BEYOND FEATURE NAMES: OPERATIONAL INTERPRETABILITY FOR TEXT REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpretable text representations should be both predictive and usable as named constructs that independent analysts can apply. Existing approaches offer different trade-offs between these goals: anonymous bases can predict well, while topic models and concept-bottleneck pipelines attach names to features, but naming alone does not establish reproducible application. We propose **LFD** (LLM-assisted Feature Discovery), which targets *operational feature interpretability* through two requirements: *conceptual clarity*, enabling intended users to apply stated definitions and reproduce model-used feature values, and *label disentanglement*, requiring the feature's meaning to be distinct from the target label. LFD discovers candidate features from contrastive, outcome-opposed text groups and combines cross-LLM examination with selection for incremental held-out predictive accuracy. Under explicit assumptions, we connect cross-rater κ to a bound on the measurement-noise contribution to final prediction risk. Across diverse text-classification tasks, LFD yields compact, named feature bases with average predictive accuracy close to TBM's. Across all human-audited datasets, LFD shows higher average inter-rater agreement and stronger correspondence between human-majority judgments and pipeline reference labels than TBM. Definition-level ratings also favor LFD under the collected disentanglement rubric. These results make the case for evaluating text representations not only by what they predict, but also by whether their named coordinates function as reproducible measurements with meaning distinct from the target.

1 INTRODUCTION

What does it mean for a learned representation of text to be *interpretable*? Existing coordinate-level methods often face a trade-off. Anonymous bases such as embeddings and their reduced-dimensional variants (e.g., via PCA) can predict well, but their coordinates are not named constructs an analyst can read or apply. Conversely, concept-bottleneck and LLM-assisted pipelines attach natural-language names to features, but names alone do not establish reproducible application: the resulting definitions may sound plausible while remaining too abstract, ambiguous, or label-entangled to support auditing or intervention (Koh et al., 2020; Ludan et al., 2024; Belinkov, 2022). A feature name is not enough. To audit a model's reliance on a coordinate, or to intervene on that coordinate, an analyst must know what the feature reliably picks out in new texts.

Social science offers a stricter standard. When researchers code documents for latent constructs such as *aggressive language*, *gaslighting*, *deceptive review*, or *implicit hate*, they do not treat a construct as usable merely because its description is intuitive. Researchers need to turn definitions into measurements that others can reproduce on new documents, compare across samples, and use to test substantive hypotheses. Understanding what a construct means is therefore not enough; its definition must support consistent use as a research variable. They write an operational codebook, ask independent coders to apply it, and report inter-rater agreement—commonly Cohen's κ (Cohen, 1960) or Krippendorff's α (Krippendorff, 2018)—as evidence that the construct is reproducibly measurable. We import this standard into representation learning: a coordinate's natural-language definition should support reproducible application across independent annotators. The criterion makes ambiguity testable: "cites a specific dollar amount" may be easier to apply consistently than

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.

"expresses subtle irony," and an observed gap can reveal interpretive latitude left by the definition. We call this requirement **conceptual clarity**: independent users should be able to apply a feature's definition consistently and reproduce its model-used values. This consistency can also improve predictive performance by reducing measurement noise in the classifier's inputs, a connection we formalize in Section 4.

Meanwhile, conceptual clarity alone is not sufficient for a useful explanation. A feature may overlap substantially in meaning with the prediction target. In our experiments, for example, TBM proposes "mock sincerity" for sarcasm detection and "emotional authenticity" for deceptive-review detection. Such features can shift part of the target judgment into the representation rather than identify distinct evidence that explains it. This problem is especially consequential in computational social science, psychology, and political science, where features are extracted not only to predict labels but to understand the constructs associated with them. For such uses, a feature must provide distinct explanatory content rather than reproduce a judgment of the dependent variable. We therefore seek a second property: **label disentanglement**, meaning that a feature's definition describes a property conceptually distinct from the target rather than restating it or asking for substantially the same judgment.

We thus propose **operational feature interpretability**, hereafter operational interpretability, as a standard with two desiderata: conceptual clarity and label disentanglement. Predictive utility remains a separate requirement for a useful representation. We instantiate this standard in **LFD** (LLM-assisted Feature Discovery), which produces explicit definitions and application procedures for use across downstream prediction models. Contrastive, outcome-opposed text groups ground proposals in concrete properties; residual discovery seeks complementary cues missing from the current representation. A separate examiner tests whether each LLM-measured definition transfers to another measurer before incremental predictive gain is considered. This addresses a *closed-loop failure mode*: when a pipeline generates and applies a concept without an external application test, predictive success can reflect an interpretation that its definition does not communicate to another reader. The examiner tests that transfer directly. Grounded proposals and incremental selection encourage a decomposition into distinct evidence rather than repeated descriptions of the target (§3.2). We evaluate the label disentanglement of the discovered features (§5.2).

Contribution. We propose operational interpretability as a standard combining conceptual clarity and label disentanglement, and introduce LFD to instantiate it. Our theory links feature measurement agreement to a downstream prediction-risk bound. Across ten classification tasks, LFD achieves average predictive accuracy close to TBM's. Human audits favor LFD over TBM in average application agreement and human–pipeline correspondence on three datasets, and in definition-level disentanglement ratings on all ten tasks.

2 Related work

The discriminative-vs-interpretable axis. Text representations range from sparse lexical models and dictionaries to topic models and dense embeddings (Tibshirani, 1996; Loughran & McDonald, 2011; Blei et al., 2003; Grootendorst, 2022; Devlin et al., 2019). Table 1 organizes these and concept-based approaches by target-informed coordinate construction and named versus anonymous coordinates. Supplied names alone do not establish operational interpretability.

Table 1: The 2×2 organization of representation approaches (schematic, typical configurations). Rows distinguish target-informed definition, selection, or learning of coordinates from task-agnostic construction, not whether the resulting features can predict. Named coordinates do not by themselves establish interpretability.

	Anonymous coordinates	**Named coordinates**
Discriminative	Outcome-weighted PCA, fine-tuned BERT	**LFD (this paper)**, TBM, CBM, LaBo, FELIX, CB-LLM
Non-discriminative	Fixed pretrained embedding, vanilla PCA	LDA, BERTopic, keyword dictionaries

Interpretability and human-grounded evaluation. Human-grounded evaluation distinguishes subjective judgments of explanation quality from task-based tests such as forward simulation (Doshi-

Velez & Kim, 2017). Our conceptual-clarity audit sets a more demanding behavioral standard than self-reported understanding: independent users must apply feature definitions to new texts and assign feature values. We assess both agreement between users and correspondence with pipeline reference labels, directly testing whether the definitions support reproducible measurement rather than merely sound understandable.

Named concepts and predictive representations. Concept bottleneck models predict through a layer of named concepts. The original framework uses human-provided concept supervision (Koh et al., 2020); it does not rely on LLM-generated concepts. Later approaches generate concepts with language models, measure them using pretrained representations, or attach concept layers to existing predictors (Yang et al., 2023; Oikarinen et al., 2023; Yuksekgonul et al., 2023). CB-LLMs learn concept activations through a task-trained backbone and bottleneck for text classification, and extend concept bottlenecks to generation (Sun et al., 2025). LFD is closer to text feature-discovery approaches that apply feature descriptions to obtain values: TBM iteratively proposes and measures concepts using an LLM (Ludan et al., 2024), while FELIX proposes features from cross-class text pairs, consolidates them by semantic clustering, and assigns values for downstream classifiers (Malberg et al., 2024). Within this family, LFD combines residual-guided discovery and incremental predictive selection with independent examination of feature measurement.

Concept leakage and imperfect correspondence between names and measured coordinates are established concerns (Mahinpei et al., 2021; Margeloiu et al., 2021; Ramaswamy et al., 2023). Our semantic-distinction criterion addresses definitions that restate or substantially overlap in meaning with the outcome. It is not equivalent to every form of information leakage in concept bottlenecks. Cross-LLM agreement can expose disagreement in application, while definition-level assessment addresses semantic overlap.

LLMs as annotators and hypothesis generators. LLMs can code text and propose hypotheses (Gilardi et al., 2023; Törnberg, 2023; Zhou et al., 2024). We separate proposal from application to subject definitions to a transfer test. The examiner is independent in access to the other rater's answers; different model providers do not imply statistically independent errors. Prior work connects inter-rater statistics to learning under noisy target labels (Bucarelli et al., 2023). We distinguish that setting from corruption of multiple input features and avoid transferring label-noise guarantees without their assumptions and training procedure (Natarajan et al., 2013; Patrini et al., 2017).

3 Operational criteria and adaptive feature discovery

3.1 Operational feature interpretability

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ contain texts and binary outcomes. A named feature f has a verbal definition ρ_f and an application procedure, implemented by an executable rule or an annotator. Application of the verbal definition produces $\phi_f^A(x) = A(x; \rho_f) \in \{0, 1\}$ for annotator A. Writing the annotator explicitly avoids assuming that a description defines the same labeling function for every reader.

Definition 1 (Operational feature interpretability)**.** *Relative to a prediction task, specified intended users, application instructions, and an input distribution, operational feature interpretability requires two properties:*

(i) ***Conceptual clarity.*** *Users can consistently apply the feature's stated definition to new inputs and reproduce the values used by its predictor, without consulting the original annotator's judgments.*

(ii) ***Label disentanglement.*** *The definition describes a text property whose meaning is distinct from the prediction target, rather than restating it or asking for substantially the same judgment.*

Let $v_f(x)$ denote the implemented value used by the predictor, obtained by executing the rule or applying the designated labeler. We assess conceptual clarity through two complementary measurements: *reproducibility*, or agreement between users' applications $\phi_f^A(x)$ and $\phi_f^B(x)$ of ρ_f, measured by $\kappa_f^{A,B}$ on the audit distribution; and *correspondence*, or agreement between users' applications and $v_f(x)$. Individual-user and human-majority comparisons answer different questions and are identified separately.

We evaluate the two interpretability properties separately, with *predictive utility* as a further objective rather than part of Definition 1. Agreement is evidence of reproducible measurement, not by itself of construct validity; users and the implementation may share the same mistaken interpretation.

3.2 LFD: FROM INITIAL CONTRASTS TO RESIDUAL DISCOVERY

LFD builds a feature basis by progressively changing the contrast posed to the proposer. The *bootstrap* phase asks which concrete text properties distinguish the outcome classes; the *residual* phase asks what additional distinctions the current representation is missing. Predictive feedback therefore guides both *where to look* for new features and *which features to retain*. Discovery and classifier fitting use the training set, incremental predictive value is assessed on the validation set, and the test set is reserved for final evaluation.

Bootstrap: turn outcome contrasts into an initial feature basis. Starting with no accepted features, we sample small groups of training texts from each outcome class and present them side by side to a proposer LLM $\mathcal{P}$. The contrast supplies examples of both what a candidate should detect and what it should distinguish from. Rather than proposing concepts from the task name alone, the proposer must articulate observable differences in these texts as Boolean feature definitions.

Candidates may capture lexical patterns, such as *"contains three consecutive exclamation marks,"* or semantic properties, such as *"invokes a statute without giving a personal narrative."* Each definition specifies how to decide feature presence in a new text, without knowing its outcome. The prompt asks for explicit inclusion rules and cues covering distinct positive-class subgroups: a feature need not describe the entire class to be useful. Candidates are measured and examined as described below, then selected for incremental held-out gain. The accepted features provide the initial representation and predictor from which residual discovery can proceed.

Residual discovery: use the current representation to direct the next contrast. An initial pool can capture salient signals yet miss other useful cues. Once it ceases to yield sufficient predictive gain, simply repeating the same broad class comparison risks rediscovering what is already captured. The residual phase instead refreshes proposals using remaining or harder training examples, as identified from the current predictor's errors. The proposer again sees outcome-opposed texts, together with existing feature definitions, and is asked for complementary cues rather than rewordings of the current basis. For illustration, if a statute-reference feature already captures one subgroup of complaints, the next proposal should seek evidence that helps distinguish the cases this feature leaves unresolved, rather than merely rename legal language.

The change in proposal context is coupled to a conditional selection criterion. A candidate's residual gain Δ_f is the increase in balanced accuracy on the validation set after adding it to the current feature set and refitting the classifier on the training set using the same fitting procedure. A cue that is predictive by itself may add little once its information is already represented; conversely, a cue useful for a previously missed subgroup can improve the combined predictor. Thus, *hard examples guide proposal, while incremental held-out gain decides admission*. Here gain means additional predictive value, not correlation with the internal pseudo-residuals of LightGBM (Ke et al., 2017). After an addition, the predictor is updated; subsequent proposal refreshes use the expanded basis and its remaining errors.

This feedback loop also explains why LFD encourages label disentanglement: text contrasts and subgroup-specific prompts favor concrete cues, while incremental selection favors complementary evidence over repeated versions of an already captured signal. These are incentives for decomposing the task, not a guarantee of semantic distinction; an initial label paraphrase could still be selected. We therefore evaluate disentanglement as a property of the discovered representation (§5.2).

A shared requirement: definitions must transfer to another measurer. Both phases use the same feature-application and examination procedure. Executable rules are applied directly with fixed preprocessing. Otherwise, a labeler $\mathcal{L}$ produces the feature values used by the predictor, and an examiner $\mathcal{E}$ independently applies the same definition. Neither sees the document's outcome or the other's answers. The intended screen is

$$\widehat{\kappa}_f^{\mathcal{L},\mathcal{E}} \geq \kappa^*, \tag{1}$$

using Cohen's κ (Cohen, 1960). We choose $\kappa^* = 0.70$, within the conventional substantial-agreement range (Landis & Koch, 1977). This tests whether the definition transfers to another LLM before it

competes on predictive gain. Cross-LLM agreement is only a proxy for human conceptual clarity; executable-rule determinism likewise does not establish that people can recover the rule's output from its verbal description. Human application audits evaluate that correspondence separately.

A worked CFPB example. For illustration, suppose the contrasted complaint groups include a routine dispute and a complaint stating *"Per FCRA §1681i, see attached supporting documentation."* The proposer might suggest *"invokes a statute or supporting documentation to justify the complaint."* The labeler and examiner independently apply this definition to the same training texts. If their agreement clears the screen, the selector asks whether adding the feature improves validation prediction beyond the current basis. A candidate can pass examination yet fail to improve prediction; predictive gain alone cannot compensate for failing the agreement screen. The next residual proposal then seeks evidence missing from the accepted basis, rather than another name for the same cue.

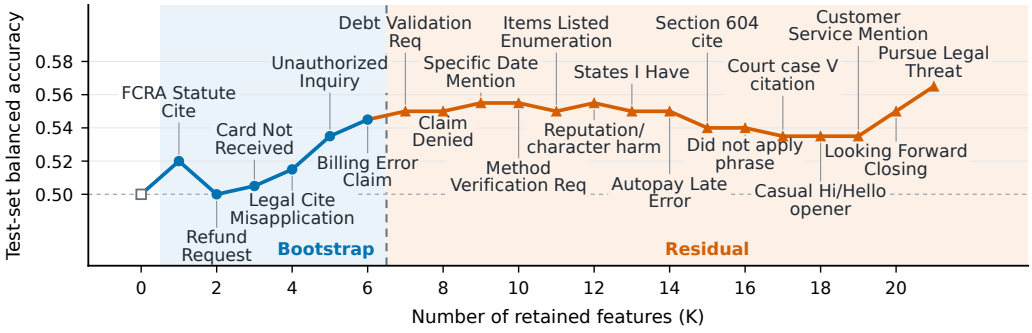

Figure 1: Building the CFPB representation one feature at a time. Each point refits LightGBM on the combined training and validation sets using the first k admitted features and reports test-set balanced accuracy. At $k = 0$, the predictor is the majority-class baseline. Blue shading and circles mark bootstrap additions; orange shading and triangles mark residual additions, beginning at $k = 7$. The final 21-feature basis reaches 0.565. Labels identify the feature added at each point; explanations are in Appendix I.

From individual cues to a predictive representation. Figure 1 shows how this process builds the CFPB representation. Early additions include *FCRA Statute Cite* ($k = 1$) and *Billing Error Claim* ($k = 6$). Residual discovery begins with *Debt Validation Req* ($k = 7$) and later adds cues such as *Pursue Legal Threat* ($k = 21$). The named coordinates make the representation inspectable: an analyst can see which distinct aspects of the complaint enter as the basis expands. Each point evaluates the joint predictor using all features admitted so far, rather than the newly added feature in isolation. Test accuracy can decline at an intermediate step because admission is based on validation gain, not test-set improvement.

Algorithm 1 gives the training shortlist, pool refresh, and stopping procedure, with implementation details in Appendix B. After discovery, feature definitions and measurement procedures are fixed. The resulting feature matrix can be reused across downstream predictors without repeating feature discovery. For the reported evaluation, the final predictor is refitted on the combined training and validation sets and evaluated on the test set. The head may be nonlinear because our interpretability requirements concern its input features; understanding the complete decision rule is a separate property.

4 Theory: Conceptual Clarity and Generalization

Requiring conceptual clarity may appear to sacrifice predictive fit, but fit based on an annotator's idiosyncratic reading need not transfer to new texts. We connect agreement to downstream risk in two steps: agreement bounds feature measurement noise, which contributes a prediction penalty beyond finite-sample estimation error.

Assumption 1 (Symmetric independent rater errors)**.** *For one fixed feature, a latent binary labeling Z has prevalence $\pi \in (0, 1)$. Two audited annotators return $A = Z \oplus \epsilon_A$ and $B = Z \oplus \epsilon_B$. For $a \in \{A, B\}$, $\mathbb{P}(\epsilon_a = 1 \mid Z = z) = \eta_a < 1/2$ for both z, and the two errors are conditionally independent given Z.*

For each LLM-measured coordinate k, apply this assumption to the deployed labeler and examiner. A population agreement lower bound $\kappa_k \geq q_k \geq 0$ implies labeler error $\eta_k \leq b_k = (1 - q_k)/2$; with equal rater error rates, $b_k = (1 - \sqrt{q_k})/2$ (Lemma 1, Appendix A). This intermediate result supplies the noise term in the downstream bound.

For a feature basis $\mathcal{F}$, let $Z_{\mathcal{F}}$ be its latent reference values and $A_{\mathcal{F}}$ its measured values. For directly executed coordinates the reference is the rule itself, so the two values coincide; this does not certify fidelity to a verbal description. Write $R^0_{\mathcal{F}}(h)$ and $R^L_{\mathcal{F}}(h)$ for the population classification errors of head h on reference and measured features, respectively. Let $\epsilon_N(\delta)$ be a uniform generalization bound, valid with probability at least $1 - \delta$, over a prespecified family of candidate bases and heads on N iid training observations, with measurement procedures fixed. Its complexity term covers feature discovery as well as head fitting, allowing a basis to be selected using these observations (Mohri et al., 2018).

Theorem 1 (Conceptual clarity and downstream prediction risk)**.** *Let $\widehat{\mathcal{F}}$ be the selected basis and $\hat{h}$ a fitted head whose empirical classification error is within α of the minimum over its head class. Let $h^\star$ minimize reference-feature risk within that same class and basis. Under Assumption 1 for each LLM-measured coordinate and the population agreement bounds above, with probability at least $1 - \delta$,*

$$\begin{aligned} R^L_{\widehat{\mathcal{F}}}(\hat{h}) - R^0_{\widehat{\mathcal{F}}}(h^\star) &\leq 2\epsilon_N(\delta) + \alpha + D_{\widehat{\mathcal{F}}}(h^\star) \\ &\leq 2\epsilon_N(\delta) + \alpha + \min\left\{1, \sum_{k \in J(h^\star)} b_k\right\}, \end{aligned} \tag{2}$$

where $D_{\mathcal{F}}(h) = \mathbb{P}\{h(A_{\mathcal{F}}) \neq h(Z_{\mathcal{F}})\}$ is the probability that measurement changes the prediction, and $J(h)$ contains the LLM-measured coordinates on which h depends.

The proof is in Appendix A. The theorem separates estimation error, optimization error, and the prediction penalty from feature measurement. Unlike estimation error, the last term need not disappear with more training data. Higher population agreement tightens its upper bound, holding the representation and head fixed. Predictive selection remains necessary: screening can also remove useful information and change the reference risk. The comparator is the best head on the selected reference features, not an oracle with access to all information in the text.

Population agreement bounds are distinct from the empirical admission threshold. Appendix A details estimation, shared-error limitations, adaptive discovery, and a margin-based refinement for linear heads.

5 Experiments

We evaluate three questions: (1) how much predictive signal can LFD express through a compact, named feature basis compared with TBM and anonymous embedding representations (§5.1); (2) whether its discovered features describe properties distinct from the prediction target (§5.2); and (3) whether independent raters can apply the definitions consistently and reproduce pipeline feature values (§5.3). We also examine reuse of a frozen basis across downstream classifiers and sensitivity to the LLMs and selection thresholds (§5.4, Appendix H).

Datasets. We evaluate on ten binary tasks across seven corpora: five domain and social-science corpora—AdParaphrase (Murakami et al., 2025), Loan Application, CFPB Complaints (Consumer Financial Protection Bureau, 2025), Deceptive Reviews (Ott et al., 2011; 2013), and Gaslighting—and two standard NLP benchmarks, iSarcasm (Abu Farha et al., 2022) and AG News (Zhang et al., 2015), the latter as four one-versus-rest tasks. The set spans topical cues and context-dependent judgments such as sarcasm and review authenticity, across English, Chinese, and Japanese. All tasks use balanced 1,000-text subsets with stratified 60/20/20 training/validation/test splits. LFD uses the training set for proposal and classifier fitting and the validation set for incremental-gain selection. After discovery, the predictor is refitted on the combined training and validation sets and evaluated on the test set. Appendix C gives domains, languages, outcomes, and configuration details.

Baselines. Following the named-versus-anonymous distinction in Table 1, we compare three groups. *Anonymous coordinates:* **Full Embedding** uses LightGBM on the full 3,072-dimensional

Table 2: Test balanced accuracy in the initial native-head comparison. TBM uses a logistic head and LFD uses LightGBM. K is LFD's width; LDA, BERTopic, TF-IDF, and PCA use the same K. TBM uses its own selected width, shown in parentheses beside its accuracy. Light-gray cells mark baselines with lower balanced accuracy than LFD in the same row.

		Int. non-disc.		*Interpretable disc.*			*Anonymous disc.*	
Dataset	K	LDA	BERTopic	TF-IDF	TBM	**LFD**	PCA	Full Emb.
AdParaphrase	22	0.510	0.515	0.485	0.540 (9)	0.665	0.535	0.630
Loan Application	3	0.745	0.615	0.695	0.700 (6)	0.775	0.670	0.790
AG News S/T	26	0.515	0.775	0.705	0.835 (9)	0.855	0.885	0.895
Gaslighting	17	0.555	0.550	0.550	0.530 (6)	0.605	0.610	0.630
AG News Sports	33	0.630	0.865	0.820	0.980 (2)	0.885	0.985	0.980
AG News World	23	0.585	0.770	0.740	0.830 (4)	0.810	0.905	0.905
AG News Business	20	0.615	0.745	0.680	0.795 (6)	0.750	0.855	0.855
CFPB	21	0.595	0.560	0.595	0.550 (5)	0.565	0.590	0.600
iSarcasm	15	0.520	0.545	0.570	0.705 (6)	0.575	0.635	0.680
Deceptive Reviews	23	0.530	0.550	0.740	0.605 (9)	0.600	0.855	0.860

`text-embedding-3-large` representation; **PCA** projects that embedding to a smaller basis. *Named, task-agnostic coordinates:* **LDA** (Blei et al., 2003) and **BERTopic** (Grootendorst, 2022) provide classical and neural topic representations. *Named, task-informed coordinates:* **TF-IDF** uses chi-squared-selected tokens, while **TBM** (Ludan et al., 2024) iteratively proposes and measures concepts with an LLM and predicts through a logistic head. TBM is the closest main-table comparator to LFD's explicit feature discovery. PCA, TF-IDF, LDA, and BERTopic use LFD's task-specific width K; TBM retains its own selected width. This matches output dimensionality, not inference cost or description complexity. LFD uses LightGBM; the native-head comparison evaluates complete pipelines.

Evaluation metrics. Predictive performance is reported as **balanced accuracy** (BA). Interpretability follows the two requirements of Definition 1. **Label disentanglement** is assessed through definition similarity, feature–label association, and five-point human ratings (§5.2). **Conceptual clarity** is assessed by Cohen's κ for human–human reproducibility and human-majority–pipeline correspondence, complemented by a held-out cross-LLM audit (§5.3). These application tests are distinct from judgments of the descriptions alone.

5.1 Predictive performance

The results in Table 2 distinguish raw predictive strength from the stricter requirement that predictive signal be expressed through named, independently auditable coordinates. Anonymous representations can exploit embedding directions without having to explain what each coordinate measures, and often give the strongest raw prediction. LFD instead asks for a compact set of explicit features whose definitions can be examined through independent application. The relevant question is how much predictive signal can be expressed through coordinates that are both useful and auditable.

In the native-head comparison (Table 2), LFD and TBM have similar average predictive accuracy, with each higher on five tasks (Wilcoxon $p = 1.00$). This does not establish equivalence or isolate feature quality from the choice of prediction head.

On AdParaphrase, LFD reaches BA 0.665 with 22 features, compared with 0.630 for the full embedding. On Loan Application, just three features reach 0.775 versus 0.790. These cases illustrate how much predictive signal can be expressed through a compact set of explicit cues. The contrast with Deceptive Reviews is substantial: its selected basis reaches 0.600 versus 0.860 for the embedding, indicating predictive information not captured by these features.

Where does TBM's predictive edge come from? Inspecting the admitted concepts makes the distinction between prediction and explanatory content concrete. On AG News Sports, TBM matches the full embedding's BA of 0.980 using just two concepts: *"mention of athlete or sports team"* and *"mention of sports terminology."* Their native scores have absolute held-out label correlations of 0.953 and 0.914, respectively, and their definitions explicitly connect these cues to sports versus non-sports

news. On iSarcasm, *"mock sincerity"* is defined through a tone conveying *"irony or sarcasm"*; AG News World includes *"Global political and social issues."* These task-proximal definitions can retain substantial classification judgment inside a named coordinate. The concern is what the definition asks a reader to decide, not high label correlation itself. LFD instead seeks more specific, complementary cues whose application can be examined separately. Accuracy, semantic distinction, and reproducible application therefore answer different questions about a representation.

Reusing the discovered representation. We reuse each frozen LFD basis across three downstream predictors. Mean BA is 0.717 for L1-logistic, 0.708 for a pruned decision tree, and 0.707 for LightGBM. The discovered representation therefore supports both linear and tree-based readouts. Appendix H reports these refits and a separate embedding-derived concept-bank comparison.

5.2 Label Disentanglement

We assess label disentanglement in the features produced by LFD through definition similarity, feature–label association, and human ratings. Neither similarity nor association alone establishes that a feature is a label paraphrase.

Automated evaluation. Figure 2 reports **(a) absolute Pearson correlation** between feature values and labels, measuring predictive association rather than semantic equivalence; and **(b) cosine similarity** between descriptions and label questions, embedded with `text-embedding-3-large`. Across ten tasks, mean cosine is 0.419 for TBM and 0.370 for LFD. Because LFD descriptions include regex syntax whereas TBM descriptions are prose, this comparison has a format confound; we report descriptive values without a pooled significance claim.

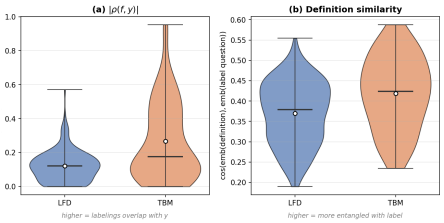

Figure 2: Two automated diagnostics for LFD features and TBM concepts, pooled across ten tasks: (a) absolute feature–label correlation on the test set; (b) description–question cosine. The latter has a format confound discussed in the text.

Human evaluation. Raters judged randomized LFD and TBM descriptions on a five-point rubric with task context. Across 34 retained sessions, LFD scores higher on all ten tasks (Figure 3), with the largest gaps on AG News World (+2.17) and iSarcasm (+1.93). These are descriptive results from small, uneven panels, using a rubric that asks whether a description is *independent of the answer*. Appendix F reports the instrument, filtering, and coverage.

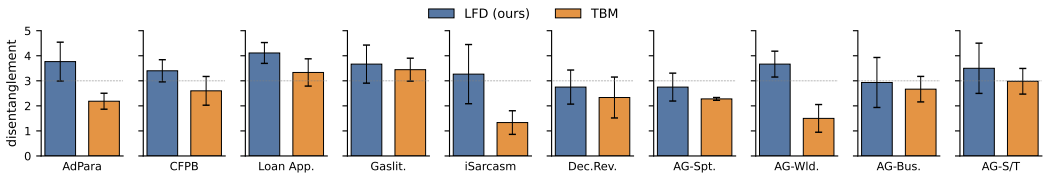

Figure 3: Human disentanglement rubric: mean rating per task. Higher is better under the collected rubric. The bar is the across-feature mean of per-feature means; the error bar is the across-feature SD, not a confidence interval. The 34 retained sessions have uneven task coverage (Appendix F).

5.3 Conceptual Clarity: Cohen's κ Across Rater Pairs

We assess conceptual clarity through independent application audits, complementing the cross-LLM admission screen (§3.2). Human–human (HH) agreement measures reproducibility; human-majority–pipeline (HM–P) agreement measures panel-level correspondence with model-used values.

Do human raters find the features conceptually clear? Human application requires substantially more annotation than rating feature descriptions: our protocol asks at least three independent raters to apply each definition to 25 texts. Given this cost, we focus the application audit on three contrasting

datasets—iSarcasm, AG News Sci/Tech, and CFPB—while the definition-level disentanglement audit covers all ten tasks. Both methods use the same agreement-independent quality checks; no exclusion depends on matching pipeline labels. Appendix E details the protocol, coverage, and uncertainty.

HH averages agreement across individual rater pairs. LFD has higher average HH agreement on all three tasks (Table 3), supporting more consistent human application. Agreement nevertheless varies across definitions and application procedures: executable features provide stronger human-reproducibility evidence on iSarcasm and Sci/Tech than their smaller LLM-measured subsets (Appendix E).

Table 3: Human application agreement, averaged over features with defined κ. HH averages human pairs; HM–P compares human majority with pipeline references (TBM scores thresholded at > 0). Undefined statistics are omitted without imputation.

Task	Method	HH κ	HM–P κ
iSarcasm	LFD	0.522	0.829
	TBM	0.269	0.261
AG News S/T	LFD	0.629	0.792
	TBM	0.361	0.365
CFPB	LFD	0.618	0.855
	TBM	0.178	0.168

Does LLM understanding align with humans? We assess alignment operationally through HM–P, comparing the human majority with the feature values used by the predictor. For LFD, these are binary outputs of executable rules or an LLM labeler; TBM scores are thresholded at > 0 to match the binary survey. LFD has higher average HM–P agreement on all three tasks (Table 3), supporting closer panel-level recovery of pipeline values. The TBM audit tests feature presence, not recovery of its full native scores. A majority can match the reference despite individual disagreement, so HH and HM–P provide complementary evidence.

5.4 Sensitivity and ablations

The CFPB threshold sweep gives non-monotone BA 0.510–0.565 for $\kappa^* = 0.50$–0.85 and does not establish 0.70 as optimal (Appendix G). Fixed-basis ablations and a separate post-discovery association-filter analysis complement this adaptive-discovery comparison (Appendix H).

Sensitivity to LLM choice. Three complementary studies distinguish proposal, measurement, and end-to-end sensitivity (Appendix H.4). With the labeler and examiner fixed, four proposers yield different feature bases and BA ranges of 0.090 on AdParaphrase and 0.040 on CFPB. To isolate measurement from discovery, six raters then apply 76 fixed definitions to the same 200 held-out documents per definition. Median agreement is similar for stronger same-family and cross-family pairs ($\kappa = 0.79$ versus 0.80), whereas GPT-3.5 Turbo has median $\kappa = 0.30$ against stronger peers. In this cohort, changing model family is therefore less consequential for agreement than substituting the weaker rater. Finally, nine full-discovery runs replace the proposer, labeler, or examiner one at a time with GPT-3.5 Turbo on three tasks. All nine yield smaller bases; BA decreases in eight and is unchanged in one. Proposer substitution lowers BA by 0.065–0.145; labeler substitution ranges from no change on CFPB to a 0.265 drop on Sci/Tech; examiner substitution lowers BA by 0.030–0.050. The roles matter differently: the proposer supplies candidate features, the labeler produces their values, and the examiner assesses agreement for admission. Screening cannot recover useful candidates that were never proposed, and reproducibility still depends on the raters. The tested substitutions identify model capability, and its role in the pipeline, as an important source of sensitivity.

6 Conclusion

We propose operational interpretability for text representations, requiring conceptual clarity and label disentanglement. LFD discovers compact, named features through contrastive proposals, cross-LLM examination, and incremental predictive selection. Our results show the potential to combine predictive utility with more consistent human application.

The broader goal is to make learned text features usable as measurements that independent users can apply, audit, and reuse in substantive research. Future work should test their transfer across domains and user populations. A feature's name is the starting point, not the explanation.

AI USE STATEMENT

Language models are components of proposal, measurement, and examination; their roles are specified in Appendix C.1. Generative AI also assisted with manuscript revision, consistency checking, proof checking, and implementation of the human-audit reanalysis. The authors reviewed all AI-assisted work and take responsibility for the final content.

ETHICS STATEMENT

Human raters only applied supplied feature definitions to texts. No personal information about the raters was collected, and the study involved no experimental intervention. Financial and social text may be sensitive; redistribution must respect source permissions. Named features can expose spurious associations and can also be exploited to game a classifier.

REPRODUCIBILITY STATEMENT

Appendix A provides proofs of the theoretical results. Appendices B, C.1, and E document selection, configuration, and human reanalysis. Code, prompts, feature definitions, split identifiers, and predictions will be released upon publication.

REFERENCES

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 802–814. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.semeval-1.111.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Maria Sofia Bucarelli, Lucas Cassano, Federico Siciliano, Amin Mantrach, and Fabrizio Silvestri. Leveraging inter-rater agreement for classification in the presence of noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3439–3448, 2023.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Consumer Financial Protection Bureau. Consumer complaint database. `https://www.consumerfinance.gov/data-research/consumer-complaints/`, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3146–3154, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, fourth edition, 2018.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, 2011.

Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. Interpretable-by-design text understanding with iteratively generated concept bottleneck. *arXiv preprint arXiv:2310.19660*, 2024.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

Simon Malberg, Edoardo Mosca, and Georg Groh. FELIX: Automatic and interpretable feature engineering using LLMs. In *Machine Learning and Knowledge Discovery in Databases*, pp. 230–246, 2024. doi: 10.1007/978-3-031-70359-1_14.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? In *ICLR 2021 Workshop on Responsible AI*, 2021.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.

Soichiro Murakami, Peinan Zhang, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. AdParaphrase: Paraphrase dataset for analyzing linguistic features toward generating attractive ad texts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1426–1439, 2025.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep K. Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, 2013.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *International Conference on Learning Representations*, 2023.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 309–319, 2011.

Myle Ott, Claire Cardie, and Jeffrey T. Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 497–501. Association for Computational Linguistics, 2013.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Chung-En Sun, Tuomas Oikarinen, Berk Ustun, and Tsui-Wei Weng. Concept bottleneck large language models. In *International Conference on Learning Representations*, 2025.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Petter Törnberg. ChatGPT-4 outperforms experts and crowd workers for annotating political Twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19187–19197, 2023.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pp. 649–657, 2015.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In Lotem Peled-Cohen, Nitay Calderon, Shir Lissak, and Roi Reichart (eds.), *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4science-1.10.

A AGREEMENT AND DOWNSTREAM PREDICTION RISK

A.1 FROM AGREEMENT TO FEATURE MEASUREMENT NOISE

Lemma 1 (Agreement and individual reliability)**.** *Under Assumption 1, let* $r_a = 1 - 2\eta_a$ *and* $m = 2\pi - 1$*. Population agreement satisfies*

$$\kappa^{A,B} = \frac{r_A r_B (1 - m^2)}{1 - r_A r_B m^2} \leq r_A r_B. \tag{3}$$

For $q \in [0, 1]$*, if* $\kappa^{A,B} \geq q$*, then* $\max(\eta_A, \eta_B) \leq (1 - q)/2$*. If additionally* $\eta_A = \eta_B = \eta$*, then* $\eta \leq (1 - \sqrt{q})/2$*.*

Proof. Under Assumption 1, transform Z, A, B to signs in $\{-1, 1\}$. Their means are respectively m, $r_A m$, and $r_B m$. Conditional independence gives $\mathbb{E}[AB] = r_A r_B$ in sign notation. Consequently, observed agreement p_o and chance-expected agreement p_e satisfy

$$p_o = \frac{1 + r_A r_B}{2}, \qquad p_e = \frac{1 + r_A r_B m^2}{2}.$$

Substitution in Cohen's definition of κ (Cohen, 1960) gives Equation 3. Since $0 < r_A r_B \leq 1$ and $m^2 < 1$,

$$\frac{1 - m^2}{1 - r_A r_B m^2} \leq 1.$$

Thus $q \leq r_A r_B \leq \min(r_A, r_B)$ implies $\eta_a = (1 - r_a)/2 \leq (1 - q)/2$ for each audited rater. With equal errors, $r_A r_B = (1 - 2\eta)^2$ and the sharper bound follows. □

At population agreement 0.70, the respective per-rater error bounds are 0.15 and approximately 0.082. Neither is a multiplicative generalization penalty.

A.2 PROOF OF THEOREM 1

Setup and the uniform event. Fix before observing the training data a family $\mathfrak{F}$ of candidate feature bases, with a head class $\mathcal{H}_\mathcal{F}$ for each basis. All measurement procedures are fixed functions of the text; if randomized, their randomness is included in the iid observation with the same law at training and deployment. These assumptions exclude an unmodeled change of labeler, instructions, or deployment distribution. The uniform event is

$$\sup_{\mathcal{F} \in \mathfrak{F},\ h \in \mathcal{H}_\mathcal{F}} \left| R^L_\mathcal{F}(h) - \widehat{R}^L_\mathcal{F}(h) \right| \leq \epsilon_N(\delta).$$

For example, if the family contains a total of M basis–head pairs, Hoeffding's inequality for each pair followed by a union bound gives $\epsilon_N(\delta) = \sqrt{\log(2M/\delta)/(2N)}$ with probability at least $1 - \delta$. For an infinite class, a suitable uniform bound for the full composed predictor class can be used instead (Mohri et al., 2018; Vapnik, 1998). The finite-family example is not a claim that the realized candidate pool can be counted after looking at the data.

The event holds simultaneously for every basis and head, hence also for the data-selected $\widehat{\mathcal{F}}$ and its reference-risk minimizer $h^\star$. No independence between the selected basis and the training sample is required. However, one cannot replace the complexity of the full class by that of the head on a basis chosen from the same data. Independent post-discovery training would permit such conditioning; the reported pipeline instead reuses the training and validation sets and therefore requires the joint-class treatment. We do not estimate this full search complexity or report a numerical risk certificate for the experiments.

Proof. Abbreviate the selected basis by $\mathcal{F}$. On the uniform event, ordinary empirical fitting with optimization gap α gives

$$\begin{aligned} R^L_\mathcal{F}(\hat{h}) &\leq \widehat{R}^L_\mathcal{F}(\hat{h}) + \epsilon_N \\ &\leq \widehat{R}^L_\mathcal{F}(h^\star) + \alpha + \epsilon_N \\ &\leq R^L_\mathcal{F}(h^\star) + \alpha + 2\epsilon_N. \end{aligned}$$

Couple measured and reference features on the same test text. For any fixed head, its zero–one losses agree whenever its predictions agree, so $|R^L_{\mathcal{F}}(h) - R^0_{\mathcal{F}}(h)| \le D_{\mathcal{F}}(h)$. Combining this with the preceding inequality proves the first bound.

Lexical coordinates coincide with their executable references. If none of the semantic coordinates on which $h^\star$ depends is mis-measured, its prediction cannot change. Thus

$$D_{\mathcal{F}}(h^\star) \le \min\left\{1, \sum_{k \in J(h^\star)} \eta_k\right\} \le \min\left\{1, \sum_{k \in J(h^\star)} b_k\right\},$$

where the last step uses Lemma 1. The union bound needs no independence across feature coordinates; conditional independence is required only between the two audited raters for each application of the lemma. □

Risk, optimization, and the comparator. The left side compares actual deployed prediction risk with the best reference-feature head on the same selected basis. It is not the training–test gap alone. Relative to the full-text Bayes risk, one must also retain the approximation error of that basis and head class. The gap α is empirical suboptimality under the classification loss in the theorem. Surrogate-loss optimization by LightGBM does not make this gap zero or establish that it is small. Under a balanced target distribution, classification error equals $1 - \mathrm{BA}$. At other prevalences, a balanced-risk version uses equal class weights throughout, including the agreement/noise assumptions and the corresponding concentration bound.

A.3 Classifier sensitivity and sharper noise penalties

The coordinate-sum bound is a worst-case fallback and can be vacuous for many noisy features. The first inequality in Theorem 1 instead retains the probability of a changed prediction. A useful refinement should account for the head rather than assume that every coordinate error changes the outcome.

For a linear threshold head $h(z) = \mathbf{1}\{w^\top z + a \ge 0\}$, define the reference margin $M(z) = |w^\top z + a|$ and let I_{sem} index the semantic coordinates. For any $t > 0$, a changed prediction requires either $M(Z_{\mathcal{F}}) \le t$ or a score perturbation larger than t. Since $|w^\top(A_{\mathcal{F}} - Z_{\mathcal{F}})| \le \sum_k |w_k| \mathbf{1}\{A_k \ne Z_k\}$, Markov's inequality and the coordinate error bounds give

$$D_{\mathcal{F}}(h) \le \min\left\{1, \inf_{t>0}\left[\mathbb{P}\{M(Z_{\mathcal{F}}) \le t\} + \frac{\sum_{k \in I_{\mathrm{sem}}} |w_k| b_k}{t}\right]\right\}. \tag{4}$$

No independence across coordinates is needed. This bound can replace the coordinate-sum bound when $h^\star$ is a linear threshold head: measurement noise is weighted by coefficient magnitude, while the margin distribution determines how readily score changes alter decisions. It is a conditional refinement for linear heads, not a margin guarantee for the nonlinear heads or a measured certificate for the reported runs.

For a general bounded loss, an analogous sensitivity-weighted risk bound uses coordinate bounded differences $c_k(h) = \sup_{v,y} |\ell(h(v), y) - \ell(h(v^{(k)}), y)|$, where $v^{(k)}$ flips coordinate k. Changing coordinates one at a time gives $|R^L_{\mathcal{F}}(h) - R^0_{\mathcal{F}}(h)| \le \sum_k c_k(h)\eta_k$. For zero–one loss the worst-case c_k are zero or one, so this alone does not remove dimension dependence. For an L-Lipschitz loss of a linear score, $c_k \le L|w_k|$ gives a weighted bound on that loss, not automatically on classification error.

A.4 Scope of the assumptions

Scope of the reliability interpretation. The bounds concern the audited pair; they do not bound an unaudited third rater's error or certify operational interpretability for human users. Two models may agree because they share an error, and model-family diversity does not establish conditional independence. The lemma therefore does not make $\kappa^* = 0.70$ a universal reliability threshold or establish that it controls downstream overfitting.

Population versus empirical agreement. The lemma does not turn an observed admission threshold into a population guarantee. Definitions are selected adaptively and the sampling variability of $\widehat{\kappa}$ depends on the joint label frequencies. For a frozen representation, independent audit should report per-feature prevalence, agreement, uncertainty, and the fraction of features for which agreement is estimable. A simultaneous guarantee over selected features needs an appropriate simultaneous uncertainty analysis.

Dependent and asymmetric errors. Without the assumptions, agreement does not identify latent correctness. Two raters that always share the same wrong interpretation can agree perfectly. As an illustration only, under balanced latent prevalence, equal error rates $0 \le \eta < 1/2$, and a mixture chosen independently of Z in which Bernoulli errors are shared with probability $0 \le \zeta < 1$ and independent otherwise,

$$\kappa = \zeta + (1-\zeta)(1-2\eta)^2.$$

For $\kappa > \zeta$, this implies

$$\eta = \frac{1 - \sqrt{(\kappa-\zeta)/(1-\zeta)}}{2}.$$

There is no universal multiplicative correction for error correlation without an error model. Weighted κ does not, by itself, identify asymmetric or instance-dependent errors.

Why the downstream noise penalty is additive. Even with unlimited training data, feature measurement can leave irreducible prediction error. For example, let a balanced latent bit determine the outcome exactly, and let two independent raters each flip it with probability $\eta = (1 - \sqrt{0.70})/2$. Their population agreement is 0.70, but a classifier observing only the labeler's bit has minimum test error $\eta \approx 0.082$, while the reference-feature oracle has zero error. An excess-risk bound consisting only of a constant times a vanishing estimation term cannot hold for this comparison. Here the prediction-change term in Theorem 1 equals η, exactly matching the irreducible deployed error.

Theorem 1 does not rely on noise inversion. For comparison, the loss-correction route faces a separate difficulty. For independently flipped coordinates with rates $0 \le \eta_k < 1/2$, let $T_k = \left(\begin{smallmatrix} 1-\eta_k & \eta_k \\ \eta_k & 1-\eta_k \end{smallmatrix}\right)$. Inversion of the joint corruption matrix has operator norm

$$\left\| \left(\bigotimes_{k=1}^{K} T_k \right)^{-1} \right\|_{\infty} = \prod_{k=1}^{K} (1-2\eta_k)^{-1}.$$

This model requires assumptions across coordinates beyond Assumption 1. Moreover, ordinary fitting on measured features does not minimize a noise-corrected loss. Scalar target-label corrections (Natarajan et al., 2013; Patrini et al., 2017) therefore do not justify a dimension-independent generalization penalty for the classifiers used here. This does not preclude sharper results for restricted heads; it explains why a scalar label-noise correction cannot justify the original multiplicative claim for a general joint classifier.

A.5 Direct transfer stability

A direct transfer audit freezes definitions and the predictor, replaces the annotator, and measures changes in values, predictions, and accuracy. The following bound describes this audit without latent-noise assumptions.

Proposition 1 (Fixed-predictor annotator transfer)**.** *Fix definitions, a classifier h, and a loss in $[0,1]$. Let $A(X), B(X) \in \{0,1\}^K$ measure the same texts and let $R_A(h) = \mathbb{E}[\ell(h(A(X)), Y)]$. Then*

$$\begin{aligned} |R_A(h) - R_B(h)| &\le \mathbb{P}\{h(A(X)) \neq h(B(X))\} \\ &\le \mathbb{P}\{A(X) \neq B(X)\} \\ &\le \min\left\{1, \sum_{k=1}^{K} \mathbb{P}(A_k(X) \neq B_k(X))\right\}. \end{aligned}$$

Proof. Losses agree when predictions agree and otherwise differ by at most one. Identical feature vectors yield identical predictions for a fixed classifier. The last inequality is a union bound; it requires no independence of errors. □

For balanced accuracy, apply the argument conditional on each outcome class and average with equal class weights. This bounds annotator-transfer sensitivity, not latent correctness or excess risk of a newly trained classifier. It motivates measuring feature disagreement and prediction changes on the same frozen audit texts. The sum bound may be loose, making the directly observed prediction-change rate a useful additional statistic. Prediction stability is not a substitute for feature-level correspondence: an insensitive classifier can remain stable despite substantial measurement disagreement.

B SELECTION ALGORITHM

Algorithm 1 separates measurement, eligibility, gain evaluation, and selection. Its maximization is over eligible candidates whose gains have already been computed. Training shortlisting is an efficiency device, not a guarantee of finding the global best validation candidate. The cross-LLM screen is motivated as a proxy for independent application, not a validated predictor of human agreement or a certification of operational feature interpretability. Human reproducibility and correspondence with model-used values require separate evaluation.

Bootstrap and residual proposal context. Bootstrap constructs the initial candidate pool from the training set. As gains plateau, the residual phase concentrates the proposal context on training texts with larger current prediction errors. Hardness uses the absolute difference between the predicted probability and the outcome, not only a misclassification indicator. Configured quantile, top-count, and sample-size controls determine the retained context. New proposal calls occur when the candidate pool is refreshed, with the accepted definitions supplied as context; accepting one feature does not require an immediate new LLM call. The phase transition, further concentration of the context, and termination depend on gain, patience, and proposal budgets.

Defined agreement and successful examination. A failed examination or undefined agreement does not satisfy Equation 1. In particular, identical constant label vectors have undefined Cohen's κ and are not reported as empirical $\kappa = 1$. A feature that is constant on the training texts cannot raise validation accuracy and therefore cannot be admitted. The chosen admission threshold is neither a universal cutoff for human operational interpretability nor a construct-validity certificate.

Sampling variants and run provenance. The reported pipeline uses class-conditioned hybrid minibatches, passing label-grouped examples to the proposer. All ten runs draw class-balanced minibatches of 64 texts (32 per class); no run pairs individual texts across classes. The contrast grounds feature proposal in observed texts; it does not control all other text attributes or identify a causal effect.

Selection has no backward-pruning step.

C DATASETS AND EXPERIMENTAL CONFIGURATION

Table 4: Tasks; 1,000 texts per task.

Dataset	Domain	Language	Outcome
AdParaphrase	Advertising copy	Japanese	Copy preference
Loan Application	P2P lending	Chinese	Loan default
CFPB	Consumer complaints	English	Relief obtained
Deceptive Reviews	Hotel reviews	English	Authenticity
Gaslighting	Social media	Chinese	Comment moderation
iSarcasm	Social media	English	Sarcasm
AG News ($\times 4$)	News	English	Topic, one vs. rest

C.1 REPRODUCIBILITY DETAILS

Splits are stratified 60/20/20 with seed 42. The model stack is `grok-4-fast-reasoning` for proposal, `gpt-4o-mini` for labeling, and `deepseek-chat` for examination, with temperature 0. Embeddings use `text-embedding-3-large` (3,072 dimensions).

Algorithm 1 LFD with separate proposal, labeling, and examination roles

Require: Training and validation sets; proposer $\mathcal{P}$, labeler $\mathcal{L}$, examiner $\mathcal{E}$; κ^*, minimum gain δ_g; shortlist size m, feature cap, proposal and patience budgets

$\mathcal{F} \leftarrow \emptyset$; initialize candidate pool and phase
while budget remains and stopping conditions are unmet **do**
  if the current phase requires new proposals **then**
    Draw class-conditioned minibatches from the current proposal context
    Add $\mathcal{P}$'s lexical and semantic proposals to the pool
  end if
  for each candidate f requiring measurement **do**
    if f has an executable lexical rule **then**
      Compute its values with fixed rule and preprocessing
      Record executable status; check nondegeneracy and duplication
    else
      Label with $\mathcal{L}$ and examine with $\mathcal{E}$ on the training set
      Admit to eligible pool only with a completed, defined $\widehat{\kappa}_f^{\mathcal{L},\mathcal{E}} \geq \kappa^*$
    end if
  end for
  Rank eligible, unselected candidates by incremental training score
  $\mathcal{V} \leftarrow$ at most m training-ranked candidates
  for $f \in \mathcal{V}$ **do**
    Fit the classifier on the training set using $\mathcal{F} \cup \{f\}$
    Compute Δ_f on the validation set
  end for
  if $\mathcal{V} \neq \emptyset$ **then**
    $f^\star \leftarrow \arg\max_{f \in \mathcal{V}} \Delta_f$
    if $\Delta_{f^\star} \geq \delta_g$ **then**
      $\mathcal{F} \leftarrow \mathcal{F} \cup \{f^\star\}$; update predictor and proposal context/phase
      Remove $f^\star$ from the candidate pool
    else
      Update patience counter, phase, and candidate refresh
    end if
  else
    Update patience counter, phase, and candidate refresh
  end if
end while
Freeze definitions; refit final predictor on combined training and validation sets
return features and predictor; evaluate on the test set

Selection and readout settings. Table 5 lists the settings shared by all ten runs. The final predictor uses these LightGBM settings without early stopping.

Comparator protocol. TBM uses `gpt-4o-mini` for generation and measurement, trial size 50, relative trial-improvement threshold 0.01, and at most 15 iterations. Cross-annotator comparison must preserve native response categories and mappings; regenerating the mapping can change the feature being compared. Native and common-head results should be separate, with equivalent validation opportunities and fitting data for common heads.

Available readout analyses. The completed follow-up record supplies LFD-only classifier-sensitivity results, retained in Appendix H. These refits use a 400-estimator LightGBM configuration and are not a matched-head comparison between LFD and TBM. Neither set of results supports universal classifier invariance.

Uncertainty. Compute test uncertainty from saved predictions on the same instances, respecting source-entity clusters where applicable. A nonsignificant difference does not establish equivalence. Discovery-seed variability is distinct from refitting heads on one feature matrix.

Table 5: Settings shared by all ten LFD runs.

Parameter	Value
Minimum validation gain δ_g	0.001
Minibatch size	64 (32 per class)
Proposal batches per refresh	3
Maximum proposals per batch	25
Validation shortlist m	10
Feature cap	40
LightGBM estimators	400
LightGBM leaves	64
Learning rate	0.05
Minimum child samples	1% of fitting rows
Column subsampling	0.8
Class weighting	None
Seed	42

D Feature specifications and prompt documentation

A feature release should include its full definition, executable rule if present, inclusion/exclusion cases, labeling instructions, and lane metadata. Truncated regex strings and abbreviated procedures cannot support independent reapplication.

Table 6: Illustrative measurement distinctions. A suggestive name does not add conditions absent from a rule.

Example	Procedure	Interpretation
Repeated exclamation marks	Match three or more consecutive ! characters	Detects punctuation, not sarcasm itself.
Score-shaped number pair	Match one/two digits, a hyphen, and one/two digits	Can match non-score strings; sport context is additional.
Statutory citation	Match specified statute names or identifiers	Does not establish relevance or correctness of a citation.
Documentation as support	Determine whether referenced documentation supports the complaint	Requires semantic application and observed agreement.

The lexical proposal prompt asks for 3–5 features distinguishing two outcome groups, with a name, operational definition, and Python-compatible regex. The semantic prompt emphasizes narrative, rhetorical, credibility, or institutional properties and omits a regex. The examiner prompt requests binary application of the supplied definition. These summarize roles; they are not complete executable templates. The full prompt templates will be released with the code.

E Human application audit and reanalysis

Sources and quality checks. The reanalysis uses the original iSarcasm v2, AG News Sci/Tech, and CFPB response collections. It does not combine iSarcasm v1 with v2. The two CFPB batches cover all 21 LFD and five TBM definitions; an older export repeating one batch's responses is excluded to avoid duplication. Both methods' binary-reference columns come from the same task-level instance pools. Retain completed sessions with duration at least 300 seconds and every recorded attention item answered correctly. No exclusion depends on agreement with pipeline labels. Three Sci/Tech sessions fail attention checks and are excluded. All iSarcasm sessions pass them.

Panels and estimands. For each feature, use the first three retained sessions by recorded completion time, with input order resolving ties. This rule does not use agreement. Most features have only three raw responses. Require at least ten common texts with valid labels from all three sessions; eligible panels contain 22–25 texts. Different features and methods need not have identical sampled texts. The reference pool has 100 texts per task, sampled for the original audit rather than representative deployment. TBM pipeline values greater than zero are treated as feature presence, matching the

Table 7: Quality-control counts. Sessions are not verified unique people. Attention exclusions are counted after duration exclusions.

Task	Method	Raw	Too fast	Failed attention	Retained
iSarcasm	LFD	46	2	0	44
	TBM	18	1	0	17
AG News S/T	LFD	76	6	3	67
	TBM	27	2	0	25
CFPB	LFD	75	9	0	66
	TBM	15	0	0	15

binary survey. Its HM–P statistic therefore assesses correspondence with this coarsened reference, not exact reproduction of the full native scores used by its predictor.

Compute Cohen's κ for each of the three human pairs and average defined values within a feature. Method means weight features equally. Equal constant vectors have undefined κ and perfect raw agreement; report both without imputation. Human-majority–pipeline agreement uses the same three-person panel and hence has no majority ties. The pipeline reference may be a deterministic rule or a semantic labeler. It should not universally be called an LLM judgment.

Human–human agreement measures reproducibility of application, whereas human-majority–pipeline agreement measures correspondence with the pipeline reference (binary presence for TBM). These provide distinct, complementary evidence about conceptual clarity, one requirement of operational feature interpretability, for the audited users and texts. Majority correspondence does not establish that an individual user can reproduce the coordinate, and neither statistic certifies construct validity.

Coverage. On iSarcasm, one LFD feature and one TBM feature lack three retained sessions. An additional LFD majority–pipeline statistic is undefined because both vectors are constant. On Sci/Tech, eight LFD features lack three retained sessions and two more have undefined HH κ from equal constant labels. Two TBM features lack three retained sessions. Some otherwise eligible panels have only two defined human-pair kappas. All CFPB definitions have three retained sessions. Two lexical features (Casual Hi/Hello opener and Claim Denied) have undefined HH and HM–P κ from constant labels; Court case V citation has defined HH but undefined HM–P κ. Thus CFPB has 19 of 21 defined LFD HH values and 18 defined HM–P values; all five TBM values are defined for both. Feature-level coverage and reasons are included in the analysis artifact.

Table 8: Human–human measurement reproducibility by LFD lane. Executable metadata determines the lane, not an observed value of κ.

Task	Lane	Mean HH κ	Defined/total
iSarcasm	Lexical	0.578	12/13
	Semantic	0.185	2/2
AG News S/T	Lexical	0.736	11/20
	Semantic	0.394	5/6
CFPB	Lexical	0.615	18/20
	Semantic	0.669	1/1

All three task-level LFD mean HH agreements are below 0.70. This is the chosen LLM admission threshold, not a validated cutoff for human interpretability. Six of fourteen defined iSarcasm LFD feature means and eight of sixteen Sci/Tech means reach 0.70. Neither iSarcasm semantic mean and only one of five defined Sci/Tech semantic means reaches that level. On CFPB, ten of nineteen defined LFD means reach 0.70; its sole semantic feature does not. The small semantic panels and missingness prevent strong population conclusions, but the pooled result should not obscure these observations.

Uncertainty and sensitivity. The artifact supplies percentile intervals from 10,000 feature resamples with seed 20260918. For the LFD-minus-TBM difference in mean HH κ, the intervals are $[-0.026, 0.545]$ on iSarcasm, $[0.044, 0.488]$ on Sci/Tech, and $[0.130, 0.725]$ on CFPB. These are descriptive feature-resampling intervals conditional on observed definitions and panels. They do not account for shared participants, shared texts, feature discovery, or deployment representativeness, and

are not definitive population significance tests. With only two iSarcasm semantic features, resampling is particularly limited.

Table 9: Same-collection sensitivity to the duration threshold. Both analyses enforce actual attention checks and exclude no one for disagreement.

		At least 300 sec.		No duration exclusion	
Task	Method	HH κ	Defined	HH κ	Defined
iSarcasm	LFD	0.522	14	0.569	15
	TBM	0.269	5	0.297	6
AG News S/T	LFD	0.629	16	0.638	21
	TBM	0.361	7	0.412	9
CFPB	LFD	0.618	19	0.611	19
	TBM	0.178	5	0.178	5

Reproduction. The accompanying aggregate artifact includes the analysis script, input-file hashes, per-feature statistics, and exclusion counts. It exports neither raw responses nor participant identifiers. Running the script recomputes the tables from the existing exports. The implementation was checked against an independent contingency-table κ calculation on all 1,024 pairs of length-five binary vectors, including degenerate cases, and against the recorded attention-field format.

Further validation. Ground-truth concept annotations and controlled edits would add evidence about fidelity and sensitivity. CEBaB, for example, supplies annotated aspect concepts and edited reviews (Abraham et al., 2022). Such evaluation complements agreement; the current experiments do not establish causal validity or complete recovery of latent concepts.

F Definition-level rubric

This is a separate study from application agreement. Ratings of descriptions do not directly test users' ability to apply those descriptions or reproduce pipeline values. The exports have 35 sessions, of which 34 are completed and last at least 300 seconds. One short AG News Sports response is excluded. There are no attention-item columns, so no attention filtering is claimed. Missing feature ratings are omitted. First average available ratings within a feature, then average equally over features within a task/method. The feature sample covers all available TBM concepts and a subset of LFD features on most tasks, rather than all admitted features uniformly.

Table 10: Definition-rubric coverage and descriptive means on a five-point scale. The favorable anchor describes independence from the answer.

Task	Sessions	LFD features	TBM features	LFD mean	TBM mean
AdParaphrase	3	10	9	3.767	2.185
AG News Business	3	10	6	2.933	2.667
AG News S/T	3	10	9	3.500	2.981
AG News Sports	10	14	2	2.749	2.278
AG News World	3	10	4	3.667	1.500
CFPB	3	10	5	3.400	2.600
Deceptive Reviews	2	10	9	2.750	2.333
Gaslighting	3	10	6	3.667	3.444
iSarcasm	1	15	6	3.267	1.333
Loan Application	3	3	6	4.111	3.333

Sci/Tech TBM features have two or three ratings each. Sports LFD features have nine or ten; Sports TBM concepts have nine. Other features have the listed task-level count. The direction favors LFD on all ten tasks, but the small, uneven panels do not support strong per-task inferential claims.

Table 11: CFPB threshold sensitivity. These separate discovery runs do not isolate screening a common candidate pool. Threshold 0.70 is the configured default, not an established optimum.

κ^*	Selected features	Test BA
0.50	14	0.565
0.60	16	0.555
0.65	21	0.510
0.70	21	0.565
0.75	17	0.540
0.80	10	0.555
0.85	8	0.560

G SUPPLEMENTARY SENSITIVITY RESULTS

Counts do not decrease monotonically over all thresholds, and no accuracy collapse appears at 0.85. Separate CFPB runs restricted to one lane give lexical-only BA 0.540 with $K = 19$ and semantic-only BA 0.500 with $K = 0$.

Incremental embedding value. Paired-bootstrap intervals for adding an embedding to the LFD basis are AdParaphrase -0.035 ($[-0.114, 0.046]$), CFPB $+0.049$ ($[-0.022, 0.120]$), and Loan Application $+0.010$ ($[-0.033, 0.051]$). These intervals include zero and allow meaningful improvements; they do not establish saturation of predictive information. These come from a separate refit analysis whose LFD-only estimates differ from Table 2.

Proposal variation. The completed four-proposer comparison is reported alongside the fixed-definition rater audit and full-discovery role substitutions in Appendix H.4. These studies distinguish model choice from repeated-discovery variability; classifier-seed variation alone does not establish discovery stability.

H COMPLETED FOLLOW-UP STUDIES

These studies extend the original evaluation with post-discovery association analysis, lane-separated analyses, additional predictive comparisons, model-role substitutions, and a third human audit. We preserve their study-specific denominators and evaluation settings. The feature-level analyses cover 202 LFD features and 62 TBM concepts. For Deceptive Reviews they use a 22-feature basis from an earlier run on the same splits, whereas Table 2 reports a 23-feature basis; the other nine bases are identical. The frozen-pool and downstream-head refit studies use a 400-tree LightGBM reference with mean BA 0.707, whereas Table 2 reports the original mean 0.708. These follow-up estimates do not replace the original task-level results.

H.1 LABEL ASSOCIATION AND POST-DISCOVERY FILTERING

The discovered LFD basis has pooled mean $|\rho(f, y)| = 0.121$ across 202 features, compared with 0.267 across 62 TBM concepts. LFD has lower mean association on seven of ten tasks; Loan Application, Deceptive Reviews, and CFPB favor TBM. On AG News Sports, TBM's two concepts, "mention of athlete or sports team" and "mention of sports terminology," have $|\rho| = 0.953$ and 0.914, respectively; LFD's 33 features average 0.150. This is a system-level comparison, not an isolated causal test of the proposal objective or incremental-gain selector. Statistical association is also distinct from whether a definition semantically restates the target.

In a separate post-discovery analysis, we assess the predictive cost of imposing a feature–label association cap on a fixed LFD basis. This analysis does not alter the discovery procedure. We froze each task's LFD pool, filtered it at $\tau \in \{0.5, 0.3, 0.1\}$ using training-set association, refit on the surviving features, and evaluated BA on the untouched test set (Table 12). Every surviving feature conforms on the sample used for filtering. This enforces the empirical association bound, not a population bound or semantic non-paraphrase. The predictive cost is threshold-dependent: the strictest setting removes more than half the basis and substantially reduces BA.

Table 12: Post-discovery association filtering over ten tasks. Removal is averaged per task; maximum association is the worst task's surviving maximum on the training set. BA is evaluated on the test set.

τ	Mean K	Mean removed	Max $\lvert\rho\rvert$	Mean BA	ΔBA
Unfiltered	20.2	0.0%	0.575	0.707	0.000
0.5	20.1	3.3%	0.400	0.683	-0.023
0.3	19.7	4.9%	0.296	0.674	-0.032
0.1	8.6	56.6%	0.100	0.551	-0.155

Table 13: Held-out cross-LLM agreement by feature group. The numerator counts defined agreement estimates; the denominator counts all features.

Feature group	Defined / total	Mean κ	Median κ
LFD semantic	24/25	0.60	0.66
TBM concepts	55/62	0.22	0.18

H.2 Clarity and predictive contribution by lane

Table 13 separates semantic definitions from executable lexical rules. The admission operating point $\kappa^* = 0.70$ applies to the admission sample; the semantic median 0.66 is re-measured on fresh held-out documents. The relevant held-out comparison is therefore 0.66 versus 0.18 for TBM, not whether the held-out point estimate repeats the admission threshold. The remaining 177 features are executable lexical rules. Their deterministic execution is reported separately rather than assigned an empirical cross-rater κ of one; it does not establish human understanding, correctness, or semantic validity. Undefined agreements remain in coverage counts but not in means or medians.

Keeping the discovered basis fixed, we refit with and without its semantic lane (Table 14). Semantic features increase held-out BA on five of the eight tasks containing them, tie on one, and reduce it on two. Their median share of final-model split gain is 18.5%, reaching 42–45% on AG News Sports and Sci/Tech. This is a conditional contribution analysis of an existing basis, not semantic-only rediscovery; split-gain importance is model-specific and not a causal attribution.

H.3 Interpretable heads and an embedding-derived concept bank

On the accepted feature bases, an L1-logistic head and a pruned single decision tree obtain mean BA 0.717 and 0.708, respectively, versus 0.707 for the refit LightGBM reference (Table 15). Logistic regularization and tree pruning are selected by five-fold cross-validation within training data. The intentionally shallow trees trade predictive performance for shorter rule sets. These results show that competitive observed performance need not require a boosted-tree head; they do not establish statistical equivalence of heads.

The completed embedding-derived comparison generates approximately 80 generic concept phrases per task, excluding outcome-category names, and uses document–phrase embedding cosine similarities as coordinates. Its L1-logistic head reaches mean BA 0.754 and exceeds LFD on seven of ten tasks; matching bank size to LFD's K gives 0.722. Thus, named embedding coordinates can retain predictive information unavailable to the more restricted Boolean representation. This is a custom concept-bank implementation.

To test whether a phrase explains its coordinate, an examiner applied the ten highest-weighted concepts on each of three tasks to 100 held-out documents. Table 16 compares those binary judgments with cosine coordinates binarized at a median split. The mean $\kappa = 0.14$ indicates weak correspondence under this particular audit. The median threshold is a diagnostic choice, not ground truth, and this coordinate-to-name agreement is not identical to rater-to-rater agreement on an explicit Boolean definition. The result concerns this implementation and does not establish a general limitation of all concept bottlenecks.

Table 14: Predictive contribution of semantic features in fixed bases. Full-basis BA reproduces the corresponding original main-table values. Loan Application and Deceptive Reviews contain no semantic features.

Task	BA full	BA lexical	ΔBA	Semantic split gain
AG News Sports	0.885	0.845	+0.040	42%
AG News Sci/Tech	0.855	0.770	+0.085	45%
AG News World	0.810	0.785	+0.025	25%
AdParaphrase	0.665	0.645	+0.020	18%
iSarcasm	0.575	0.575	+0.000	19%
AG News Business	0.750	0.730	+0.020	8%
CFPB	0.565	0.590	−0.025	12%
Gaslighting	0.605	0.620	−0.015	9%

Table 15: Downstream-head comparison on fixed LFD bases, averaged over ten tasks. The unrestricted single tree is pruned using cross-validation.

Head	Mean BA
LightGBM (400 trees)	0.707
L1-logistic	0.717
Single pruned decision tree	0.708
Depth-$\leq$ 2 tree	0.618
Depth-$\leq$ 3 tree	0.636

H.4 Proposer, rater, and model-role sensitivity

The four-proposer study holds the labeler (GPT-4o mini) and examiner (DeepSeek Chat) fixed while varying the proposer on AdParaphrase and CFPB (Table 17). BA ranges are 0.090 and 0.040, respectively; reported basis sizes span 4–22 and 5–13 features. The reported bootstrap intervals overlap, but this does not establish equivalent predictive performance or an invariant feature basis. This is a separate study: its evaluator fits on combined training and validation data, and its CFPB Grok BA of 0.560 differs from the 0.565 main-table reference used in the role-substitution study below.

We held 76 accepted definitions fixed and asked six models to label the same 200 held-out documents per definition. Table 18 reports each model's median feature-level agreement against the stronger rater peers in the tested cohort. Same-family and cross-family strong pairs have similar medians (0.79 versus 0.80), while GPT-3.5 Turbo reaches 0.30. Model family alone therefore does not explain agreement in this cohort. Neither cross-family sampling nor high agreement establishes independent errors or semantic correctness.

The complementary full-discovery experiment replaces one role at a time with GPT-3.5 Turbo, keeping the other roles and the evaluation protocol fixed (Table 19). Every substitution produces a smaller basis, and BA losses depend on role and task. The semantic definitions admitted in the weak-proposer condition have strong-model re-examination κ values 0.756 and 1.000. This limited check supports reproducible application of those admitted definitions, not an unrestricted claim that weak components can only reduce recall or cannot introduce errors. The supported robustness claim is restricted to the tested models and settings. Prompt invariance and repeated-discovery stability are not established by these role-substitution results.

H.5 Third human clarity audit: CFPB

This follow-up extends the application audit to a regulated-finance domain. CFPB also has a contrasting association profile: it is one of the three tasks on which TBM has lower mean feature–label correlation than LFD. The human task remains application of supplied definitions, not independent discovery of features or a rating of their names. The extension uses the same clarity-audit protocol on all 21 LFD features and all five TBM concepts, with each survey presenting 25 documents. We reanalyze both response batches using the same quality checks and three-person panel rule as the other tasks (Appendix E). Of 75 LFD and 15 TBM response sessions, 66 and 15 pass quality checks;

Table 16: Concept-name judgments versus median-binarized cosine coordinates: ten highest-weighted concepts and 100 held-out documents per task.

Task	Concepts	Mean κ
AG News Sports	10	0.11
Deceptive Reviews	10	0.22
AG News Sci/Tech	10	0.08
Pooled	30	0.14

Table 17: Four-proposer sensitivity with labeler and examiner fixed. Entries are held-out BA with reported 95% bootstrap intervals (1,000 replicates), using the study-specific evaluator.

Proposer	AdParaphrase	CFPB
Grok-4-fast-reasoning	$0.665\ [0.60, 0.73]$	$0.560\ [0.49, 0.63]$
GPT-4o	$0.645\ [0.58, 0.70]$	$0.575\ [0.51, 0.64]$
Claude Sonnet 4.5	$0.575\ [0.51, 0.65]$	$0.535\ [0.46, 0.60]$
DeepSeek Chat	$0.620\ [0.56, 0.69]$	$0.555\ [0.49, 0.62]$

these are session counts, not verified counts of unique people. Every definition retains a three-person panel, with 22–25 common valid document labels. Mean HH κ is 0.618 for LFD versus 0.178 for TBM; the medians are 0.733 and 0.111 (Table 20). Two constant-label LFD features have undefined HH κ, leaving 19 defined values. Mean HM–P κ is 0.855 versus 0.168, over 18 and five defined values, respectively.

The descriptive feature-resampling interval for the HH mean difference is $[0.130, 0.725]$; it is conditional on the observed definitions and panels, not a population significance test. Removing the duration exclusion while retaining attention checks gives mean HH $\kappa = 0.611$ for LFD and 0.178 for TBM. The single semantic feature has primary HH $\kappa = 0.669$, so this task cannot establish a broad semantic-lane advantage.

Table 18: Rater sensitivity on 76 fixed definitions and 200 held-out documents per definition. Agreement is measured against stronger-model peers, not against ground-truth feature values.

Rater	Median feature-level κ
GPT-4.1	0.83
DeepSeek Chat (main examiner)	0.81
Claude Opus 4.5	0.80
Claude Sonnet 4.5	0.79
GPT-4o mini (main labeler)	0.71
GPT-3.5 Turbo	0.30

Table 19: One-role-at-a-time GPT-3.5 Turbo substitutions in full discovery. Reference values are the original task results; evaluation uses the held-out test set. These are individual runs, not estimates of rerun variability.

Substitution	Task	K	Ref. K	BA	Ref. BA	ΔBA
Proposer	AdParaphrase	1	22	0.520	0.665	-0.145
	CFPB	8	21	0.500	0.565	-0.065
	AG News Sci/Tech	22	26	0.775	0.855	-0.080
Labeler	AdParaphrase	4	22	0.615	0.665	-0.050
	CFPB	12	21	0.565	0.565	$+0.000$
	AG News Sci/Tech	5	26	0.590	0.855	-0.265
Examiner	AdParaphrase	3	22	0.615	0.665	-0.050
	CFPB	5	21	0.535	0.565	-0.030
	AG News Sci/Tech	15	26	0.805	0.855	-0.050

Table 20: CFPB human clarity audit using the common three-task analysis protocol. Undefined statistics are omitted without imputation.

Method	Defined / covered	Mean HH κ	Median HH κ
LFD	19/21	0.618	0.733
TBM	5/5	0.178	0.111

I CFPB FEATURE EXPLANATIONS

Table 21 explains all 21 features in Figure 1, preserving their names and admission order. Features 1–6 enter during bootstrap; features 7–21 enter during residual discovery. The explanations summarize the archived feature definitions and detection rules. Feature 4 uses an LLM judgment; the others detect specified text patterns. These descriptions state what the rules detect, without inferring additional intent or legal validity from a feature's name.

Table 21: Feature-by-feature explanation of the CFPB trace in Figure 1.

k	Feature	Explanation of the recorded rule
1	FCRA Statute Cite	Contains *15 USC 1681*, or *FCRA* followed by *violation* or *act*.
2	Refund Request	Contains the substring *refund*.
3	Card Not Received	Contains *card*, then *never* or *did not*, then *arrived* or *received*; alternatively, *did not receive* occurs before *card*.
4	Legal Cite Misapplication	Locate citations to specific laws or statutes and assess whether their application to the facts is vague or incorrectly interpreted. The feature is present when the misapplication fails to tie the alleged violation to concrete company actions.
5	Unauthorized Inquiry	*hard* or *unauthorized* occurs before *inquiry* or *inquiries*.
6	Billing Error Claim	Contains *billing error*, allowing whitespace between the words.
7	Debt Validation Req	Contains *debt validation*, *validate debt*, or *1692g*.
8	Claim Denied	*claim* or *investigation* occurs before *denied*.
9	Specific Date Mention	Contains the redacted date format `XX/XX/` followed by four digits.
10	Method Verification Req	Contains *method of verif*, matching the beginning of phrases such as *method of verification*.
11	Items Listed Enumeration	Contains *the following items*, *these items*, *items listed*, *items I am disputing*, or *items below*.
12	Reputation/character harm	Contains *reputation*, *character*, *mode of living*, or *standing in the community*.
13	States I Have	Contains *states I*, *states that I*, *states a consumer*, *states the consumer*, or *state I have*.
14	Autopay Late Error	*autopay* occurs before *late*, *error*, or *billing*.
15	Section 604 cite	Contains *Section 604*, *604 A*, or *1681 Section 604*.
16	Did not apply phrase	Contains *did not apply*, *did not authorize*, *never applied for*, *i didnt apply*, or *i didn't apply*.
17	Court case V citation	Contains a court-case-style pattern: capitalized name words, then *v.*, *V.*, or *vs.*, then another capitalized name.
18	Casual Hi/Hello opener	The text starts with the greeting *Hi*, *Hi,*, *Hello*, or *Hey*.
19	Customer Service Mention	Contains *customer service*, allowing whitespace between the words.
20	Looking Forward Closing	Contains *look forward to* or *looking forward to* anywhere in the text.
21	Pursue Legal Threat	Contains *pursue legal*, *will pursue*, *legal action against*, or *take legal action*.