BRINGING MODEL INTERNALS INTO EXPERT ANNOTATION OF LEGAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

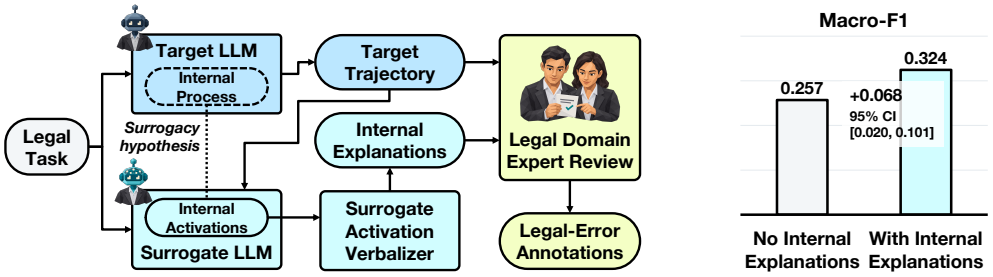
Large language models (LLMs) increasingly perform complex professional work, creating a need to examine the reasoning behind their decisions. In legal tasks, written rationales alone may provide an incomplete basis for expert oversight. We introduce an **activation-informed expert annotation framework** in which legal experts inspect generated reasoning alongside natural-language explanations derived from model activations. Three experts contributed 34 hours of review, producing 354 span-level legal annotations across 15 conversational and agentic trajectories, together with explanation-helpfulness ratings for eight trajectories. To supply explanations across models, we propose **Cross-LLM Activation Surrogacy**: an accessible surrogate model reads a target model's trajectory, and a pretrained Natural Language Autoencoder (NLA) (Fraser-Taliente et al., 2026) describes the surrogate's resulting activations. This approach requires neither access to the target's activations nor training a dedicated NLA for each target. Experts judged NLA-based explanations helpful in 41.4% of their ratings. In classification of previously identified legal errors, aligned surrogate explanations improve macro-F1 from 0.257 to 0.324 relative to providing no explanations, with gains also observed over text-only and shuffled-explanation controls. Cross-model comparisons show stronger semantic correspondence between explanations within model families than across families, while reconstruction similarity alone does not reliably track this correspondence. Together, these results establish a practical framework for activation-informed expert annotation, with initial evidence that aligned surrogate explanations support classification of expert-identified legal errors. Code and dataset are available at: https://anonymous.4open.science/r/tr3D8F.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed for complex professional tasks, including extended workflows carried out with growing autonomy (Handa et al., 2025; McCain et al., 2026). As these systems make intermediate decisions that guide subsequent actions, effective oversight requires attention to the reasoning and assumptions underlying their behavior, alongside the correctness of their final outputs. This is particularly important in legal tasks, where sound judgments depend on interpreting facts, applying relevant rules, and justifying the connections between them. Yet a model's explicitly written rationale does not necessarily reflect the internal computation underlying its response (Turpin et al., 2023; Chen et al., 2025). This motivates incorporating evidence about internal model processing into expert evaluation. We introduce an **activation-informed expert annotation framework** that pairs model-generated trajectories with natural-language explanations derived from internal model representations, or *activations*, enabling experts to annotate legal errors and separately assess the explanations' helpfulness.

Using this framework (Figure 1(a)), we collected annotations from three legal subject matter experts. Experts reviewed model-generated trajectories from PRBench (Akyürek et al., 2026) and Harvey's LAB (Harvey AI, 2026), covering conversational legal tasks and longer-horizon agentic workflows selected to match their practice areas. They identified erroneous spans, assigned legal-error categories, and provided written feedback. Separately, they rated summarized explanations of internal model processing as helpful, neutral, or misleading, with each summary claim linked to its underlying token-level explanations. Annotation spanned seven weeks and required 34 expert-hours, with experts compensated at approximately 100 USD per hour. The resulting dataset covers 15 trajectories,

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.



(a) Activation-informed expert annotation. (b) Legal-error classification.

Figure 1: **Annotation framework and downstream utility.** (a) Legal experts review target-model trajectories alongside NLA explanations of surrogate activations to annotate legal errors. The dotted link represents the hypothesis that surrogate activations provide a useful proxy for aspects of the target's internal reasoning. (b) Aligned surrogate explanations improve macro-F1 from 0.257 to 0.324 for classifying expert-identified legal errors (Section 3.2).

comprising 354 span-level legal annotations and explanation-helpfulness ratings for eight trajectories. It brings expert legal judgments and assessments of internal explanations into a shared resource for studying legal reasoning and its evaluation.

To provide experts with explanations of internal model processing, we draw on mechanistic interpretability, which investigates the representations and computations underlying model behavior (Rai et al., 2024; Bereska & Gavves, 2024). Model activations are the intermediate numerical values computed within a neural network during a forward pass; in LLMs, these include vectors representing information at particular layers and token positions. Sparse autoencoders (SAEs) decompose these activations into interpretable features (Huben et al., 2024), while circuit-tracing methods investigate interactions among features that contribute to model outputs (Ameisen et al., 2025). Complementary approaches translate activations into natural-language explanations (Pan et al., 2026; Karvonen et al., 2025). We use Natural Language Autoencoders (NLAs), which generate descriptions from activations and learn to preserve information needed to reconstruct those activations (Fraser-Taliente et al., 2026). Their natural-language output fits our annotation workflow: experts can review explanation claims alongside a trajectory and assess their relevance to legal reasoning. Released NLA checkpoints also make it possible to build on existing interpretability models, while their reconstruction component provides an additional way to evaluate the information retained in their explanations.

Applying NLAs directly to each model under review requires access to its activations and an NLA trained for that model. These requirements make extending the annotation framework to new models costly and prevent direct application to closed frontier models whose activations are inaccessible. We therefore propose **Cross-LLM Activation Surrogacy**, which reuses an existing NLA across target models. A target model generates a trajectory, which an accessible surrogate processes; the NLA's activation verbalizer then explains the surrogate's activations (Figure 1(a)). This approach is motivated by evidence of shared representational structure across independently trained models (Huh et al., 2024; Gröger et al., 2026). We hypothesize that processing the same trajectory elicits overlapping conceptual information across target and surrogate models, allowing surrogate-derived explanations to serve as useful proxies for aspects of the target's internal reasoning. This *surrogacy hypothesis* motivates their use in expert review without assuming identical internal computations across models. We test this hypothesis through controlled comparisons with target-model explanations and downstream legal-error classification.

Expert annotations reveal variation in the perceived usefulness of internal explanations: 41.4% of explanation ratings were *helpful*, 57.0% *neutral*, and 1.6% *misleading*. In downstream classification of expert-identified legal errors, aligned surrogate explanations improve macro-F1 from 0.257 to 0.324 over a no-explanation baseline (Figure 1(b)). They also outperform text-only and shuffled-explanation controls (Section 3.2). These findings support the practical value of incorporating internal explanations into legal reasoning assessment. Complementary cross-model experiments examine the basis for surrogacy: surrogate explanations show greater semantic agreement with target-model explanations within the evaluated model family than across families, while activation-reconstruction similarity alone does not consistently reflect semantic agreement.

Our contributions are fourfold:

1. **Activation-informed expert annotation framework.** We integrate explanations of internal model processing into expert review, collecting legal-error annotations, written feedback, and explanation-helpfulness ratings.
2. **An expert-annotated legal reasoning dataset.** We collect 354 span-level annotations across 15 conversational and agentic legal trajectories, with explanation-helpfulness ratings for eight trajectories.
3. **Cross-LLM Activation Surrogacy.** We reuse an existing NLA through a surrogate model, avoiding target-activation access and per-target NLA training, and empirically evaluate correspondence with target explanations.
4. **Controlled evidence of downstream utility.** We show that aligned surrogate explanations improve legal-error classification over no-explanation, text-only, and shuffled-explanation controls.

2 Activation-Informed Expert Annotation Framework

Our framework supports expert review of LLM trajectories using both generated text and surrogate activation explanations (Figure 1(a)). A target model generates a trajectory, which an accessible surrogate processes to produce activation-derived explanations through its NLA. We summarize token-level explanations into claims linked to their source explanations and present them alongside the trajectory for expert review. The framework collects two complementary annotation streams: (1) localized legal-error labels with written feedback, and (2) helpfulness ratings of the explanation claims. This design connects expert assessment of legal reasoning with assessment of the activation-derived information supplied during annotation.

2.1 Trajectory Collection

We collect trajectories from PRBench (Akyürek et al., 2026) and Harvey's LAB (Harvey AI, 2026), using Kimi-K2.6 as the target model (Moonshot AI, 2026). PRBench supplies single- and multi-turn conversational legal tasks: the model answers either a standalone legal question or the latest question in an ongoing conversation. LAB supplies longer-horizon agentic tasks involving input documents and model-produced documents.

We select tasks to match the practice areas of our three legal experts by conducting a survey prior to annotation work. To keep the review context manageable, we restrict multi-turn PRBench tasks to at most five turns of preceding context and LAB tasks to at most five input documents. For each LAB task, the review material includes the input documents and one or more model-produced output documents, alongside the trajectory.

2.2 Surrogate Explanation Generation

We use Gemma3-27B-IT (Team et al., 2025) as the surrogate for Kimi-K2.6, together with its publicly released NLA checkpoint (Fraser-Taliente et al., 2026; Lin, 2023). Each target-generated trajectory is passed through the surrogate, and the NLA's activation verbalizer produces natural-language explanations of the surrogate's token activations. The NLA also includes an activation reconstructor, which maps explanations back into activation space and supports the reconstruction-based evaluation described in Section 3.1. The explanations supplied during annotation describe the surrogate's processing of the target trajectory; their correspondence to target-model explanations is evaluated separately through controlled cross-model experiments (Section 3.1).

To make these explanations manageable for expert review, we divide each LLM turn into sentence-based spans and use Kimi-K2.6 to summarize the associated token-level explanations into one to five claims per span. Each claim cites the tokens whose explanations support it, preserving the connection between the summary and its activation-derived sources. This design addresses experts' initial feedback that reviewing hundreds of individual token explanations per trajectory was overwhelming.

2.3 Expert Annotation Protocol

Experts review each trajectory through an interface that provides the generated reasoning and answers, relevant documents, LLM-judge scores and rubric-based feedback, and summarized surrogate explanation claims. We collect two complementary forms of annotation:

1. **Legal-error annotations.** Experts highlight erroneous passages in the generated reasoning, answers, or output documents and assign one or more categories: *wrong rule*, *missed rule*, *fact misread*, *misapplied rule*, *incomplete reasoning*, or *other*. They provide written feedback explaining the error, the basis for their judgment, and an appropriate correction where applicable. When selecting *other*, experts are asked to describe the additional error category.
2. **Explanation-helpfulness ratings.** Experts rate individual summarized claims as *helpful* when they add to their understanding of the trajectory, *neutral* when they merely describe the model's observable behavior, or *misleading* when they judge the claim to introduce information absent from, or inconsistent with, the model's verbalized reasoning. These ratings capture experts' assessments of the explanations in the review context.

2.4 Dataset and Annotation Analysis

The dataset comprises 15 trajectories (nine from PRBench and six from LAB) with 19 trajectory reviews by three legal experts. Annotation was conducted over 7 weeks, requiring 34 hours of combined expert review time. Table 1 summarizes legal-error categories, Figure 2 explanation-helpfulness ratings, and Appendix Table 6 task coverage. We collect 354 span-level legal annotations across generated reasoning, answers, and output documents. Because a span can receive multiple error labels, category counts can exceed the number of annotated spans. Explanation-helpfulness reviews cover eight trajectories, with 62 claims rated by two experts. We report coverage separately for the two annotation streams.

Table 1: Legal errors: 381 label assignments on 354 annotated spans. Percentages use 354 spans as the denominator; labels may overlap.

	Span Location				
Error Type	**Reasoning**	**Answer**	**Document**	**Total**	**%**
Wrong rule	16	19	1	36	10.2
Missed rule	11	2	1	14	4.0
Fact misread	39	8	9	56	15.8
Incomplete reasoning	**93**	**53**	**34**	**180**	**50.8**
Misapplied rule	20	9	14	43	12.1
Other	45	7	0	52	14.7
Label assignments	**224**	**98**	**59**	**381**	–
Annotated spans	**207**	**95**	**52**	**354**	–

Most explanation ratings are neutral (Figure 2), followed by helpful, with misleading ratings least common. On the 62 claims rated by two experts, their judgments agree on 51 claims (82.26%): 39 are jointly rated neutral and 12 helpful. The remaining 11 claims receive differing judgments: eight helpful–neutral pairs and three helpful–misleading pairs. This agreement describes the doubly rated subset of explanation claims; the legal-error annotation agreement is not measured by this statistic.

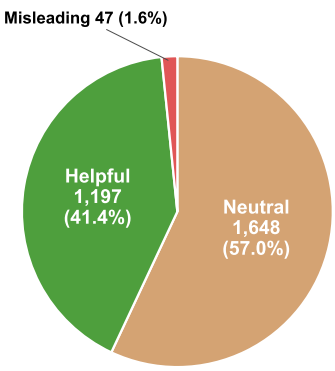
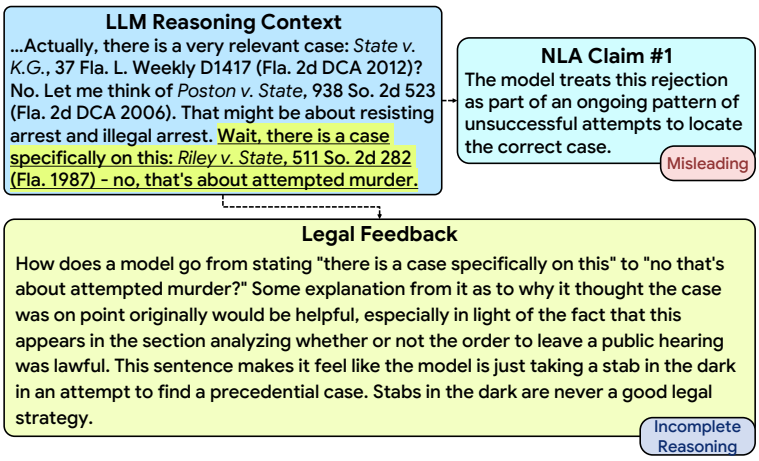


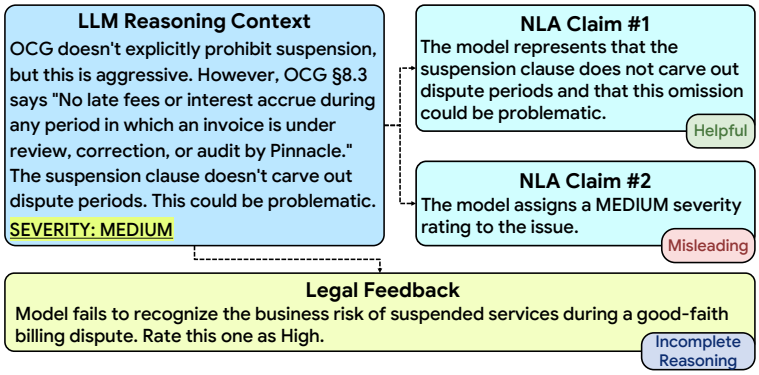
Figure 2: Distribution of NLA claim ratings ($N = 2,892$ reviews across $2,830$ unique claims).

Figure 3 illustrates the two annotation streams. In the litigation example (Figure 3a), the expert identifies *incomplete reasoning* where the model proposes and rejects precedents without explaining their relevance and rates the accompanying explanation claim as *misleading*. In the healthcare example (Figure 3b), the expert judges that the model underestimates the risk of suspended services during a billing dispute. The expert rates the claim highlighting the absence of an exception for dispute periods as *helpful* but rates the claim reporting the model's "medium" severity assessment as misleading.

The latter claim accurately reports the severity stated in the visible reasoning, although the expert considers that severity assessment incorrect. This example suggests that explanation-helpfulness



(a) **Litigation and Dispute Resolution Example**: The model randomly proposing and rejecting multiple legal precedents without explaining their relevance.



(b) **Healthcare and Life Sciences Example**: The model underestimating the risk of services being suspended during a billing dispute.

Figure 3: Two legal and NLA annotation examples from collected dataset on different legal tasks.

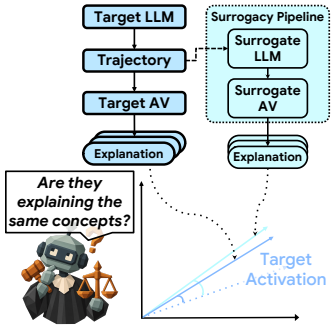
judgements can reflect disagreement with the underlying legal assessment, even when an explanation accurately describes the generated text. It illustrates the importance of distinguishing legal correctness, consistency with visible reasoning, and perceived explanation helpfulness when interpreting the collected ratings.

3 Experiments and Results

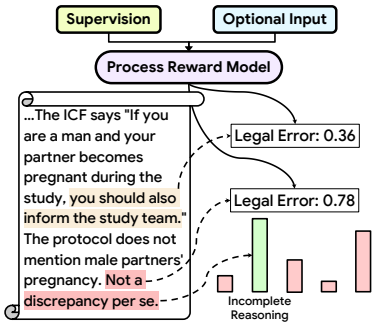
We evaluate the *surrogacy hypothesis* through cross-model explanation correspondence and downstream classification utility in our activation-informed expert annotation framework (Figure 4). First, we compare surrogate-derived explanations with target-model NLA explanations where both models' activations are accessible. Second, we test whether surrogate explanations improve classification of expert-identified legal errors. Additional analyses examine the effects of supervision sources and detection objectives.

3.1 Evaluating Cross-LLM Activation Surrogacy

We evaluate all six directed target–surrogate pairs among Gemma3-12B-IT, Gemma3-27B-IT, and Qwen2.5-7B-IT (Yang et al., 2024), using publicly released NLA checkpoints (Fraser-Taliente et al., 2026; Lin, 2023). Access to both models' activations allows controlled comparisons within the Gemma family and across the Gemma and Qwen families. We measure correspondence through target-activation reconstruction and semantic agreement between explanations.



(a) **Surrogacy correspondence.** Target-activation reconstruction and semantic agreement between target and surrogate NLA explanations.



(b) **Downstream utility.** Legal annotations provide supervision; surrogate NLA explanations are optional inputs for classifying known errors. Supporting analyses examine supervision sources and passage-level detection.

Figure 4: Overview of our experimental evaluation.

We sample eight token activations per generation from 200 WildChat and 300 LegalBench examples, yielding 4,000 activations (Zhao et al., 2024; Guha et al., 2023). Semantic evaluation uses explanations for 500 tokens from LegalBench generations. These validation datasets are distinct from the PRBench and LAB trajectories used for expert annotation (Section 2).

Target-activation reconstruction. For each target–surrogate pair, the surrogate processes target-generated text, and each model's activation verbalizer (AV) explains its own activations. We pass both explanations through the target's activation reconstructor (AR_{target}) and measure cosine similarity between each reconstruction and the original target activation. We compare reconstruction from surrogate explanations with target explanations and a shuffled-explanation control. This tests whether surrogate explanations contain information the target reconstructor can use to recover its activations.

Semantic agreement. Kimi-K2.6 judges whether target and surrogate explanations give consistent, partially consistent, or different interpretations of the token's context and legal concepts. For each pair, we obtain three judgments with randomized explanation order and model identities hidden, then assign the majority label. We report label counts and the proportion judged consistent or partially consistent. This measures agreement in explanation content, rather than establishing faithfulness to the target model's computation.

Table 2: Surrogacy evaluation across model pairs: Target-activation reconstruction cosine similarity scores and blinded LLM-judge concept agreement on **LegalBench**. Explanations with *unclear* verdict after tie are not considered. Shading in the last column scales with agreement; shuffled explanations (gray) are a control.

	Reconstruction Similarity			Legal Concept Semantic Agreement			
Target → Surrogate	Target	Surrogate	Shuffled	Same	Partial	Diff.	Same+Partial
Within family							
Gemma-27B → Gemma-12B	99.12%	98.81%	95.34%	317	79	102	79.5%
Gemma-12B → Gemma-27B	99.50%	99.44%	97.51%	330	79	91	81.8%
Cross family							
Qwen-7B → Gemma-12B	89.32%	84.06%	35.55%	196	90	214	57.2%
Qwen-7B → Gemma-27B	89.32%	85.13%	35.80%	182	91	227	54.6%
Gemma-12B → Qwen-7B	99.50%	99.17%	97.63%	174	119	207	58.6%
Gemma-27B → Qwen-7B	99.12%	98.40%	95.58%	186	97	216	56.7%

Results. Table 2 reveals three patterns:

1. **Stronger correspondence within the evaluated family.** Gemma pairs achieve reconstruction similarities of 98.81–99.44%, with 79.5–81.8% of explanation pairs judged consistent or partially

consistent. This supports within-family surrogacy, including using the smaller Gemma model as a surrogate for the larger one.

2. **Weaker semantic correspondence across families.** Combined semantic agreement falls to 54.6–58.6% for Gemma–Qwen pairs. Reconstruction is asymmetric: similarity remains high with Gemma as the target and Qwen as the surrogate, but is lower in the reverse direction.
3. **Reconstruction similarity alone is insufficient.** Shuffled explanations also yield high reconstruction similarity for Gemma targets. These scores therefore need to be interpreted alongside shuffled controls and semantic agreement.

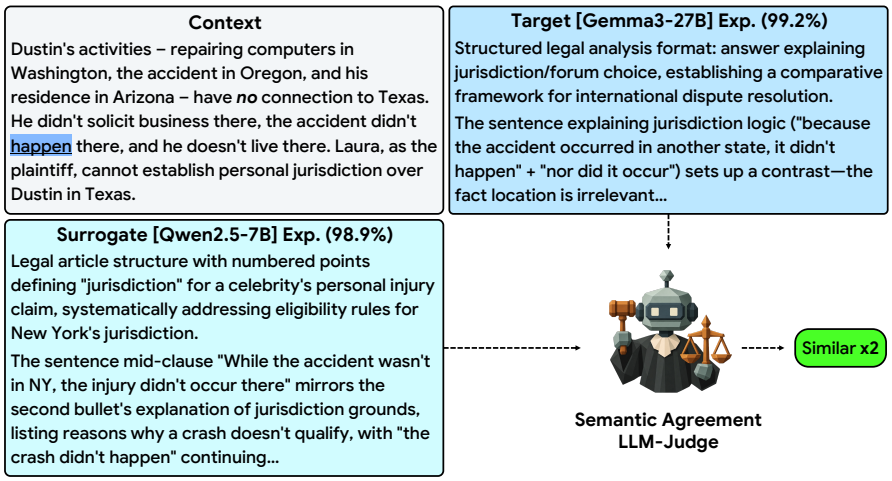


Figure 5: Cross-family surrogacy on a LegalBench personal-jurisdiction task. Gemma3-27B-IT generates the target response (deep blue); Qwen2.5-7B-IT serves as the surrogate (cyan).

Figure 5 illustrates correspondence on broad legal concepts alongside a jurisdiction-specific discrepancy. For the token *" happen"*, the Qwen explanation yields high reconstruction similarity and is judged semantically similar to Gemma's, but substitutes New York for Texas. The judge treats this as "surface-level variation," although jurisdictional distinctions can matter for legal interpretation. The example highlights the complementary roles of automated evaluation and expert review.

These results support the feasibility of Cross-LLM Activation Surrogacy, with stronger explanation correspondence within the evaluated model family than across families. We next assess whether surrogate explanations improve downstream legal-error classification.

3.2 Utility of Activation-Derived Explanations

We test whether surrogate activation explanations improve classification of known legal errors. Using the expert annotations collected through our framework, the model assigns one or more of five categories to each erroneous span: *incomplete reasoning*, *fact misread*, *wrong rule*, *misapplied rule*, and *missed rule*. This evaluates error diagnosis; locating errors automatically is examined in Section 3.3.

We compare five input conditions on 280 expert-labeled error spans from 14 document groups. Each group is held out in turn for leave-one-document-group-out evaluation, with a frozen backbone and three training seeds. All conditions share the training setup and evaluation splits. Expert-reviewed NLA claims are available for 96 spans; the remaining 184 receive identical task-and-span inputs without explanations in every condition. Reported performance covers all 280 spans.

The input conditions are:

- **A: No explanation.** The model receives only the task-and-span input.
- **B: Text-only explanation.** An explanation is generated from the observable task and span text.
- **C: Aligned NLA.** The span's surrogate activation claims are supplied without helpfulness filtering.
- **D: Helpfulness-selected NLA.** Aligned claims are filtered by a helpfulness model trained only on judgments from training-document groups.

- **E: Shuffled NLA.** Claims are shuffled within each document and matched for token budget.

C versus A tests the benefit of adding NLA explanations; C versus B compares them with text-only explanations; C versus E tests whether alignment with the span matters. D versus C tests whether helpfulness filtering improves on unfiltered claims.

Results. Aligned NLA explanations improve macro-F1 from 0.257 to 0.324 over the no-explanation baseline (Table 3), with gains in all three seeds. The paired document-bootstrap 95% interval for the improvement is $[0.0195, 0.1009]$. Document binary cross-entropy (BCE; lower is better) decreases by 0.3348 (95% interval for the difference: $[-0.6739, -0.0221]$).

Table 3: **NLA-input comparison.** All conditions use the same 280 labeled error spans; 96 have aligned NLA claims. Best in bold; the highlighted row uses aligned surrogate NLA explanations.

Input condition	Macro-F1 ↑	Doc. BCE ↓
A: No explanation	0.2568	2.0631
B: Text-only explanation	0.2596	1.8862
C: Aligned NLA	**0.3243**	**1.7283**
D: Helpfulness-selected NLA	0.3041	1.7845
E: Within-document shuffled NLA	0.2679	1.9314

In secondary comparisons, aligned NLA explanations outperform text-only explanations and shuffled claims, with macro-F1 difference intervals of $[0.0166, 0.0958]$ and $[0.0218, 0.0879]$, respectively. These results support the utility of activation-derived explanations and the importance of their alignment with the assessed span.

Helpfulness-based filtering yields a lower macro-F1 than unfiltered claims (0.304 versus 0.324), providing no additional benefit in this setting. We note that perceived helpfulness during expert review and usefulness for automated error classification are distinct evaluation criteria. And this experiment tests predicted helpfulness as a claim-selection criterion and does not measure whether explanations improve experts' judgments.

3.3 Additional Analyses of Supervision and Detection

Table 4: Leave-one-trajectory-out performance on expert-annotated spans. Intervals are 95% sample-cluster bootstrap intervals.

Condition	Macro-F1 ↑
PRM-S: synth. supervision	0.200 [0.102, 0.257]
PRM-M: synth. + expert labels	0.178 [0.084, 0.241]
PRM-P: synth. + judge labels	0.219 [0.112, 0.274]

We examine supervision sources and joint training for error detection and diagnosis through additional process reward modeling (PRM) experiments (Uesato et al., 2022; Wang et al., 2024; Lightman et al., 2024). These experiments use expert annotations for evaluation and omit NLA explanations from model inputs. Each uses a separate evaluation subset and protocol, described below.

Sources of supervision. We train a Qwen3-4B-Instruct-2507 backbone with a newly initialized five-label linear head applied to the final-token representation of a fixed assessment suffix. Inputs contain the task, span location, local context, and span text. We compare three training conditions:

- **PRM-S: Synthetic supervision.** Synthetic annotations are supplemented with "silver negatives": passages that do not overlap annotated spans and are provisionally assigned none of the five error labels.
- **PRM-M: Synthetic plus expert supervision.** Expert-labeled spans from the training portion of the 233-span dataset are added to PRM-S's supervision.
- **PRM-P: Synthetic plus judge supervision.** The same expert-selected training spans are added, with error labels supplied by Claude Sonnet instead of the experts.

PRM-M and PRM-P differ only in the source of error-category labels; span selection and training volume are held fixed. Both use expert-selected spans, so this comparison does not test the value of expert error localization.

We evaluate on 233 expert annotations across ten trajectories, holding out each trajectory in turn. PRM-S, PRM-M, and PRM-P achieve macro-F1 scores of 0.200, 0.178, and 0.219, respectively (Table 4). All paired difference intervals include zero, so no supervision condition shows a clear advantage in this setting.

Joint error detection and diagnosis. We next assess errors without supplying expert-highlighted span locations. We divide each trajectory into sentence-like candidate passages. For each passage, a joint model predicts whether it contains an error and, conditional on an error, its categories. Category predictions are emitted only when the detector flags the passage.

We compare the five-label classifier with three joint models that differ in their treatment of unannotated passages. Joint0 excludes these passages from the detection loss. Joint01 and Joint03 treat them as provisionally acceptable examples, with loss weights of 0.1 and 0.3, respectively. This tests the effect of using unannotated passages as weak negative examples.

Results. On 40 held-out expert-identified errors, joint models achieve conditional classification F1 scores of 0.160–0.190, compared with 0.018 for the five-label baseline (Table 5). Requiring the detector to also flag the passage reduces Joint0's F1 from 0.160 to 0.073 (gated F1). Increasing the detection-loss weight on unannotated passages reduces recall in every seed: treating these passages as acceptable can cause the model to miss more known errors. We therefore use Joint0, which excludes them from the detection loss, as the reference joint formulation. Detection remains a bottleneck even when conditional diagnosis improves.

Table 5: Passage detection and diagnosis on held-out expert-identified errors. *Rows*: the five-label-only classifier and joint models weighting silver passages in the detection loss by 0, 0.1, and 0.3. *Columns*: conditional F1 evaluates labels on known-error passages; gated F1 additionally requires the detector to select the passage; detection recall is the fraction of known errors flagged. Mean $\pm$ standard deviation over three seeds; best mean per column in bold.

Model	**Conditional F1** $\uparrow$	**Gated F1** $\uparrow$	**Detection recall** $\uparrow$
Label-only classifier	0.018 ± 0.031	0.018 ± 0.031	0.117 ± 0.029
Joint0 (reference)	0.160 ± 0.039	0.073 ± 0.024	$\mathbf{0.458} \pm 0.277$
Joint01	0.186 ± 0.087	$\mathbf{0.079} \pm 0.089$	0.208 ± 0.166
Joint03	$\mathbf{0.190} \pm 0.083$	0.030 ± 0.036	0.133 ± 0.123

4 Conclusion

We introduced an activation-informed expert annotation framework that pairs model-generated legal trajectories with natural-language explanations of surrogate activations. Our Cross-LLM Activation Surrogacy approach reuses an existing NLA to explain an accessible surrogate's processing of a target trajectory, enabling annotation without access to the target's activations or a dedicated NLA for that target. Through this framework, legal experts contributed localized legal-error judgments, written feedback, and explanation-helpfulness ratings, providing a resource for studying how activation-derived information can support expert oversight.

Our experiments examine explanation correspondence and downstream utility in distinct settings. Among the open-weight models evaluated in Section 3.1, surrogate and target-model explanations show stronger semantic agreement within the Gemma family than across the Gemma and Qwen families, while reconstruction similarity alone does not reliably track semantic agreement. In the Kimi-K2.6-based expert dataset, aligned surrogate explanations improve classification of already identified legal errors over no-explanation, text-only-explanation, and shuffled-explanation controls. These results support the usefulness of aligned activation-derived information for legal-error diagnosis in the evaluated setting.

A key limitation concerns the target model used in the expert study. Kimi-K2.6's activations were inaccessible in our study, so it could not be included in the correspondence experiments; correspondence between its internal processing and the Gemma surrogate explanations therefore remains unmeasured. The classification gains demonstrate the utility of surrogate explanations, while their faithfulness to Kimi-K2.6's internal processing remains an open question. Limited activation access motivates our surrogacy approach, and broader studies with more experts and model pairs are needed to assess generalization. Thus, future work should evaluate additional target–surrogate pairs where activation access permits direct comparison and measure whether explanations improve experts' annotation accuracy and efficiency. Our framework provides a practical basis for investigating these questions and the role of activation-derived explanations in expert oversight of legal LLMs.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets of legal annotations to bootstrap and assist our evaluation of the collected human expert annotations and NLA explanations in their utility in downstream process reward modeling and for preprocessing the collected human expert data for process reward modeling (3.2 and 3.3). Please refer to Appendix sections B.2 and B.3 for the exact human annotation processing, synthetic data pipeline, and how initial human annotations were used to guide it and preprocessing human expert annotations.

We have not used generative AI tools for ideating research questions, objectives, methodology, or experiments; data analysis of the results; nor for any conceptual development. And mathematical proofs and claims are not applicable to this work.

Additionally, we used generative AI tools for creating initial surveys for human legal experts regarding their use of AI in their domain and how comfortable they were working on our initial annotation pipelines and for literature review on existing annotation methods and utilities. For manuscript preparation, we used Codex and Claude to assist with drafting LaTeX commands, polishing text and structure, and creating a few of the icons used in our figures. For software, we used Claude Code and Codex to support code implementation, including the annotation interface and backend and inference pipelines throughout the legal task generation and downstream experiments. All AI-assisted output was manually reviewed, validated, and refined by the authors. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies expert oversight of legal LLMs using benchmark tasks and model-generated trajectories. Its intended benefit is to help researchers and legal experts identify reasoning errors and assess the usefulness of activation-derived explanations. The framework and resulting models are research tools; the reported results do not establish suitability for autonomous legal advice or decisions affecting individuals' rights. Incorrect legal claims, missed errors, and misleading explanations could cause harm if relied on without independent professional review.

A central risk is unwarranted confidence in the explanations. Our surrogate explanations describe an accessible model's processing of a target trajectory, and their faithfulness to Kimi-K2.6's internal computation remains unverified. Experts also see generated claims and LLM-judge feedback, which may anchor their judgments. Explanation-helpfulness ratings therefore reflect judgments made in this assisted review setting. Improvements in classification of already identified errors do not establish reliable error detection, improved human decision-making, or faithful recovery of the target model's reasoning.

Three legal experts contributed approximately 34 hours of review and were compensated at approximately USD 100 per hour. We report their contributions using anonymized expert labels. The small expert sample and tasks selected to match their practice areas limit coverage of legal perspectives, jurisdictions, and affected populations.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix B document task selection, expert review, human-annotation normalization, and synthetic annotation generation. Section 3.1 and Appendix C specify the surrogate-explanation evaluation, including sampling, controls, and statistical tests. Section 3.2 describes reward-model training and evaluation; Appendix B.2 identifies the missing settings that limit exact replay of the human-normalization model calls. To foster future research, we publicly release our code and dataset at: https://anonymous.4open.science/r/tr3D8F.

REFERENCES

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph, IV, Mohammad Mahmoudi Meymand, Gurshaan Chattha, Paula Rodriguez, Diego A. Mares Buendia, Pavit Singh, Michael Liu, Subodh

Chawla, Peter Cline, Lucy Ogaz, Ernesto Gabriel Hernández Montoya, Zihao Wang, Pavi Bhatter, Marcos Ayestaran, Bing Liu, and Yunzhong He. PRBench: Large-scale expert rubrics for evaluating high-stakes professional reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42297–42325, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1958. URL https://aclanthology.org/2026.acl-long.1958/.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, March 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*, August 2024. URL https://openreview.net/forum?id=ePUVetPKu6.

Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, Arushi Somani, Peter Hase, Misha Wagner, Fabien Roger, Vlad Mikulik, Samuel R. Bowman, Jan Leike, Jared Kaplan, and Ethan Perez. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

Debarati Das, Khanh Chi Le, Ritik Sachin Parkar, Karin De Langis, Brendan Madson, Chad M. Berryman, Robin M. Willis, Daniel H. Moses, Brett McDonnell, Daniel Schwarcz, and Dongyeop Kang. LawFlow: Collecting and simulating lawyers' thought processes on business formation case studies. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=MsgdEkcLRz.

Kit Fraser-Taliente. NLA models. Hugging Face model collection, 2026. URL https://huggingface.co/collections/kitft/nla-models. Model collection maintained by kitft; accessed September 25, 2026.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. Natural language autoencoders produce unsupervised explanations of LLM activations. *Transformer Circuits Thread*, 2026. URL https://transformer-circuits.pub/2026/nla/index.html.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic Representation Hypothesis: An Aristotelian view. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.14486.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 44123–44279, 2023. doi: 10.52202/075280-1915. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/89e44582fd28ddfea1ea4dcb0ebbf4b0-Abstract-Datasets_and_Benchmarks.html.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. Which economic tasks are performed with AI? Evidence from millions of Claude conversations, 2025. URL https://arxiv.org/abs/2503.04761.

Harvey AI. Harvey lab: The legal agent benchmark, 2026. URL https://github.com/harveyai/harvey-labs/tree/v1.0. Announcement: https://www.harvey.ai/blog/introducing-harveys-legal-agent-benchmark.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic Representation Hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Adam Karvonen, James Chua, Clément Dumas, Kit Fraser-Taliente, Subhash Kantamneni, Julian Minder, Euan Ong, Arnab Sen Sharma, Daniel Wen, Owain Evans, and Samuel Marks. Activation oracles: Training and evaluating LLMs as general-purpose activation explainers. *arXiv preprint arXiv:2512.15674*, 2025. URL https://arxiv.org/abs/2512.15674.

Belinda Z. Li, Zifan Carl Guo, Vincent Huang, Jacob Steinhardt, and Jacob Andreas. Training language models to explain their own computations. *arXiv preprint arXiv:2511.08579*, 2025. URL https://arxiv.org/abs/2511.08579.

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 *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Johnny Lin. Neuronpedia: Interactive reference and tooling for analyzing neural networks, 2023. URL https://www.neuronpedia.org. Software available from neuronpedia.org.

Miles McCain, Thomas Millar, Saffron Huang, Jake Eaton, Kunal Handa, Michael Stern, Alex Tamkin, Matt Kearney, Esin Durmus, Judy Shen, Jerry Hong, Brian Calvert, Jun Shern Chan, Francesco Mosconi, David Saunders, Tyler Neylon, Gabriel Nicholas, Sarah Pollack, Jack Clark, and Deep Ganguli. Measuring AI agent autonomy in practice. Anthropic, February 2026. URL https://www.anthropic.com/research/measuring-agent-autonomy.

Moonshot AI. Kimi K2.6: Advancing open-source coding. Moonshot AI, April 2026. URL https://www.kimi.ai/blog/kimi-k2-6. Accessed September 25, 2026.

Rajiv Movva, Smitha Milli, Sewon Min, and Emma Pierson. What's in my human feedback? learning interpretable descriptions of preference data. In *International Conference on Learning Representations*, volume 2026, pp. 59420–59455, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/613c72b51ebd8f44cd4126252348c624-Abstract-Conference.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Alexander Pan, Lijie Chen, and Jacob Steinhardt. LatentQA: Teaching LLMs to decode activations into natural language. In *International Conference on Learning Representations*, volume 2026, pp. 78726–78759, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7f19b99e63762d20e9df91144056f1ee-Abstract-Conference.html.

Gonçalo Santos Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48393–48421. PMLR, 2025. URL https://proceedings.mlr.press/v267/paulo25a.html.

Keenan Pepper, Alex McKenzie, Florin Pop, Stijn Servaes, Martin Leitgab, Mike Vaiana, Judd Rosenblatt, Michael S. A. Graziano, and Diogo de Lucena. Learning self-interpretation from interpretability artifacts: Training lightweight adapters on vector-label pairs. *arXiv preprint arXiv:2602.10352*, 2026. URL https://arxiv.org/abs/2602.10352.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models, 2024. URL https://arxiv.org/abs/2407.02646.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Ke-Ming Lu, Ke-Qin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhen-Ru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *International Conference on Learning Representations*, volume 2024, pp. 34590–34605, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/9421261e06f1a63a352b068f1ac90609-Abstract-Conference.html.

A Related Work

Expert annotation and process supervision. Human feedback supports language-model alignment through demonstrations and preference judgments over generated responses, as in InstructGPT (Ouyang et al., 2022). Process supervision extends feedback to individual reasoning steps, enabling reward models to assess intermediate reasoning (Lightman et al., 2024). In the legal domain, LawFlow collects workflows from trained law students to study legal reasoning and compare human and model behavior (Das et al., 2025). Complementing these collection efforts, sparse autoencoders have been used to identify interpretable features that characterize response differences and annotator preferences in existing feedback datasets (Movva et al., 2026). Our framework incorporates activation-derived explanations into the annotation process itself: legal experts inspect trajectories alongside surrogate explanations and provide both localized legal-error judgments and explanation-helpfulness ratings. This connects expert process supervision with explicit assessment of the interpretability signals available during review.

Activation verbalization. Sparse autoencoders decompose model activations into sparsely activated features, whose meanings can be investigated through activating examples and automated natural-language descriptions (Huben et al., 2024; Paulo et al., 2025). Activation verbalization methods instead generate natural-language descriptions or answers directly from activation inputs. LatentQA trains a decoder on activation–question–answer examples (Pan et al., 2026), while Activation Oracles extend this approach through diversified training and evaluation across tasks (Karvonen et al., 2025). Self-interpretation methods also explore decoding activations with a frozen language model, including lightweight adapters trained on vector–label pairs (Pepper et al., 2026). Natural Language Autoencoders (NLAs) learn to verbalize activations through a natural-language bottleneck from which a reconstructor recovers the original activation; following supervised initialization, the verbalizer and reconstructor are trained with an activation-reconstruction objective (Fraser-Taliente et al., 2026). We use existing NLAs to supply activation-derived explanations within our expert annotation framework, and study how domain experts assess these explanations and how they contribute to downstream legal-error classification.

Cross-model representation alignment and explanation transfer. The Platonic Representation Hypothesis (Huh et al., 2024) proposes that independently trained models converge toward shared representational structure. Subsequent work identifies model-scale confounders in similarity measurements and finds that agreement in local neighborhood structure persists after calibration, while apparent global convergence is substantially reduced (Gröger et al., 2026). These findings motivate investigating cross-model explanation transfer, while leaving its reliability on individual trajectories an empirical question. Related work trains language models to explain their own computations using supervision derived from interpretability tools, finding that self-explanation generally outperforms explanation by a different model (Li et al., 2025). This finding cautions against assuming that explanations transfer reliably across models. Our surrogacy approach reuses an existing NLA by passing target-generated trajectories through a surrogate and explaining the surrogate's own activations. We therefore evaluate the correspondence and utility of these proxy explanations through target-activation reconstruction, semantic agreement with target explanations, and downstream legal-error classification.

B Data collection and annotation details

The expert study covers 15 trajectories and the synthetic training set of 38 more trajectories; expert feedback was also standardized for downstream experiments.

B.1 Expert review and coverage

Review material and protocol. To match the practice areas of three legal experts, we selected nine PRBench (Akyürek et al., 2026) and six Harvey LAB (Harvey AI, 2026) trajectories. PRBench multi-turn tasks had at most five preceding turns; LAB tasks had at most five input documents. Kimi-K2.6 (Moonshot AI, 2026) generated the trajectories. The experts saw the task materials, model reasoning and answers, any output documents, rubric-based judge feedback, and summarized natural-language

Table 6: Expert-reviewed trajectories by benchmark and primary legal area.

Benchmark	Primary legal area	Trajectories
PRBench	Regulatory and administrative law	3
	Civil rights	2
	Criminal law	2
	Intellectual property	1
	Family law	1
LAB	Healthcare and life sciences	2
	Litigation and dispute resolution	2
	Corporate M&A	1
	Intellectual property	1
Total		15

Table 7: Active expert-review hours by task type and primary legal area through September 20, 2026. A dot indicates no review.

Primary legal area	A	B	C	Total hours	Trajectories	Reviews
PRBench: single-turn						
Regulatory and administrative law	0.75	1.36	·	2.11	2	3
Civil rights	1.01	·	·	1.01	1	1
Criminal law	0.43	·	·	0.43	1	1
Intellectual property	·	·	0.78	0.78	1	1
Family law	0.74	·	·	0.74	1	1
PRBench: multi-turn						
Regulatory and administrative law	1.10	1.26	·	2.36	1	2
Civil rights	·	1.69	·	1.69	1	1
Criminal law	1.28	·	·	1.28	1	1
LAB: agentic						
Healthcare and life sciences	·	7.05	·	7.05	2	2
Litigation and dispute resolution	3.50	2.08	2.29	7.88	2	3
Corporate M&A	·	3.93	·	3.93	1	1
Intellectual property	2.53	2.49	·	5.02	1	2
Total	11.33	19.87	3.07	34.27	15	19

autoencoder (NLA) explanations. They marked legal-error spans and wrote feedback. Separately, they rated the helpfulness of individual NLA claims.

We passed the expert-reviewed trajectories through Gemma-3-27B-IT as a surrogate. Its public activation verbalizer (`kitft/nla-gemma3-27b-L41-av`) produced per-token explanations at temperature 0.7 with at most 160 new tokens. Kimi-K2.6 summarized them into one to five token-cited claims per span at temperature 0 with at most 16,384 output tokens. Synthetic span annotation did not use these claims.

Coverage and time. The 15 trajectories received 19 reviews over seven weeks; four had a second review. Experts logged 34.27 active hours. Table 6 counts each trajectory once under its primary legal area, even if the task has several practice-area tags.

Table 7 reports active time by the trajectory's primary legal area. The timer excludes pauses but does not separate legal-span work from NLA-claim assessment. Totals use unrounded seconds, so displayed cells may differ from a displayed total by 0.01 hour.

Annotation output. Experts marked 354 legal spans: 207 in reasoning, 95 in answers, and 52 in output documents (Table 8). Some spans had more than one error label, yielding 381 label assignments. Experts also made 2,892 NLA-claim reviews on 2,830 distinct claims; two experts reviewed 62 of those claims. Counts use each annotation's latest non-cleared state through September 20.

Table 8: Expert annotation output through September 20, 2026. Legal-span columns count annotated spans; the NLA column counts claim reviews.

Expert	Reviews	Reasoning	Answer	Document	Legal total	NLA reviews	Active hours
A	8	67	28	6	101	721	11.33
B	9	125	57	39	221	2,108	19.87
C	2	15	10	7	32	63	3.07
Total	19	207	95	52	354	2,892	34.27

B.2 Processing expert span annotations

For downstream experiments, we converted experts' free-text span feedback into structured fields. We kept their labels, selected text, and span boundaries. The source included current annotations and an earlier pilot; for each annotation, we used its latest active version. We excluded cleared or superseded records, NLA-claim ratings, sample preferences, and document ratings. We checked every selected span against the frozen text and token boundaries.

For each span, the normalizer saw the expert's annotation, task, input documents, and trajectory content through the end of the selected span. It did not see later output, NLA summaries, trajectory-judge output, or external search results. It could rephrase feedback but had to preserve the expert's legal propositions, labels, actors, addressees, obligations, permissions, and certainty without adding a legal judgment. A separate verifier checked semantic preservation, support in the supplied context, and the exact span anchor. The verifier input concealed the normalizer's identity.

We calibrated the prompts on 16 development examples using GPT-5.6-Sol. The bulk campaign used GPT-5.6-Terra at high reasoning effort; later collections used GPT-5.6-Luna at high or maximum effort. Bulk and refresh campaigns used version-3 normalizer and verifier prompts. We validated outputs against the annotation schema and reviewed cases needing editorial repair or legal judgment before promotion. Sealed test examples were excluded.

Output fields and replay limits. Each processed annotation has a finding and basis, with a prediction or recommended correction when applicable. The error labels are wrong rule, missed rule, misapplied rule, fact misread, incomplete reasoning, and other; other requires the expert's custom label. Sound reasoning, descriptive reasoning, and insufficient signal are non-error labels and cannot appear with an error label. The record also keeps the selected span and validated offsets. This downstream subset has a different cutoff from the 354 expert annotations reported above.

Campaign records include the deployed model, reasoning effort, service tier, prompt versions, and outputs. They omit temperature, maximum output tokens, maximum turns, API surface, and hosted checkpoint revision for the external normalization sessions. The saved records show the transformations and checks but cannot reproduce the model calls exactly.

B.3 Synthetic span-annotation generation

Task selection and trajectories. For synthetic annotation, we selected PRBench and LAB tasks in U.S. federal and state law that required review, analysis, comparison, or extraction. PRBench multi-turn context had at most five turns; LAB tasks had at most five input documents and retained the benchmark's 200-turn agent limit. Kimi-K2.6 generated trajectories through chat completions at temperature 0 with at most 32,768 output tokens per call.

Selection and annotation. Candidate spans came from model reasoning, answers, and output documents. They followed sentence, list-item, table-row, or paragraph boundaries and contained between 16 and 512 tokens without splitting a token. For each trajectory, a selector chose up to 30 material spans using the task, input documents, and trajectory content available through each candidate span. It did not see NLA explanations or trajectory-judge results. Input documents were shown in full up to 400,000 characters; longer documents were split into 12,000-character chunks.

For each selected span, a proposer wrote labels, a finding, and a basis, adding a prediction or recommended correction when supported. It could make up to three primary-law searches, open

results from permitted official legal sources, and read task documents within ten tool rounds. A separate verifier then passed the annotation, requested one revision, sent it for legal review, or rejected it. A new verifier session checked any revision. Accepted records cite supporting task text or retrieved evidence.

The proposer and revision step used eight approved few-shot examples from normalized human annotations in the training partition. Together, they covered all six error labels and both reasoning and answer spans. Each showed the task, selected text, and at most 6,000 characters of model text available through the selected span, without input-document excerpts. The selector and verifier received no examples, and development records were not used. The examples demonstrated annotation content and format, not legal authority for other tasks.

Kimi-K2.6 handled selection, proposal, and verification through chat completions at temperature 0.5 with at most 32,768 output tokens per call. Primary-law search used GPT-5.4-nano through the Responses API at temperature 0, `xhigh` reasoning, and at most 5,120 output tokens.

Release and validation. An annotation entered the synthetic training export only after verification and release checks. We excluded exact duplicate targets or content and clustered near duplicates without automatically removing them. The release contains 759 annotations on 38 trajectories: 473 from 17 LAB trajectories and 286 from 21 PRBench trajectories. All 759 are in the training partition. We saved cited web evidence as snapshots. Each record has labels, feedback, token and character offsets for its span, and evidence references. Export validation checks that the offsets reproduce the frozen source text.

Human and synthetic annotations share a span format and have separate releases. Archived task IDs, trajectories, prompts, few-shot examples, model settings, span-selection rules, verifier decisions, evidence snapshots, code, and environment specifications document how the synthetic export was built.

B.4 Annotation interface

The annotation interfac used to inspect and review the legal trajectories, showing the trajectory (center), NLA claims and citations (right), and expert annotations (yellow underlines).

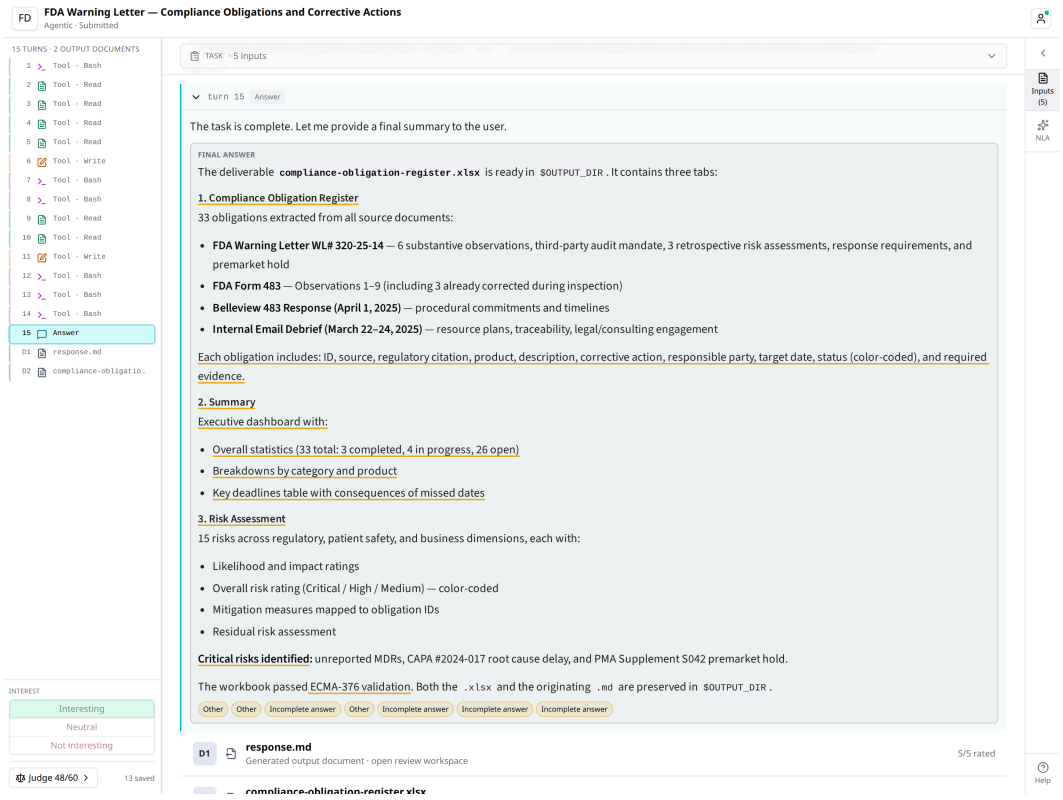


Figure 6: Submitted FDA Warning Letter assignment. Underlines mark saved legal span annotations in the final answer.

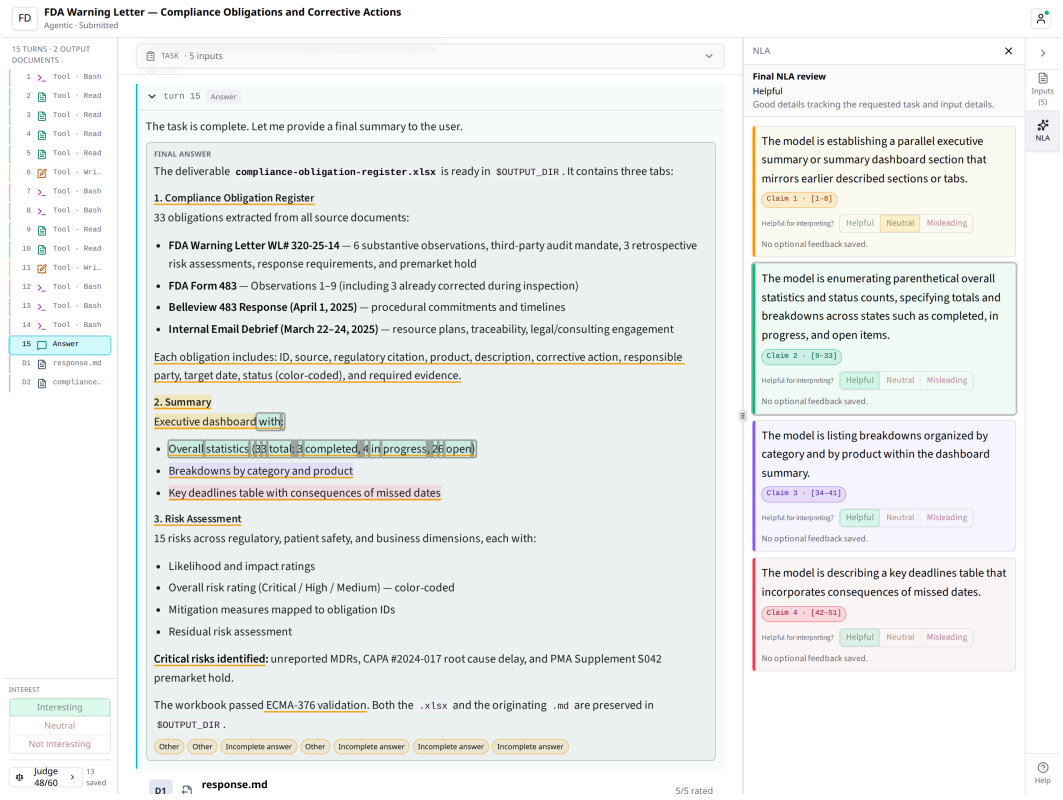


Figure 7: NLA review of the same assignment. Claim ratings appear beside the highlighted answer text.

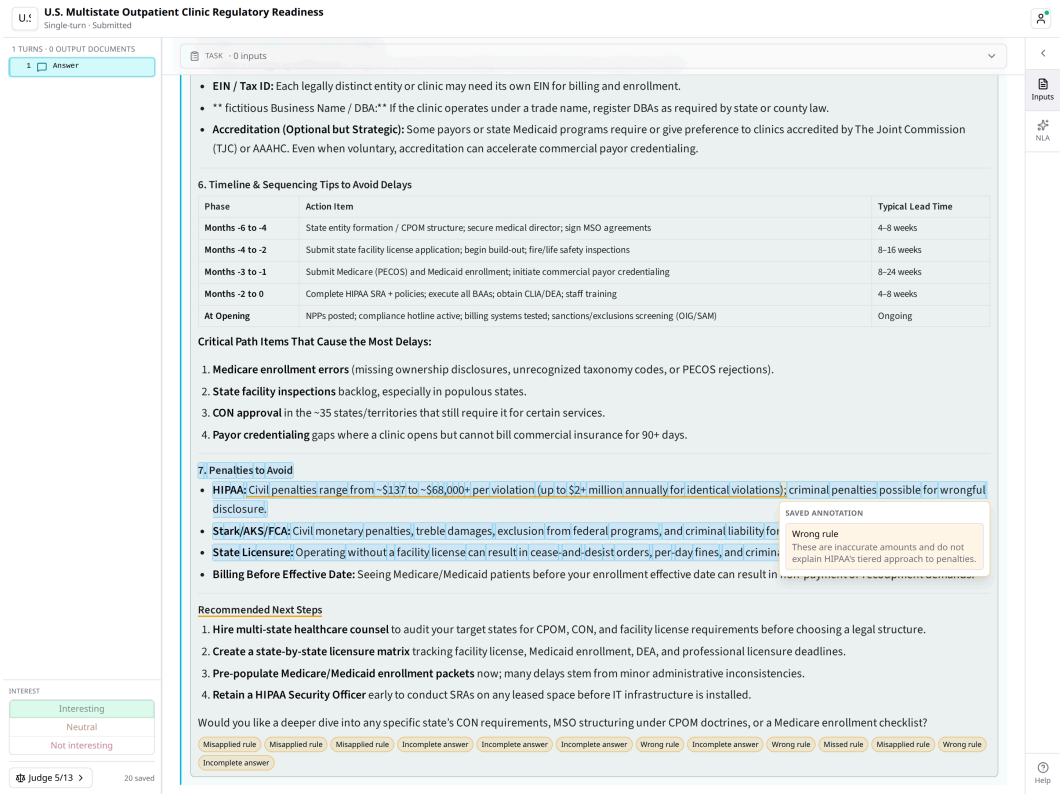


Figure 8: Saved legal span feedback on a clinic-regulation answer. The annotator flags inaccurate HIPAA penalty figures and the missing tiered structure.

Table 9: Models and public NLA configurations. Here ℓ is the residual extraction layer and d is the residual width.

Model	ℓ	d
Gemma-3-12B-IT	32	3840
Gemma-3-27B-IT	41	5376
Qwen2.5-7B-Instruct	20	3584

C Surrogate Explanation and Validation Details

C.1 Surrogate-NLA reconstruction

We test whether a surrogate model's activation explanation can recover the target model's activation at the same position in a generated response. We use the released Natural Language Autoencoder (NLA) activation verbalizers (AVs) and activation reconstructors (ARs) without further training.

Scoring direction. At a selected assistant token t, let $h_A^{(t)}$ and $h_B^{(t)}$ be the residual activations of target model A and surrogate model B at their NLA layers. With AV_M and AR_M denoting model M's verbalizer and reconstructor, we compare

$$\text{target explanation:} \quad r_{AA}^{(t)} = \mathrm{AR}_A\Big(\mathrm{AV}_A(h_A^{(t)})\Big), \tag{1}$$

$$\text{surrogate explanation:} \quad r_{AB}^{(t)} = \mathrm{AR}_A\Big(\mathrm{AV}_B(h_B^{(t)})\Big). \tag{2}$$

We score both reconstructions against $h_A^{(t)}$. A surrogate explanation counts as successful only if the target AR maps it toward the target activation. We do not use reconstruction in the surrogate's activation space as evidence of surrogacy.

Models. We evaluate all six directed pairs among Gemma-3-12B-IT, Gemma-3-27B-IT, and Qwen2.5-7B-Instruct using the public NLA checkpoints from Fraser-Taliente (2026). Table 9 lists each NLA layer and residual width. The released configuration records the exact model and checkpoint revisions.

Evaluation data. The evaluation contains 200 WildChat prompts and 300 LegalBench prompts. From WildChat (Zhao et al., 2024), we take the first 200 English single-turn prompts in dataset order whose first user message is non-empty, no longer than 3000 characters, and not marked redacted or toxic. From the LegalBench test split (Guha et al., 2023), we sample `hearsay`, `personal_jurisdiction`, `proa`, `sara_entailment`, and `successor_liability`. We initially take up to 60 examples from each task, then redistribute unused slots across the remaining tasks. The final task counts are $68/50/68/67/47$. We prefix each LegalBench prompt with "Answer the legal question. Give brief reasoning, then state the final answer clearly."

All prompts follow the system message "You are a helpful assistant. Give brief reasoning before the final answer." We generate target responses greedily in `bfloat16`, with a limit of 512 new tokens.

Position selection. We select up to eight positions per target response. We exclude the first 10 whitespace-delimited tokens because their NLA decodes are unstable, then check the first occurrence of each legal or reasoning term below:

```
therefore, because, conclusion, liable, liability, contract,
statute, rule, court, claim, duty, breach, permission,
prohibited, required, final answer, answer.
```

For each term, we choose the last eligible token start at or before its occurrence. If none exists, we choose the first eligible token. We sample any remaining positions without replacement using seed $17 + i$ for source example i. For responses shorter than 10 tokens, we use the available positions. Each pair has 4000 positions, except the two runs with Gemma-3-12B as target; one response has only seven eligible positions, giving 3999 in total.

Activation extraction and explanation. We replay the complete target response through both models and extract `bfloat16` residual activations at each model's NLA layer. Each replay uses that model's chat template and tokenizer. We map response character offsets to assistant-token indices separately for each model, avoiding any assumption of one-to-one token alignment between Gemma and Qwen. We exclude prompt and template tokens.

We decode both target and surrogate explanations with the official NLA inference implementation using temperature 0.7 and at most 200 new tokens. The target AR reconstructs an activation from each explanation. We normalize predictions and target activations according to the released NLA configuration. The main comparison reports cosine similarity; the baseline tests use the equivalent normalized MSE, $\mathrm{MSE} = 2(1 - \cos)$.

Controls. The target AR scores each control against the same target activation:

- **Target explanation:** the target NLA's explanation, used as an upper reference rather than a perfect ceiling.
- **Shuffled explanation:** a random permutation of the surrogate explanations across all scored positions, seeded with 17. We redraw any permutation with a self-pair. Each explanation is therefore used once, and no position receives its own surrogate explanation. This control tests whether unrelated NLA text can reconstruct the target activation.
- **Generic text:** the fixed sentence "This text discusses a legal or conversational question and gives a brief answer."
- **Visible transcript:** the final 1000 characters of the user message and of the assistant text before the selected position. It contains visible context but no activation-derived explanation.

Concept agreement. A blinded Kimi-K2.6 judge provides a separate measure of semantic agreement. For each model pair, we sample 100 LegalBench positions from each task using seed 17, for 500 positions in total. The judge evaluates each explanation pair twice, reversing the order on the second pass. A third blinded judgment resolves disagreements. Labels are `same`, `partial`, `different`, or `unclear`.

Statistical analysis. The source example is the statistical unit for comparisons with the visible transcript, shuffled explanation, and generic text. We first average position-level MSE within each example, then run two-sided paired t-tests on $\mathrm{MSE}_{\mathrm{control}} - \mathrm{MSE}_{\mathrm{surrogate}}$; positive values favor the surrogate. Holm correction covers all 18 combinations of six model pairs and three controls. For concept agreement, a two-sided exact binomial test compares `same`+`partial` with `different` after excluding `unclear`.

Compute and reproducibility. We ran the experiments on one host with two NVIDIA A6000 GPUs. Transformers handled activation replay, vLLM generated target responses, and SGLang decoded the AV explanations. The artifact includes the software environment, resolved model revisions, prompts, pairwise configurations, and analysis scripts.

C.2 Significance relative to reconstruction controls

Table 10 reports the mean MSE reduction of the surrogate explanation relative to each control. Every test includes all 500 source examples, including the two runs with 3999 positions. All 18 differences are positive, with Holm-adjusted $p_{\mathrm{Holm}} < 10^{-300}$.

The surrogate has lower reconstruction error than every control for all six model pairs. The gap is largest for a Qwen target with a Gemma surrogate. This matches the asymmetry in Table 2; reconstruction scores should still be read alongside semantic agreement.

C.3 Concept-agreement significance

Table 11 reports the judge-label counts and exact-binomial tests. All six model pairs have more `same`+`partial` than `different` judgments. Agreement is highest within the Gemma family, at 79.5–81.8%, and ranges from 54.6% to 58.6% for the four cross-family pairs. Every comparison

Table 10: Mean paired reduction in target-activation reconstruction MSE, $\mathrm{MSE}_{\mathrm{control}} - \mathrm{MSE}_{\mathrm{surrogate}}$. Positive values favor the surrogate. Each test uses source examples as the statistical unit; all Holm-adjusted p-values are below 10^{-300}.

Target → Surrogate	Transcript	Shuffled	Generic
Within family			
Gemma-27B → Gemma-12B	0.0444	0.0693	0.0943
Gemma-12B → Gemma-27B	0.0252	0.0386	0.0385
Cross family			
Qwen-7B → Gemma-12B	0.8342	0.9701	0.9169
Qwen-7B → Gemma-27B	0.8555	0.9865	0.9382
Gemma-12B → Qwen-7B	0.0198	0.0310	0.0331
Gemma-27B → Qwen-7B	0.0363	0.0565	0.0862

Table 11: Concept agreement between target and surrogate explanations. The agreement rate is $(\texttt{same} + \texttt{partial})/N$. We exclude `unclear` judgments from N and test against a null agreement probability of 0.5 using a two-sided exact-binomial test. The final column gives Holm-adjusted p-values across the six model pairs.

Target → Surrogate	N	Same+partial	Different	Rate	p_{Holm}
Within family					
Gemma-27B → Gemma-12B	498	396	102	79.5%	3.29×10^{-41}
Gemma-12B → Gemma-27B	500	409	91	81.8%	2.26×10^{-48}
Cross family					
Qwen-7B → Gemma-12B	500	286	214	57.2%	0.00442
Qwen-7B → Gemma-27B	500	273	227	54.6%	0.0441
Gemma-12B → Qwen-7B	500	293	207	58.6%	0.000556
Gemma-27B → Qwen-7B	499	283	216	56.7%	0.00618

remains significant after Holm correction across the six tests. The weakest result is Qwen-7B → Gemma-27B ($p_{\mathrm{Holm}} = 0.0441$).

The binomial test measures whether agreement labels outnumber disagreement labels under this judge and labeling scheme. It does not measure the quality of the explanations or the reliability of the judge itself. We therefore report the agreement rates and label counts with the p-values.

D Process Reward Model Training and Evaluation Details

D.1 Explanation-utility classification

Data and inputs. Section 3.2 uses 280 normalized expert spans with at least one of the five error labels, after merging identical spans and excluding conflicting sound/error annotations. They form 14 task/document groups, with trajectories sharing the same task and input documents kept together. This experimental collection includes pilot annotations and differs from the 354-span descriptive study. The 233-span supervision comparison in Section 3.3 uses a separate collection of ten trajectories, excluding its pilot/development partition.

Inputs contain the full task, the selected span, up to 1200 preceding and 400 following characters, and up to three input-document excerpts of 1200 characters each, retrieved by lexical relevance to the task and span. Retrieval does not use annotation labels or feedback. Claims are aligned by trajectory, surface, and character overlap, then ordered by cited-token overlap. They cover 96 spans; the remaining 184 receive identical unaugmented inputs across conditions.

Explanation conditions. Condition B uses Qwen3-4B-Instruct-2507 with its user-message chat template, greedy decoding, and at most 384 new tokens. The following instruction precedes the observable inputs:

> Describe what reasoning the candidate passage performs and how it relates to the task and supplied evidence. Use only the observable text below. Be specific about factual references and reasoning transitions; distinguish what is stated from what is inferred. Write approximately 180–220 words of explanation. Output only the explanation, without an introduction or error-category labels.

Condition C uses the surrogate claims described in Appendix B.1. Condition E shuffles claims within the same trajectory, excluding overlapping explanation spans and identical claim texts. For each covered span, B–E use a common explanation field and token budget: the minimum of 256 and the available lengths in B, C, and E. Explanations are truncated to this budget using the classifier tokenizer. Generation and budget selection do not use expert error labels or feedback.

Condition D prioritizes aligned claims using a logistic helpfulness predictor with 512-dimensional normalized signed hashed word features. Targets are the fraction of helpful ratings per claim. Training weights documents equally and uses 200 gradient steps, learning rate 2, and L_2 coefficient 0.001, with an unregularized intercept. Each selector excludes both the outer test group and the input's own group, and selects ranked claims within the same token budget.

Classifier training. We freeze Qwen3-4B-Instruct-2507 and extract its 2560-dimensional final hidden state at the last token of the suffix `<assessment>`. Classifier inputs use a plain-text template, without a chat wrapper, and a 4096-token limit. The backbone runs in BF16; features and heads use FP32. Features are standardized using training-fold means and standard deviations, with a minimum scale of 0.1. A linear head with five sigmoid outputs is trained for 200 full-batch AdamW steps (learning rate 0.003; weight decay 0.01). Weights are initialized from $\mathcal{N}(0, 0.02^2)$ and biases to zero. Binary cross-entropy averages equally over labels and document groups, without class reweighting. Each of 14 groups is held out once; five conditions and seeds 0, 1, and 2 yield 210 heads. All runs use the final head and a fixed prediction threshold of 0.5.

Evaluation. Macro-F1 pools held-out predictions within each seed, averages F1 over the five labels, then averages seeds. Document BCE averages within each document and then equally over documents and seeds. Paired 95% percentile intervals use 10,000 document-bootstrap resamples with seed 0 and the same sampled groups across conditions. For F1, each draw pools confusion counts within each seed before averaging label F1 and seeds. For BCE, seeds are averaged within each document before resampling. Resampling uses saved predictions without retraining.