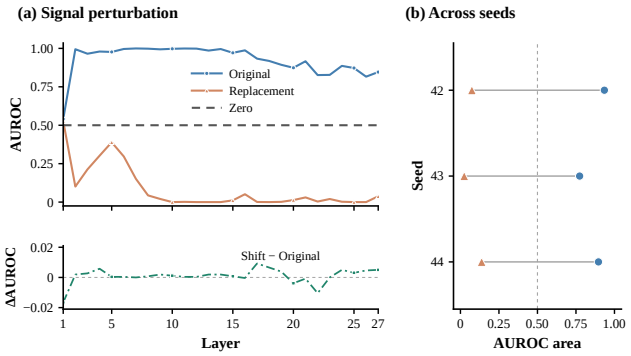


Figure 4: Signal dependence of the valve readout (BW250, Llama). (a) Layer-wise AUROC under replacement and zeroing (top) and changes after circular shifts (bottom). (b) Full-depth AUROC before and after replacement across three seeds.

4.5 Cross-state readout transfer

We test cross-state transferability by applying a valve direction fitted in one training state to the hidden representations of another. The experiment uses three BW250 Llama training seeds (42, 43, and 44). Source directions are applied directly at corresponding network layers without realignment or refitting. All six directed transfers among the seeds use the same 114 evaluation records, with each state's readout using its own directions as a reference.

Figure 5 shows valve-versus-rest full-depth AUROC areas of 0.850–0.961 across the six transfers. The selected expression thus retains strong class discrimination across seeds on the shared evaluation records. Transfer is asymmetric: the area is 0.959 for $42 \rightarrow 43$ and 0.851 for $43 \rightarrow 42$.

4.6 Same-space intervention

We change hidden representations in BW250 Llama seed 42 to test local effects of the fault-expression space on diagnostic outputs. With input signals and background prompts fixed, we intervene at the previously selected layer 2 and the decision position before answering. We translate along class-centroid differences within the jointly controlled space, examining piston-fault-to-valve-fault interventions and the reverse direction. Layer selection and centroid estimation use 57 records, while evaluation uses another 57 disjoint records, including 19 source-class records per direction. Effects

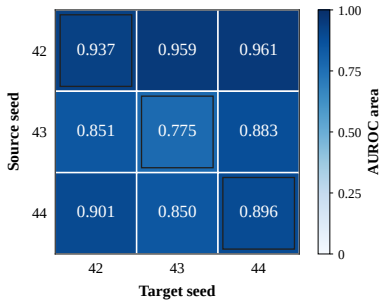


Figure 5: Cross-seed valve-readout transfer (BW250, Llama). Cells show full-depth AUROC; rows and columns index source and target seeds. Bordered diagonal cells use each state's own directions.

are changes in target-relative-to-source answer scores; positive values indicate a greater preference for the target class.

Figure 6(a) shows that output responses vary with direction and dose. For piston-to-valve intervention, the effect is $+0.0044$ nats/token at dose 1, with a 95% interval containing zero. At dose 2, it is -0.0154 nats/token, with the interval below zero. The reverse intervention gives -0.0088 nats/token at dose 1, also with the interval below zero, whereas its mean effect at dose 2 is near zero. Increasing the translation magnitude therefore does not produce a monotonic target-directed response.

Figure 6(b) compares the unit-dose main effect with five norm-matched controls. For piston-to-valve intervention, the paired-difference point estimates are positive against every control, but only the interval against the orthogonal-complement control lies entirely above zero. The reverse intervention likewise shows no consistent advantage over all control types.

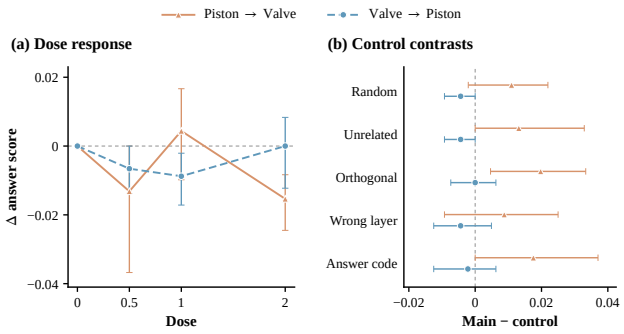


Figure 6: Answer-score effects in nats/token (BW250, Llama seed 42, layer 2). (a) Dose responses. (b) Main-minus-control effects at unit dose. Bars show 95% intervals from paired resampling of operating-condition groups.

Overall results. The valve-fault case shows class correspondence, signal dependence, and cross-seed retention, but these properties are not equally clear in all configurations. Signal replacement can increase or decrease readouts, and cross-state transfer does not generally retain strong discrimination. Same-space interventions can locally affect answer scores, but their effects vary with direction and dose.

5 Discussion

Abstract-code supervision separates training answers from the fault expressions used for analysis, allowing us to examine internal class information beyond answer vocabulary. In the selected BW250 case, the valve readout changes with vibration inputs and transfers across training seeds. This class discrimination therefore cannot be explained solely by a fixed prompt and is not confined to one

checkpoint. Semantic specificity is a separate question that requires comparison with controls such as unrelated expressions; class separation alone does not establish it.

Strong readout does not necessarily provide reliable target-directed control. Readout assesses class discrimination under natural inputs, whereas control tests whether downstream computation produces the intended answer change after an artificial change in representation. The J-space concept-replacement experiments by Gurnee and colleagues also show that replacement success varies with concept category and intervention strength. Our intervention translates class centroids at a single layer's decision position, unlike their coordinate swaps across intermediate layers and token positions. The connection is therefore the need to test intervention reliability, rather than evidence that both approaches identify the same mechanism.

Same-space interventions provide local evidence that selected coordinates affect answer scores, but their specific role in normal diagnosis remains unresolved. Subspace interventions may change outputs through alternative computational paths. Further work needs to identify the physical signal features supporting fault readouts and trace how related information reaches the answer during unperturbed inference. Independently varying equipment background can also help distinguish the contributions of vibration inputs and semantic prompts.

6 Conclusion

We introduce J-FDCS to separate abstract-code supervision from analysis with natural fault expressions. It jointly removes the linear overlap of fault-expression directions with specified task-control word directions and answer-code directions, enabling readout and intervention in the same space. We evaluate Qwen, Llama, and Mistral on five public rotating-machinery benchmarks and the in-house BW250 drilling-pump dataset. Fault-expression readouts vary across models, tasks, and training states. The selected BW250 valve-fault case shows clear class correspondence, signal dependence, and cross-seed retention. Same-space interventions further reveal local effects of selected coordinates on answer scores, while showing that readability does not imply reliable target-directed control. J-FDCS provides an analytical approach beyond diagnostic accuracy for studying fault-related representations in signal-grounded language models.

Reproducibility statement

Sections 3–4 describe the methods and evaluation procedures, and the appendices provide data partitions, experimental settings, and statistical details. Numerical checks use saved file-level scores and paired resampling records. They do not independently rerun training, direction fitting, or historical window aggregation.

Code and data availability. We plan to release the J-FDCS implementation, experiment configurations, and analysis scripts after the paper is accepted. The anonymous project page is `https://anonymous.4open.science/r/J-FDCS-129A`. Public benchmarks are available from the original sources cited in the paper. The in-house BW250 vibration data are available on reasonable request, subject to data-access requirements.

Statement on AI use

Generative AI tools, including ChatGPT and Codex, were used to assist with language polishing and translation, literature retrieval and synthesis, research ideation and execution, and drafting sections of the manuscript. Research assistance included discussions of hypotheses and experimental design, code development and debugging, experiment orchestration, source verification, statistical recomputation, and result interpretation. These tools also assisted with figure preparation; all experimental result figures were drawn from saved numerical results. The authors reviewed the AI-assisted materials and take full responsibility for the methods, code, experimental results, figures, tables, and text of this paper.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Shijin Chen, Zeyi Liu, Xiao He, Dongliang Zou, and Donghua Zhou. Multi-mode fault diagnosis datasets of gearbox under variable working conditions. *Data in Brief*, 54:110453, 2024. doi: 10.1016/j.dib.2024.110453.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen, Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang, T. Ben Thompson, David Abrahams, Subhash Kantamneni, Emmanuel Ameisen, Joshua Batson, and Jack Lindsey. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/workspace/index.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Huan Huang and Natalie Baddour. Bearing vibration data under time-varying rotational speed conditions. Mendeley Data, 2019. URL https://doi.org/10.17632/v43hmbwxpm.2. Version 2.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/680b2a8135b9c71278a09cafb605869e-Abstract-Conference.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Christian Lessmeier, James Kuria Kimotho, Detmar Zimmer, and Walter Sextro. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. *PHM Society European Conference*, 3 (1), 2016. doi: 10.36001/phme.2016.v3i1.1577.

Qi Li, Xinran Zhang, Jinfeng Huang, Hongliang He, Feibin Zhang, Zhaoye Qin, and Fulei Chu. VSLLaVA: A pipeline for large multimodal foundation models in industrial vibration signal analysis. *Advanced Engineering Informatics*, 76:105023, 2026. doi: 10.1016/j.aei.2026.105023.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Meta. Llama-3.2-3B-Instruct: Model card. Hugging Face, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct. Released September 25, 2024. Accessed September 23, 2026.

Mistral AI. Mistral-7B-Instruct-v0.3: Model card. Hugging Face, 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Accessed September 23, 2026.

Siyu Shao. Mechanical datasets: Gearbox gearset. GitHub repository, 2018. URL https://github.com/cathysiyu/Mechanical-datasets/tree/9732a2820f4cdf91fa7d1d0cc95442376660f5dd/gearbox/gearset. Commit 9732a2820f4cdf91fa7d1d0cc95442376660f5dd. Accessed September 23, 2026.

Siyu Shao, Stephen McAleer, Ruqiang Yan, and Pierre Baldi. Highly accurate machine fault diagnosis using deep transfer learning. *IEEE Transactions on Industrial Informatics*, 15(4):2446–2455, 2019. doi: 10.1109/TII.2018.2864759.

Wade A. Smith and Robert B. Randall. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mechanical Systems and Signal Processing*, 64–65: 100–131, 2015. doi: 10.1016/j.ymssp.2015.04.021.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115v2. Version 2, revised January 3, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

A EXPERIMENTAL SETUP

A.1 DATA SPLITS AND THE BW250 EXPERIMENTAL SETUP

Table A.1 lists the record counts and window budgets for each data split. Direction fitting and supervised probes use the same fitting records but different numbers of windows.

Table A.1: Data splits and window budgets.

Dataset	Training records	Fitting records	Readout evaluation	Codebooks	Windows: direction / probe / readout
BW250	828	114	114	1	1 / 19 / 3
CWRU	8	4	4	4	2 / 10 / 10
MCC5-THU	24	12	12	4	2 / 59 / 59
Ottawa	9	9	9	1	2 / 19 / 19
Paderborn	240	30	30	1	2 / 6 / 6
SEU	3	3	3	1	2 / 10 / 13

SEU counts time blocks; other datasets count records or files. Window budgets are per file and codebook for direction fitting, probe fitting, and readout evaluation, respectively.

In Paderborn, the specimens used for parameter training do not overlap with the held-out specimens, but direction fitting and readout evaluation use different records from the same held-out specimens. Ottawa reuses individual bearings across trajectories. SEU uses designated time blocks from the 20_0 records for training and fitting and designated time blocks from the 30_2 records for readout evaluation; diagnostic scores are computed on the fitting time blocks. The record-level holdouts in BW250 and CWRU do not constitute tests on independent equipment.

The BW250 test rig consists of a drilling pump, a drive system, a circulating-water system, and a sensor acquisition chain (Figure A.1). We use only single-channel vibration data for three classes: healthy, piston fault, and valve fault.

A.2 PROMPTS, CODE MAPPINGS, AND ANALYSIS EXPRESSIONS

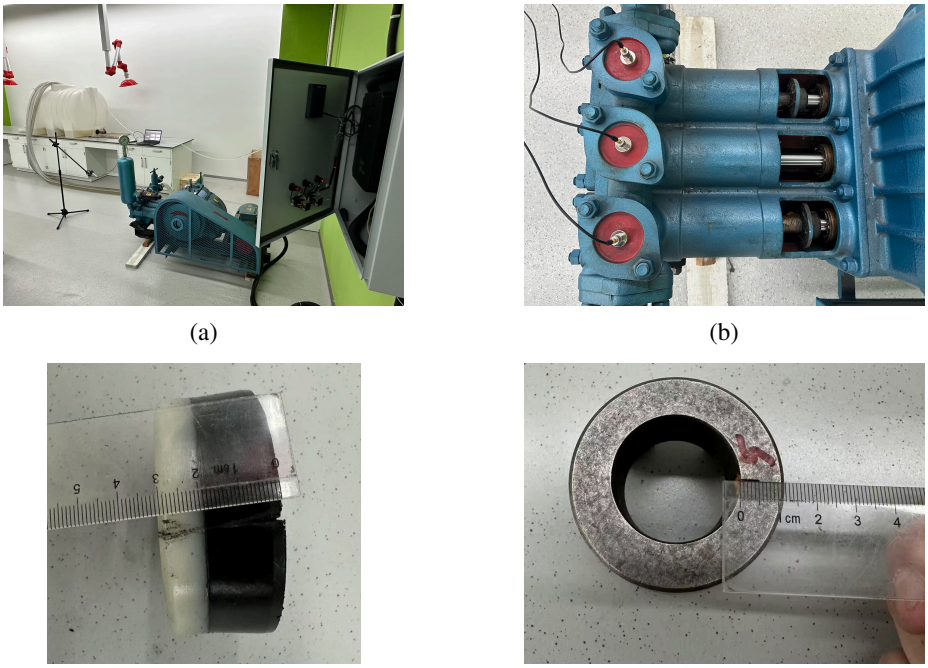
Table A.2 lists the analysis expressions in class order. BW250, Ottawa, Paderborn, and SEU assign A/B/C in this order. CWRU and MCC5-THU use four cyclic mappings, with CB1–CB4 assigning A/B/C/D, B/C/D/A, C/D/A/B, and D/A/B/C, respectively. For the four-class tasks, the user prompt specifies only the codebook identifier. The supervised targets define the mapping from classes to codes, which is not listed in the prompt.

Table A.2: Fault expressions in class order.

Dataset	Analysis expressions in class order
BW250	healthy; piston; valve
CWRU	healthy; inner race; rolling element; outer race
MCC5-THU	healthy; gear wear; gear pitting; tooth crack
Ottawa	healthy bearing; outer race fault; inner race fault
Paderborn	healthy bearing; outer race fault; inner race fault
SEU	healthy; chipped; missing

Both expressions and answer codes include one leading space, and their complete token sequences are retained for each model. BW250 uses single-token directions for Qwen and Llama, while the remaining configurations use full-sequence directions. Both jointly adjust for task controls and answer codes.

Within each dataset, the system and user message bodies are identical across the three models, and each configuration keeps its context fixed during training, direction fitting, and evaluation. The English message bodies are given below without repeating model-specific chat delimiters. CWRU and



(a) (b) (c) (d)

Figure A.1: BW250 test rig and damaged components.

MCC5-THU include natural fault descriptions, so code supervision does not exclude fault semantics from the input.

User messages (verbatim).

BW250	Classify the input vibration signal. Return exactly one code: A, B, or C.
CWRU, MCC5-THU	Classification codebook: CB1. Classify the vibration signal. Return exactly one code: A, B, C, or D.
Ottawa, SEU	Classify the input vibration signal. Respond with exactly one code: A, B, or C.
Paderborn	Classify the input vibration signal. Return exactly one standardized diagnostic code: A, B, or C.

The CWRU and MCC5-THU messages show CB1 verbatim; the other codebooks replace only CB1 with CB2, CB3, or CB4.

System messages (verbatim).

BW250	You are a drilling-pump condition classification model.
Ottawa	You are a variable-speed rolling-bearing signal classifier. A housing-mounted accelerometer records a monitored test bearing while shaft speed changes. An impact-generating discontinuity attached to the housing reference can produce recurring impacts as rolling elements pass. An impact-generating discontinuity rotating with the shaft travels through the load zone. Without a localized impact-generating discontinuity, these localized impacts are absent. Use the vibration evidence to classify bearing condition.
Paderborn	You are a rolling bearing diagnostic model.

SEU	You are a drivetrain gearbox diagnostic model. The input is one vibration direction measured on a planetary gearbox. Gear meshing can produce repeated vibration excitation.

System messages (continued).

CWRU

You are a rolling-bearing fault diagnosis model.
Equipment knowledge:
Rolling-bearing equipment knowledge
The monitored rolling-element bearing supports a rotating motor shaft.
Vibration is measured by an accelerometer attached to the motor housing near the drive end.
Healthy
Healthy denotes normal bearing operation without a deliberately seeded inner race, rolling element, or outer race defect; it does not mean that no acceleration is recorded.
The measured response can include ordinary operational vibration and depends on rotation speed and structural transmission.
Inner race
When a rolling element encounters a local inner race defect during operation, the contact can produce an impact-like response.
The repetition rate depends on rotation speed and bearing geometry, while sliding can shift the observed characteristic frequency.
Rolling element
A local rolling element defect can contact other bearing elements repeatedly as the bearing moves.
Ball-spin and cage-related frequencies describe relevant timing, while geometry and sliding affect the observed response.
Outer race
An outer race defect is stationary relative to the housing, and rolling elements repeatedly encounter its location.
Its vibration response depends on the defect position relative to the load zone and on transmission through the bearing and housing.
The localized faults in this equipment were introduced by electrical-discharge machining, so these statements concern manufactured point defects rather than naturally propagated damage.

MCC5-THU

You are a gearbox fault diagnosis model.
Equipment knowledge:
Gearbox equipment knowledge
The monitored equipment is a two-stage parallel gearbox, and the selected observation is axial vibration acceleration at the intermediate-shaft bearing seat.
A deliberately modified 36-tooth gear is located on the intermediate shaft.
Healthy
Healthy is a recorded gearbox operating state rather than the absence of a vibration signal.
The measured response can vary with speed and torque.
Gear wear
Gear wear changes the tooth profile and contact relationship.
The resulting change in time-varying mesh stiffness can alter the dynamic response, with its observation depending on speed and torque.
Gear pitting
Gear pitting is localized material loss on a tooth surface.
Pitting-aware models represent its effect through changes in pair mesh stiffness; the selected records also vary in speed and torque.
Tooth crack
A tooth crack changes local tooth compliance, and crack-aware models calculate the corresponding change in mesh stiffness.

MCC5-THU (continued)
The broad tooth crack description is retained because the available equipment evidence does not establish one exact crack location for every selected record.

Each model uses its native chat template. Mistral combines the system and user text into a single instruction.

A.3 Training and preprocessing settings

Task adaptation uses QLoRA with 4-bit NF4 double quantization (Dettmers et al., 2023), a microbatch size of 1, gradient accumulation of 8, a learning rate of 2×10^{-4}, weight decay of 0.01, and a warmup ratio of 0.03. The LoRA adapter (Hu et al., 2022) has rank 16, a scaling parameter of 32, and dropout of 0.05. The target modules are the q/k/v/o and gate/up/down projections. The analysis uses the final task-adapted checkpoint saved for each configuration. The number of training epochs and sample presentations varies across datasets.

The signal encoder is a ResNet1D-18 with GroupNorm that outputs 32 embeddings of dimension 512. The projector maps these embeddings to the hidden spaces of Qwen, Llama, and Mistral, with dimensions 2048, 3072, and 4096, respectively. Encoder pretraining uses masked waveform modeling, and projector pretraining uses masked signal embedding reconstruction. Both stages incorporate multiview temporal contrastive learning. The mask ratio is 0.4, and both the contrastive weight and temperature are 0.1. The respective learning rates are 3×10^{-4} and 1×10^{-4}, with weight decay of 1×10^{-4} in both stages. Both stages use seed 42 and bf16, and the resulting interface is reused across task seeds within the same configuration. Projector pretraining freezes the encoder and language model, and task adaptation is configured to freeze the signal interface.

Pretraining augmentations include amplitude scaling by 0.9–1.1, time shifts of up to 5%, and frequency masking of 2%–8%. Table A.3 lists the window settings, interface batch sizes, and training step counts in the reports.

Table A.3: Window settings and interface pretraining.

Dataset	Window length / stride (samples)	Encoder: batch / steps	Projector: batch / steps
BW250	25,600 / 12,800	64 / 6,630	8 or 4 / 1,000
CWRU	12,000 / explicit starts	16 / (40, 96, 80, 44)	8 / 72 or 80
MCC5-THU	12,800 / explicit starts	128 / (300, 300, 290, 300)	8 / 1,000
Ottawa	200,000 / 100,000	16 / 171	4 / 380
Paderborn	64,000 / 32,000	64 / 600	4 / 1,000
SEU	65,536 / 32,768	8 / 60	4 / 63 or 70

Encoder counts in parentheses follow folds 1–4. BW250 projector batches are 8 for Qwen and 4 otherwise. CWRU projector steps are 72 for Llama folds 3–4 and 80 otherwise; SEU uses 63 for Llama and 70 otherwise. Encoder/projector schedules allow up to 30/10 epochs, with a 1,000-step projector cap.

For preprocessing, BW250 uses the cylinder 1 vibration channel, CWRU uses drive-end vibration, MCC5-THU uses vibration in the x direction, Ottawa uses Channel_1, Paderborn uses vibration_1, and SEU uses the second column of the original file. CWRU and MCC5-THU extract windows at saved, non-overlapping explicit start points rather than sliding continuously at a nominal stride. Channel means and standard deviations are estimated only from an unlabeled fitting subset within the parameter-training records, excluding internal validation records. Ottawa task training uses trajectory A. Paderborn windows retain sample-based units, and the physical sampling rate of SEU remains unestablished.

Record scope. Task adaptation uses QLoRA supervised fine-tuning, with training budgets adjusted to the data split and model configuration. All analyses use the final task-adapted checkpoint for the corresponding configuration. Table A.3 reports the main interface pretraining and task adaptation parameters.

B Analysis and intervention details

B.1 Direction construction and geometric controls

Direction fitting follows the two constructions in Section 3.2. The single-token operator aggregates source positions and downstream target positions. The full-sequence operator differentiates mean raw logits under each candidate expression's own prefix, without using the sample's true class as the fitting target. This derivative differs from the complete-answer log probability used to evaluate interventions.

Let Q_ℓ denote the orthonormal basis after joint adjustment, and let U_ℓ denote an orthonormal random, unrelated-expression, or orthogonal-complement control basis of the same rank. We retain the class coordinates of the main space, $B_\ell = \widetilde{V}_\ell Q_\ell$, and construct $\widetilde{V}_\ell^{\mathrm{ctrl}} = B_\ell U_\ell^\top$, giving

$$\widetilde{V}_\ell^{\mathrm{ctrl}} \big(\widetilde{V}_\ell^{\mathrm{ctrl}}\big)^\top = B_\ell B_\ell^\top = \widetilde{V}_\ell \widetilde{V}_\ell^\top . \tag{B.1}$$

The controls therefore preserve the lengths and pairwise inner products of the class directions, without training a separate classifier in the control space. The answer-code reference is constructed independently and need not match this geometry. Unavailable constructions are marked as unevaluated rather than assigned zero scores.

B.2 Readout, correspondence assignments, and case evaluation

Qwen, Llama, and Mistral retain 35, 27, and 31 source layers, respectively. At each layer, window scores are first averaged within files, followed by class-wise AUROC and PR-AUC and their macro-averages. PR-AUC uses the implementation's average precision calculation. Depth is normalized to $[0, 1]$ in native layer order, and trapezoidal integration gives the full-depth area without selecting the best evaluation layer. Metrics are calculated separately for each codebook and then averaged, without pooling raw scores. The supervised ridge probe uses all fitting windows and their labels; Table A.1 lists the window budgets.

The complete correspondence analysis evaluates all $K!$ class-to-expression assignments. It records the correct assignment's area minus the median area of incorrect assignments, its midrank, and whether it ranks uniquely first. The tie tolerance for assignment ranking is 10^{-12}. The random and unrelated-expression references contain 128 and 120 fixed instances, respectively. For each instance and assignment, full-depth areas are first averaged across codebooks.

The 20 Hz, window-9 case in Figure 2(a,b) is selected based on an early inspection of the fixed windows. Figure 3(a) uses the layer selected by the procedure in Appendix B.4 and displays only the 57 records excluded from layer selection. Figure 3(b) fixes the same valve direction and varies the positive-class label. It is not a $K!$ assignment test of the complete expression set and is not directly compared with multiclass macro-averaged references. These cases are post hoc analyses rather than independent blind tests.

B.3 Signal changes and transfer across model states

Zeroing and an 11-position circular left shift both act on the 32 signal embeddings, with the background prompt held fixed. Cross-class replacement uses a fixed donor from the next class under the same operating condition. It retains recipient labels and is compared with the original readout on the same pairable subset. The four conditions in Figure 4 use the same 108 BW250 records. For zeroing, the constant prefix is recomputed with a fixed batch shape to avoid treating batch-dependent numerical differences as class information.

Transfer across seeds is restricted to the same dataset, model, and data split. Directions from the source model state are applied directly at the corresponding layer of the target model state, without realignment or refitting. Transferred and target-self readouts use the same files; Figure 5 uses 114 BW250 evaluation records. Different seeds share the evaluation set. Layers, seeds, data folds, and codebooks do not count as independent physical replicates.

B.4 Interventions in the same space and paired statistics

Layer selection and translation use 57 layer-selection records. A candidate layer's multiclass macro-averaged AUROC must exceed all three fixed controls: random, unrelated-expression, and orthogonal-complement. Among eligible layers, we select the layer with the largest minimum gap over the controls; ties favor the shallower layer. The final layer is excluded to allow reversal at the next layer. Model states without an eligible layer do not enter causal evaluation and are not assigned a zero effect. Class centroids are estimated from the same layer-selection records. Following Equation 4, source-to-target translations are applied at the decision position before answering, with doses of 0, 0.5, 1, and 2. Evaluation uses a disjoint set of 57 records.

The five intervention controls are random, unrelated-expression, orthogonal-complement, wrong-layer, and answer-code interventions. For each ordered class direction, their translation norms match that of the main intervention. Each control uses its frozen instance at the corresponding layer. Reversal at the next layer uses that layer's own analysis basis and its target-to-source centroid difference. It measures the residual answer-score shift after both operations and is not interpreted as exact recovery of the original hidden state.

For estimation and intervals, each direction includes 19 evaluation records from the source class. Each record first averages three windows. Let d_i denote the change in target-relative-to-source answer score for file i, g_i its operating-condition group, and n_g the number of files in group g. The point estimate is the file-weighted mean. Resampling draw b samples four operating-condition groups $g_1^{(b)}, \ldots, g_4^{(b)}$ with replacement and calculates

$$\widehat{\Delta} = \frac{1}{N}\sum_{i=1}^{N} d_i, \qquad \widehat{\Delta}^{(b)} = \frac{\displaystyle\sum_{j=1}^{4}\sum_{i:g_i=g_j^{(b)}} d_i}{\displaystyle\sum_{j=1}^{4} n_{g_j^{(b)}}}. \tag{B.2}$$

Repeatedly sampled groups retain all their files with the corresponding multiplicity. The main intervention and controls use the same 2,000 resampling draws. The 2.5th and 97.5th percentiles give 95% intervals without adjustment for multiple comparisons. Control comparisons first subtract the control effect from the main effect for each file, then resample under the same rule. A positive difference indicates a larger signed main effect, not a larger absolute effect. The point estimate is also not an equally weighted average of the four group means. The BW250 57/57 split follows the existing setup, and these analyses are not blind tests.

C Supplementary results

This appendix reports overall readout, training comparisons, and intervention results. Apart from answer-score effects, all areas are full-depth macro-averaged AUROC areas, rather than the single-expression metric for the valve-fault case in the main text.

C.1 Readout references and training comparisons

Table C.1 combines the three readouts and paired training differences. Ridge-probe areas approach 1 on BW250, CWRU, Ottawa, and Paderborn, whereas natural fault-expression readouts vary across configurations. The probe learns a classification rule from fitting labels, so the two readouts do not provide equivalent semantic evidence. Readout changes from the task-unadapted reference to correct-code training have both positive and negative signs. BW250 label-permutation training serves only as a supplementary comparison within the same model and seed.

C.2 Signal changes and transfer across model states

Table C.2 summarizes cross-class replacement and transfer across seeds within the same data split, grouped by dataset. Replacement differences subtract the original area on the same pairable subset from the replacement area and retain recipient labels for evaluation. All zeroing areas are 0.5. Area

Table C.1: Readout and training comparisons: medians across seeds and splits.

Dataset	Model	J-FDCS	Ridge	Code lens	Δ_{10}	Δ_{12}
BW250	Qwen	0.541	1.000	0.538	-0.027	-0.044
BW250	Llama	0.527	1.000	0.784	-0.269	$+0.083$
BW250	Mistral	0.257	1.000	0.848	-0.335	-0.162
CWRU	Qwen	0.552	1.000	0.535	$+0.049$	N/A
CWRU	Llama	0.480	1.000	0.542	-0.072	N/A
CWRU	Mistral	0.591	1.000	0.500	-0.089	N/A
MCC5-THU	Qwen	0.514	0.684	0.500	-0.040	N/A
MCC5-THU	Llama	0.490	0.675	0.500	-0.029	N/A
MCC5-THU	Mistral	0.548	0.654	0.500	$+0.027$	N/A
Ottawa	Qwen	0.450	1.000	0.633	$+0.179$	N/A
Ottawa	Llama	0.580	1.000	0.724	$+0.074$	N/A
Ottawa	Mistral	0.756	1.000	0.817	$+0.193$	N/A
Paderborn	Qwen	0.501	1.000	0.503	-0.028	N/A
Paderborn	Llama	0.483	1.000	0.573	$+0.067$	N/A
Paderborn	Mistral	0.417	1.000	0.343	-0.012	N/A
SEU	Qwen	0.319	0.821	0.500	-0.135	N/A
SEU	Llama	0.353	0.788	0.556	-0.244	N/A
SEU	Mistral	0.483	0.858	0.618	$+0.131$	N/A

Here, Δ_{10} is the median state-wise S1–S0 area difference using shared S0; Δ_{12} is the median same-seed S1–S2 difference for BW250. Differences are computed before taking medians. N/A denotes unevaluated comparisons.

differences under the prescribed 11-position circular shift range from -0.032 to $+0.060$. Zeroing and shifting use each configuration's designated evaluation records, whereas Figure 4 restricts all four conditions to the same 108 BW250 records.

Table C.2: Signal replacement and transfer. Transfer areas: median [minimum, maximum].

Dataset	n	Replacement decrease/increase	Median replacement difference	N	Median transfer area [range]
BW250	9	7 / 2	-0.132	18	0.535 [0.177, 0.690]
CWRU	36	19 / 17	-0.025	72	0.499 [0.271, 0.733]
MCC5-THU	36	18 / 18	$+0.002$	72	0.501 [0.413, 0.596]
Ottawa	9	6 / 3	-0.144	18	0.545 [0.383, 0.838]
Paderborn	54	28 / 26	-0.005	108	0.486 [0.178, 0.818]
SEU	9	2 / 7	$+0.130$	18	0.469 [0.275, 0.775]

Here, n counts paired states and N directed seed transfers; rows pool models and splits. Decrease/increase refers to replacement minus original area. Transfers use no refitting, alignment, or bidirectional averaging.

C.3 BW250 INTERVENTIONS IN THE SAME SPACE

Figure C.1 shows dose curves for all nine BW250 model states, including Mistral seed 42. All nine states satisfy the layer-selection criterion in Appendix B.4. Each state contains six ordered class directions, giving 54 direction–state combinations. Main effects at unit dose range from -0.0450 to $+0.0285$ nats/token. Table C.3 summarizes the main intervention minus each of the five matched controls. Six direction–state combinations have point estimates above all five controls. No combination has paired intervals above zero against every control.

After reversal at the next layer, the absolute file-averaged effect at unit dose decreases in 16 combinations and increases in 38, using unrounded means. This includes a numerically negligible decrease below 10^{-8} nats/token for the Mistral seed 42 piston-to-valve direction. This comparison measures the residual answer-score shift after two operations, not the extent of hidden-state recovery.

Table C.3: Main minus control at unit dose (nats/token; 54 combinations per control).

Control	Paired difference range	Interval entirely > 0	Interval entirely < 0	Interval includes 0
Random	$[-0.059, +0.022]$	10	8	36
Unrelated expression	$[-0.072, +0.137]$	10	11	33
Orthogonal complement	$[-0.054, +0.043]$	8	14	32
Wrong layer	$[-0.110, +0.054]$	7	7	40
Answer code	$[-1.300, +0.018]$	3	37	14

Counts use unrounded endpoints of the unadjusted paired 95% bootstrap intervals in Appendix B.4. Positive differences indicate larger signed main effects.

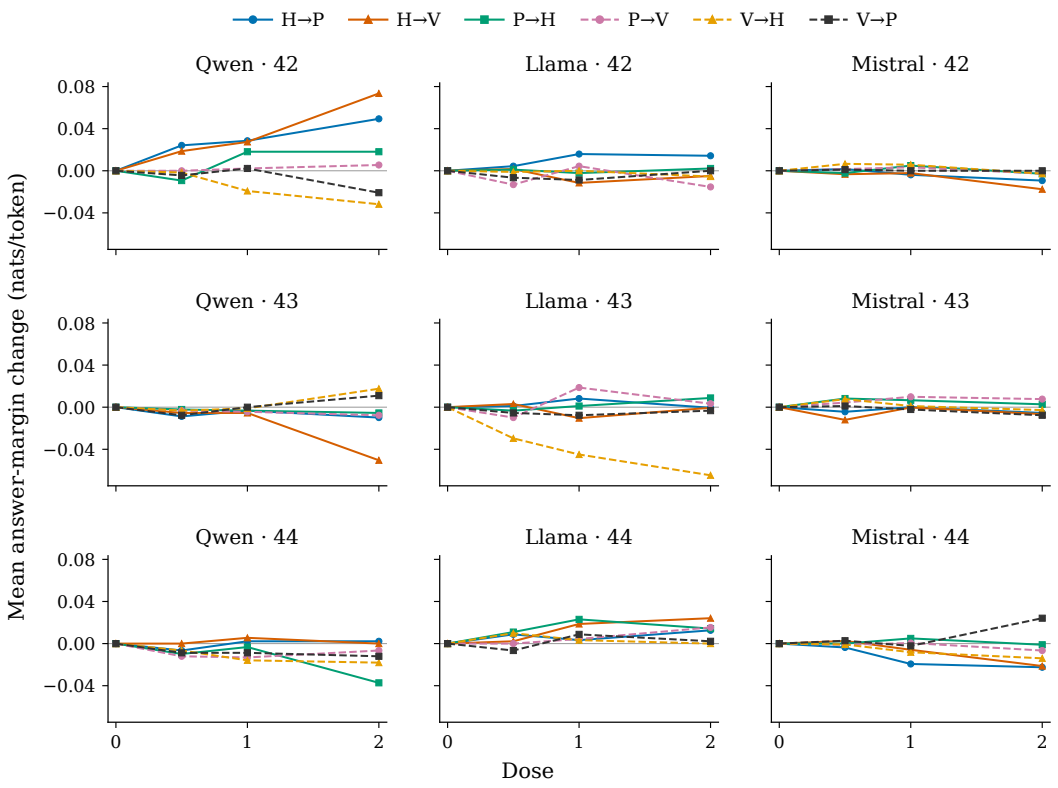


Figure C.1: BW250 dose responses. H/P/V: healthy/piston fault/valve fault. Curves show file-averaged effects without intervals.