EVIDENCE CONTRACTS FOR DOMAIN-MAPPED PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A statistically predictive feature need not be admissible evidence for the domain interpretation attached to a prediction. To address this tension, we introduce source-qualified evidence contracts: a framework that enforces evidentiary constraints on an interpretable semantic head while leaving an unrestricted predictor intact. Four-state observations (positive, negative, unknown, conflicting) compile into bounded admissible concept completions. We establish a conditional error bound under concept-dependent reporting shift, distinguishing expected-band invariance from probability certification and demonstrating via counterexample why ordinary calibration is insufficient. In a five-seed replication at matched coverage, probability certificates cap maximum mean selective error at 1.66% across seven valid-contract regimes, outperforming a same-contract MLP (4.19%) and a typed MLP (3.08%). However, exhaustive validation reveals that learned heads fail the theorem's pointwise assumption, precluding a nominal guarantee in practice. Supported by a text-extraction diagnostic and an exact arithmetic baseline on 5,070 real-world records, our work rigorously delineates admissible evidence, robustness assumptions, and measured selective utility.

1 INTRODUCTION

A mental-health severity predictor may be statistically useful without making a justified clinical claim. "I cannot sleep now," a report about past insomnia, and a report about someone else's insomnia all contain symptom-related language. Missing audio is different from an observed denial. A change in gaze or prosody may corroborate a behavioral interpretation without establishing a particular disorder. These distinctions concern *what an observation is allowed to support*, not merely whether it predicts a target.

Feature attribution and construct perturbation identify aspects of model dependence. They do not automatically establish that the affected features are admissible evidence for the interpretation attached to the prediction. Conversely, excluding an inadmissible feature can reduce accuracy when that feature happens to be predictive. Treating these as separate questions is important in human review, where an explanation can turn a useful association into an unjustified symptom claim.

We study this distinction through *evidence contracts*. A registry determines which source observations may constrain a construct, using role, speaker, temporality and quality. The contract compiles these observations into a set of permitted concept values while keeping contradictions visible. A learned semantic head operates within that set. An unrestricted predictor remains separate, so that predictive dependence is not silently presented as evidentiary support. The semantic head is a representation of the specified contract, not a substitute for a clinician or a guarantee that the registry is correct.

The original dual-map system motivates separating a registry's evidentiary assignments from a model's learned dependence. We turn that distinction into a computational restriction, then study reporting shift: which facts are observed can change even when the relationship between complete concepts and outcomes does not.

Our contributions are: (1) a source-qualified contract with a separate unrestricted predictor and conflict-aware bounded completion; (2) a probability certificate with an explicit pointwise assumption, a proof of reporting-mechanism robustness, and a counterexample to replacing that assumption

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.

with ordinary calibration; and (3) a five-seed local replication with matched-coverage controls, a real-item semi-synthetic audit and an executable text-to-contract diagnostic. Original clinical results and earlier component experiments remain separately attributed in the separate supplement.

2 Related work and positioning

Concept-based models. Concept bottleneck models predict through interpretable variables and support interventions (Koh et al., 2020). TCAV measures concept sensitivity (Kim et al., 2018); post-hoc concept bottlenecks support concept-based interpretation and editing (Yuksekgonul et al., 2023). Completeness-aware explanations study the sufficiency of concept representations (Yeh et al., 2020), while concept-residual models expose the tradeoff between predictive information and leakage (Zabounidis et al., 2023). Concept leakage is also studied by Mahinpei et al. (2021); Havasi et al. (2022). Our distinction concerns permission to use a source as evidence. We do not identify an unrestricted residual with unexplained variance or claim that source restrictions make the learned concepts causally correct.

Intervals, monotonicity and abstention. Nonnegative or lattice-based models can enforce monotonicity (You et al., 2017). CREDENCE explicitly represents concept uncertainty with intervals and studies routing (Mukherjee et al., 2026); intervals and uncertainty-guided routing are therefore not new by themselves. Our intervals encode admitted observations and unresolved conflict, rather than an estimated epistemic–aleatoric decomposition. They bound a learned head over permitted completions and do not provide statistical coverage for an outcome. Selective classification already studies error versus coverage (Geifman and El-Yaniv, 2017). A contract-disagreement statistic must earn incremental predictive utility; it is not an abstention guarantee by construction.

Evidence states and validation. The distinction between positive, negative, unknown and inconsistent information has a long history in four-valued logic (Belnap, 1977). We use that distinction, rather than claim to invent it. The specific object here combines source eligibility, a semantic computation, an independent predictive path and tests of their separation. Sanity checks for explanations motivate testing properties independently of visual plausibility (Adebayo et al., 2018). BioScope supplies annotated linguistic scope (Vincze et al., 2008); SNLI, MultiNLI, ANLI and HANS provide complementary language-inference and transfer tests (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020; McCoy et al., 2019). None supplies depression or PTSD severity labels.

3 Evidence contracts

3.1 Observations, permissions and four states

Let o_j be a source observation and c a domain construct. A registry records the source pointer, modality, construct, evidentiary role, speaker, time applicability, polarity and quality rule. Its eligibility function $e_{jc} \in \{0, 1\}$ determines whether observation j may constrain construct c. A clinical registry may reserve direct symptom support for current, participant-attributable statements and mark behavioral information as non-specific corroboration. These are domain assumptions to validate, not rules learned from severity labels.

For the implemented binary-concept prototype, aggregate eligible support and refutation:

$$S_c = \bigvee_j (e_{jc} \wedge s_{jc}), \qquad R_c = \bigvee_j (e_{jc} \wedge r_{jc}). \tag{1}$$

Their four states compile to bounds and a conflict indicator:

$$\ell_c = \mathbb{1}[S_c \wedge \neg R_c], \quad u_c = 1 - \mathbb{1}[R_c \wedge \neg S_c], \quad b_c = \mathbb{1}[S_c \wedge R_c]. \tag{2}$$

The permitted set is $\mathcal{C}(E) = \{c : \ell \leq c \leq u\}$. Positive-only and negative-only observations fix a binary coordinate; unknown and conflicting observations leave it unrestricted. Conflict remains separately visible. A contradiction is not averaged into an ordinary absence. Earlier completion models used only the bounds, so conflict affected their audit but not their point prediction; Round 3 explicitly supplies the conflict flag to completion.

Table 1: The contract preserves evidentiary distinctions before prediction. Intervals are modeling constraints, not confidence intervals.

State	(S, R)	$[\ell, u]$	Interpretation
Observed support	$(1, 0)$	$[1, 1]$	Eligible affirmative observation.
Observed refutation	$(0, 1)$	$[0, 0]$	Eligible negative observation.
Unknown	$(0, 0)$	$[0, 1]$	No qualifying observation.
Conflict	$(1, 1)$	$[0, 1]$	Both polarities; retain conflict flag.

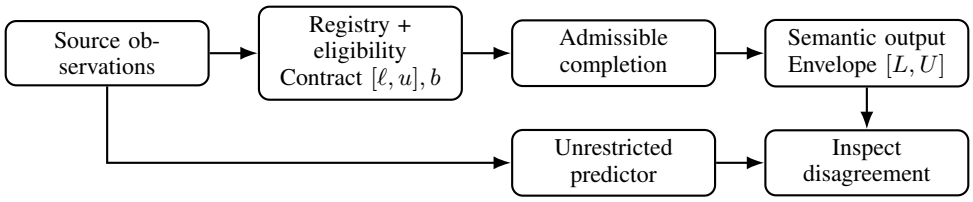

Figure 1: Two computational paths with different meanings. Only the contract path supplies the semantic claim. An unrestricted prediction can disagree; it cannot change the contract or convert residual dependence into admitted evidence. The envelope bounds the specified semantic head, not clinical truth.

The implementation accepts these structured inputs. It does not yet infer an entire clinical registry from raw interviews. In particular, an utterance denying a symptom does not necessarily establish that the symptom is globally absent, and temporally distinct reports need not contradict. The binary model is a controlled abstraction of a qualified observation within a specified interval. It does not license automatic clinical conclusions from keywords.

3.2 An admissible semantic path and a separate predictive path

We distinguish a semantic completion from an observation. The midpoint version uses $\widehat{c} = (\ell+u)/2$. Round 3 uses conflict-aware completion; Rounds 1–2 omit b. Both permit completion only inside the contract:

$$\widehat{c} = \ell + (u - \ell) \odot \sigma(g_\phi([\ell; u; b])). \tag{3}$$

Known coordinates cannot be overwritten by task optimization. Unknown coordinates may be imputed using the admitted information, but the output record marks them as imputed. The completion network cannot shrink the contract used for auditing and receives no unrestricted residual input.

The semantic head is a cumulative-link ordinal model with nonnegative weights $w_c = \mathrm{softplus}(v_c)$, score $a(c) = \beta + w^\top c$, ordered thresholds $t_0 < \cdots < t_{K-2}$ and temperature $T > 0$:

$$q_k(c) = P_{\mathrm{sem}}(Y > k \mid c) = \sigma((a(c) - t_k)/T), \quad \mu_{\mathrm{sem}}(c) = \sum_{k=0}^{K-2} q_k(c). \tag{4}$$

Adjacent differences of q_k give nonnegative class probabilities summing to one. The implementation orders thresholds using cumulative positive gaps. This is a standard compact monotone head; the contract specifies its permissible information, not a new ordinal likelihood.

An unrestricted predictor $p_{\mathrm{raw}}(y \mid x)$ may use information outside the contract. In the prototype it is an MLP over structured inputs; the original clinical model uses multimodal encoders and quality-aware fusion. The two heads are not interchangeable implementations. They can disagree without one output being silently rewritten as the other.

3.3 Output envelopes and precise guarantees

Define an expected-band envelope and an unrestricted-head disagreement:

$$L(E) = \mu_{\mathrm{sem}}(\ell), \qquad U(E) = \mu_{\mathrm{sem}}(u), \tag{5}$$

$$d(E, x) = [L(E) - \mu_{\mathrm{raw}}(x)]_+ + [\mu_{\mathrm{raw}}(x) - U(E)]_+. \tag{6}$$

The envelope requires two evaluations of the semantic head. It does not require one perturbation per construct. Extraction and the unrestricted forward pass have separate costs.

Proposition 1 (Conditional containment) *For every $c \in \mathcal{C}(E)$, $L(E) \leq \mu_{\text{sem}}(c) \leq U(E)$.*

Proof. Nonnegative weights imply $a(\ell) \leq a(c) \leq a(u)$. Each sigmoid is increasing, and summation preserves the inequalities. The learned completion lies inside the same box. □

Proposition 2 (Contract relaxation) *If $\ell' \leq \ell$ and $u' \geq u$, then $[L(E), U(E)] \subseteq [L(E'), U(E')]$.*

Proof. Apply coordinatewise monotonicity to each endpoint. □

Contract relaxation includes withdrawing all observations for a construct. It does *not* include every possible deletion of a source statement: deleting one side of a conflict can resolve the conflict and narrow the permitted set. This qualification prevents an incorrect claim that less text must always mean a wider envelope.

Proposition 3 (Noninterference) *Holding the compiled contract fixed, changing unrestricted inputs cannot change the semantic distribution or its envelope.*

Proof. Neither the completion function nor the semantic head has a computational dependency on those inputs. □

These three properties are conditional computational guarantees, not clinical truth guarantees. An erroneous extractor can compile an incorrect contract. An incorrect registry can enforce an incorrect meaning. An incomplete construct set can exclude useful information. The envelope bounds expected outputs of the specified model, not integer outcomes with a stated coverage probability. Noninterference originates in information-flow security (Goguen and Meseguer, 1982). A masked conventional model can also satisfy noninterference; the property must not be mistaken for exclusive evidence of architectural novelty.

3.4 Probability certificates under reporting shift

Expected-band invariance does not control decision error. Instead, fix an ordinal cut k and $\alpha \in (0, 1/2)$, and certify $Y > k$ when $q_k(\ell) \geq 1 - \alpha$, or $Y \leq k$ when $q_k(u) \leq \alpha$. Otherwise abstain. Let c^* be the complete concepts and $\eta(c) = \Pr(Y > k \mid c)$.

Proposition 4 (Conditional reporting-mechanism robustness) *Assume (i) $c^* \in \mathcal{C}(E)$ almost surely; (ii) $Y \perp E \mid c^*$; and (iii) for every supported c, $\eta(c) \geq q_k(c)$ whenever $q_k(c) \geq 1 - \alpha$, and $\eta(c) \leq q_k(c)$ whenever $q_k(c) \leq \alpha$. Each nonempty certified decision group has conditional error at most α, under any concept-dependent observation mechanism satisfying (i)–(ii).*

Proof. On a positive certificate, monotonicity and (i) imply $q_k(c^*) \geq q_k(\ell) \geq 1 - \alpha$. By (ii)–(iii), $\Pr(Y \leq k \mid E, c^*) = 1 - \eta(c^*) \leq \alpha$. Average over the certified group. The negative case is symmetric. □

The assumption is pointwise one-sided conservativeness, *not* score calibration or a small expected calibration error. For example, let complete concepts $(1, 0)$ and $(0, 1)$ occur equally, with outcome probabilities .8 and 1. The monotone head $q(c) = \sigma(\log(9)(c_1 + c_2))$ predicts .9 for both, hence is score-calibrated. Reveal the first type fully and hide both facts for the second. With $\alpha = .1$, only the first type is certified and its error is .2. Contracts remain correct and observation depends only on concepts; condition (iii) fails.

If the pointwise inequalities instead hold with uniform slack ϵ, the proof gives error at most $\min(1, \alpha + \epsilon)$. This is a population assumption, not a finite-sample calibration guarantee. Changes in $\Pr(Y \mid c)$ or incorrect extraction are outside the result. Conformal methods address complementary coverage objectives, including missing-data patterns (Zaffran et al., 2023); we do not claim that all conformal guarantees fail with missingness.

3.5 WHY SIGNED ACTIVATION AND GROUP ABLATION ARE INSUFFICIENT

A scalar encoding $S_c - R_c$ maps both unknown $(0, 0)$ and conflict $(1, 1)$ to zero. Any deterministic downstream function receiving only that scalar must give the same output for both states. Thus an application requiring different audit actions for conflict and absence cannot implement them through that representation alone. Keeping both bits or an equivalent typed encoding avoids this information loss.

A second issue concerns duplicate computational paths. Consider $f(u, z) = u$ with $u = z = c$. Ablating the organized descriptor z changes nothing, although changing the source construct and recomputing both u and z changes the output. Descriptor ablation correctly reports dependence on that path, but does not establish source-level independence. More generally, a source-linked audit must identify every descendant representation of an edited observation. These counterexamples do not invalidate ablation; they delimit its interpretation and motivate a contract defined before the unrestricted representation is used.

3.6 IMPLEMENTED TRAINING OBJECTIVE

For the prototype, let $p_\theta(y \mid \widehat{c})$ denote the semantic head and $q_\psi(y \mid x)$ the unrestricted head. Their implemented objective is

$$\mathcal{L}_{\mathrm{ECN}} = -\frac{1}{B}\sum_{i=1}^{B}\left[\log p_\theta(y_i \mid \widehat{c}_i) + \log q_\psi(y_i \mid x_i)\right]. \tag{7}$$

The completion network and semantic head are optimized jointly; eligibility and contract bounds are fixed input computations. The semantic temperature $T = \mathrm{softplus}(\rho) + .01$ is learned by this objective, not fitted afterward on test labels. Each comparator has its own cross-entropy loss and disjoint parameters; the implementation sums all five model losses in one optimizer. There is no ordinal-distance penalty, disagreement penalty or learned deferral threshold in these runs. This specifies the new prototype, not the original two-disorder training procedure.

4 EVALUATION DESIGN

We separate structural conformance, controlled predictive behavior, external language-component evaluation and the original clinical study. No newly trained model is assigned the original system's reported clinical performance.

Controlled mechanism study. The generator samples 12 binary concepts from a per-example prevalence drawn uniformly in $[.05, .95]$. A weighted sum, with weights evenly spaced from .6 to 1.4, is discretized at 2.4, 4.8, 7.2 and 9.6 into five bands. Fifteen percent of concept observations are unavailable in the source-like condition. A shortcut feature matches normalized target band with probability .95; otherwise it is uniform noise. These synthetic relationships intentionally permit predictable distribution shifts; they are not a model of clinical prevalence.

Five predictors receive the same generated batches: the contract model, an unrestricted MLP, an MLP on the compiled midpoint (same-evidence baseline), an MLP receiving support/refutation/eligibility, and a context-blind model receiving support alone. The semantic head uses nonnegative weights; the other classifiers use 64 hidden units. The learned completion has 64 hidden units. Adam uses learning rate .003 and batches of 2048. Each round trains three seeds, each on one million freshly generated examples. This is three million simulated draws per round, not three million distinct patients or semantic concept configurations.

Round 1 uses midpoint completion and source-like training. After examining its failure modes, Round 2 adds bounded learned completion and shared augmentation across the five models. Training batches sample source-like, missing, historical, conflicting and independent-shortcut conditions. Test cases use a new generator seed and 20% stronger specified corruption rates. They include reversed and independent shortcuts, missingness, historical cues, conflict and corrupted evidence. Each condition contains 10,000 cases shared across model seeds. Because the first round informed development, both rounds are exploratory rather than independent confirmation. We retain both in the released records.

Structural audit. For each Round-2 seed, 10,000 historical-context examples are perturbed only in ineligible observation values and unrestricted residuals. We measure any probability change exceeding 10^{-6}. We also test whether admissible completions leave their envelope and whether withdrawing whole construct observations narrows it. These tests verify an implementation property, not label accuracy. Exhaustive binary completions of a small five-concept fixture independently test containment.

Frozen reporting-shift replication (Round 3). After receiving an external reconstructed pilot, we froze a local protocol before this implementation's new test run; this is not a public preregistration. Five seeds (101–105) each train on one million fresh samples, adding independent $N(0, 1)$ score noise before banding. Training conditions and rates match our original code: baseline missingness .15, missing .65, historical .35, conflict .25, corruption .15. The conflict-aware contract, original completion, midpoint head, full-contract MLP $[\ell, u, b]$, typed MLP and unrestricted MLP receive identical batches; Adam, batch size and hidden width remain unchanged.

The primary event is $Y > 1$, with $\alpha = .1$; .05 is secondary. Missing-positive/missing-negative rates are $(.60, .05)$ for symptom MNAR, $(.85, 0)$ for extreme MNAR and $(.05, .60)$ for absence MNAR. We cross extreme/absence MNAR with shortcut reversal, include source-like and reversed tests, and separately test corrupted evidence. Each condition has 10,000 new cases shared across seeds. Baselines retain exactly as many cases as the contract, ranked by binary confidence, with stable tie handling. This label-free matched-count comparison is not a deployable fixed-threshold policy. Paired seed intervals describe training variability, not independent population sampling. All complete binary concept vectors are checked against the known Gaussian-noise $\eta(c)$ to audit Proposition 4's pointwise assumption.

Integrated text and real-item diagnostics. We freeze a supplied clause-scoped compiler inspired by ConText (Harkema et al., 2009) and connect its six-construct output to an ordinal fixture head. Its original 23-item suite was written with the rules; 18 additional stress cases probe fixed rules. Both are author-written diagnostics in the spirit of CheckList (Ribeiro et al., 2020), not clinical accuracy estimates. The fixture establishes executable wiring, not learned clinical prediction.

NHANES 2017–2018 supplies real questionnaire responses (CDC, 2020). We retain 5,070 of 5,533 records with valid first-eight-item scores (0–3), split 3,802/1,268 participants with seed 619, and simulate reporting masks. The target is the known eight-item sum ≥ 10, not an independent diagnosis. A 32-unit MLP trains on ten source-mask variants per training participant; all variants remain within the split. The essential comparator is exact interval scoring: omitted items lie in $[0, 3]$, so lower/upper sums certify the target without learning. This unweighted algorithm audit makes no population prevalence estimate; the file contains no interview transcripts.

5 Results

5.1 Structural conformance is distinct from predictive superiority

Across the three Round-2 seeds, the contract path and the same-evidence MLP both have zero noninterference violations on the 10,000 paired cases. The unconstrained typed MLP changes on 99.96% and the unrestricted head on 100%. These changes are allowed for a purely predictive head; they are violations only if its output is presented as depending exclusively on the admitted evidence. No tested contract completion leaves its envelope, and no whole-construct withdrawal narrows it. These observations are consistent with the structural propositions, not a test of clinical correctness.

The contract model is less affected than the unrestricted head by independent shortcut shifts, but that advantage is not unique: the same-evidence and typed controls perform better in this condition. Learned completion helps some missing/context settings relative to a midpoint-only baseline, without establishing broad superiority. Corrupted evidence directly attacks the contract's assumptions and degrades its predictions.

Envelope disagreement is also not a validated error detector. In Round 1, mean error-detection AUROC under independent shortcuts is approximately .620 for disagreement versus .657 for confidence. In Round 2 it is approximately .547 versus .723. Confidence is better across every reported condition in these runs. Consequently, the evidence supports using disagreement as an inspectable

Table 2: Round-2 synthetic macro-F1, mean ± sample standard deviation across three training seeds on shared test cases. These deviations describe training variability, not participant uncertainty. Same-evidence and typed MLPs are strong controls. These results do not establish a general accuracy advantage for the contract head.

Condition	Contract	Masked MLP	Typed MLP	Unrestricted
iid	0.753 ± 0.004	0.800 ± 0.011	0.820 ± 0.001	0.946 ± 0.005
reversed	0.754 ± 0.006	0.802 ± 0.009	0.821 ± 0.003	0.712 ± 0.027
independent	0.753 ± 0.004	0.800 ± 0.008	0.816 ± 0.001	0.692 ± 0.013
missing	0.394 ± 0.009	0.324 ± 0.026	0.420 ± 0.003	0.777 ± 0.018
history	0.592 ± 0.001	0.570 ± 0.013	0.359 ± 0.014	0.721 ± 0.032
conflict	0.654 ± 0.001	0.650 ± 0.003	0.659 ± 0.002	0.910 ± 0.013
corrupt	0.503 ± 0.002	0.490 ± 0.005	0.485 ± 0.003	0.774 ± 0.014

Table 3: Round-3 binary selective error (%), mean across five seeds at matched coverage. Full MLP receives $[\ell, u, b]$. Coverage is the contract's retained fraction (%). Corrupted evidence violates contract correctness.

Condition	Coverage	Contract	Full MLP	Typed	Raw
Source-like	51.49	0.35	0.08	0.09	0.02
Reversed	51.20	0.49	0.13	0.14	9.54
Symptom MNAR	26.85	0.82	1.42	1.28	0.00
Extreme MNAR	28.39	1.66	3.92	2.96	0.15
Absence MNAR	43.26	0.35	1.25	0.81	0.23
Extreme + reversed	29.29	1.37	4.19	3.08	11.90
Absence + reversed	43.18	0.37	1.10	0.60	27.00
Corrupted	39.28	3.03	2.67	2.59	0.05

computational discrepancy, not replacing confidence-based selection or claiming safer retained predictions.

5.2 A SELECTIVE ADVANTAGE UNDER CROSSED REPORTING SHIFT

Across the seven correct-contract regimes, the maximum mean error is 1.66% for probability certification, versus 4.19% for the full-contract MLP, 3.08% for the typed MLP and 27.00% for the unrestricted head. Under extreme MNAR plus reversal, the paired improvement over the typed MLP is 1.71 percentage points (95% seed-t interval [1.26,2.17]); over the full-contract MLP it is 2.82 [2.48,3.16]. On source-like data the contract loses to both controls. This is a bounded empirical selective advantage, not broad accuracy superiority. Confidence selection on the midpoint-head ablation has worst-regime mean error 1.83%; the original-completion head has 4.25%. Full per-seed and secondary-α records are released.

Crucially, the fitted heads violate Proposition 4's pointwise condition on 103–141 of the 4,096 binary vectors at $\alpha = .1$. Uniform slack ranges .123–.131. Low observed error therefore does not establish the nominal guarantee for these learned heads. An oracle Gaussian conditional-probability head supplies an assumption-satisfied check, separately reported in the logs. Corruption raises contract error to 3.03% and reverses its advantage over the matched MLPs; this invalidates the assumptions without demonstrating error above .1 in that particular test.

Incremental error detection depends on adaptation. For extreme MNAR plus reversal, unrestricted five-class error AUROC is .834 for confidence alone. A confidence-plus-contract detector trained on independent source-like calibration cases obtains .799, while two-fold within-regime cross-fitting obtains .912. The latter uses labeled target cases and does not establish source-to-target reliability. On source-like tests, confidence gives .878 and the source-trained combination .884. These comparisons distinguish useful target-adapted information from a deployable shift guarantee.

5.3 TEXT EXTRACTION AND REAL CONCEPTS EXPOSE DIFFERENT BOUNDARIES

The supplied text diagnostics reproduce keyword/window/clause-rule scores 7/23, 11/23 and 21/23. On the 18 frozen-rule stress cases the corresponding counts are 5/18, 8/18 and 11/18. Failures include uncertainty ("Perhaps I am depressed") and pseudo-negation ("not only depressed"); some assert incorrect eligibility or polarity rather than merely abstaining. One supplied gold label marks "now I sleep fine" unknown; we preserve that original evaluation but score the explicit denial as refutation in the stress set. This label audit prevents presenting 21/23 as external accuracy.

On the 1,268 held-out NHANES participants, exact interval scoring has zero errors in all four simulated masks, with coverage 85.65% source-like, 78.63% symptom MNAR, 76.50% extreme MNAR and 21.69% absence MNAR. Matched-coverage MLP errors are .147%, .140%, .062% and .218%, respectively (five seeds). Because the endpoint is defined by a known item sum, the exact baseline is sufficient: these results validate the arithmetic and demonstrate real-pattern abstention costs, not an advantage for neural learning or a transcript-to-diagnosis system.

5.4 ORIGINAL CLINICAL RECORD

The original clinical system's quantitative record is preserved unchanged in Section A of the separate supplement, including its 276-participant cohort and formative panel. Those results concern a separate implementation, not the contract models evaluated here; its source code and participant outputs were not supplied. No new clinical performance is inferred from that record.

6 DISCUSSION AND LIMITATIONS

What the distinction buys. A contract lets a user ask a precise question: could this semantic output change solely because an ineligible observation changed? It also exposes whether a concept value was observed or imputed and whether conflict has been collapsed into missingness. These are testable representation properties. They remain valuable questions when a masked baseline satisfies the same property or an unrestricted predictor is more accurate. The main empirical lesson is that structural conformance, prediction and confidence-based selection should be evaluated separately.

What is not guaranteed. Our containment proof is conditional on a specified binary concept box and a monotone head. It is neither causal identification nor clinical validation. Eligibility errors, false symptom extraction, cultural variation and an incomplete registry remain substantive failure modes. Full end-to-end perturbations must recompute every descendant representation; the structured tests hold extracted records fixed. Behavioral descriptors are not independently diagnostic. Neither millions of NLI pairs nor synthetic draws increase the original number of clinical participants.

Limits of the comparisons. The reporting-shift result follows an external pilot and a locally frozen replication, not independent discovery. The pointwise probability assumption is stronger than empirical calibration and fails for the fitted synthetic heads. Both text suites are handcrafted. NHANES preserves real concept correlations but uses simulated missingness and a deterministic score endpoint; no language-to-clinical-outcome link is claimed. Matched test coverage and labeled-target cross-fitting are retrospective evaluation tools. Neither certificates nor source restrictions establish diagnostic validity, fairness or clinical benefit.

7 CONCLUSION

Evidence contracts separate what a model predicts from what admitted observations support. Probability certificates yield a reporting-mechanism bound under explicit pointwise assumptions; score calibration alone is insufficient. The five-seed study finds a selective advantage under crossed reporting shifts, while exhaustive assumption checks prevent turning that empirical result into an unsupported guarantee. Text diagnostics and real-item scoring expose extraction failures and the strength of exact domain baselines. The contribution is this testable separation of structural sufficiency, statistical assumptions and measured predictive behavior.

ETHICS STATEMENT

The new experiments use public research language corpora and simulated structured data; they recruit no human participants and perform no clinical intervention. Dataset-specific licenses and source attribution apply. No E-DAIC recordings or participant-level clinical data are redistributed. The original clinical study reports institutional ethics approval and informed consent; the reconstruction does not independently verify those records or extend them to a new clinical study. The models are research artifacts, not autonomous diagnosis, triage or crisis systems. Source qualification does not itself establish clinical validity, anonymity or fairness.

REPRODUCIBILITY STATEMENT

The accompanying code contains the contract implementation, property tests, all three synthetic rounds, the BioScope component experiment, public NLI baselines and machine-readable completed results. Dataset revisions, hashes, exclusions and hyperparameters are retained. Tables for new experiments are generated from result files. Original clinical summaries are preserved separately and clearly identified; the supplied package did not include their executable implementation or participant outputs. All new training used one local Apple Silicon machine with 16 GB memory and no paid compute.

AI ASSISTANCE STATEMENT

OpenAI assisted with manuscript language drafting and LaTeX preparation. The authors are responsible for the claims, the interpretation of newly executed and source-reported results, and the full research-history disclosure.

REFERENCES

Julius Adebayo et al. Sanity Checks for Saliency Maps. NeurIPS, 2018.

Nuel D. Belnap. A Useful Four-Valued Logic. In Modern Uses of Multiple-Valued Logic, pp. 8–37, 1977.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. EMNLP, 2015.

Centers for Disease Control and Prevention. NHANES 2017–2018: Mental Health—Depression Screener (DPQ_J), data documentation and public-use data. 2020. https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2017/DataFiles/DPQ_J.htm.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. NeurIPS, 2017.

Joseph A. Goguen and Jose Meseguer. Security Policies and Security Models. In IEEE Symposium on Security and Privacy, pp. 11–20, 1982. https://doi.org/10.1109/SP.1982.10014.

Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of Biomedical Informatics, 42(5):839–851, 2009. https://doi.org/10.1016/j.jbi.2009.05.002.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing Leakage in Concept Bottleneck Models. In NeurIPS 35, 2022. https://papers.nips.cc/paper/2022/hash/944ecf65a46feb578a43abfd5cddd960-Abstract-Conference.html.

Been Kim et al. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). ICML, PMLR 80:2668–2677, 2018.

Pang Wei Koh et al. Concept Bottleneck Models. ICML, PMLR 119:5338–5348, 2020. https://proceedings.mlr.press/v119/koh20a.html.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and Pitfalls of Black-Box Concept Learning Models. arXiv:2106.13314, 2021.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL, 2019.

Tanmoy Mukherjee, Thomas Bailleux, Pierre Marquis, and Zied Bouraoui. Credal Concept Bottleneck Models for Epistemic–Aleatoric Uncertainty Decomposition. arXiv:2604.24170, 2026. Preprint.

Yixin Nie et al. Adversarial NLI: A New Benchmark for Natural Language Understanding. ACL, 2020.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In ACL, pp. 4902–4912, 2020. https://aclanthology.org/2020.acl-main.442/.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 9(Suppl 11):S9, 2008.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. NAACL, 2018.

Chih-Kuan Yeh et al. On Completeness-aware Concept-Based Explanations in Deep Neural Networks. NeurIPS, 2020.

Seungil You, David Ding, Kevin Canini, Jan Pfeifer, and Maya Gupta. Deep Lattice Networks and Partial Monotonic Functions. NeurIPS, 2017.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc Concept Bottleneck Models. ICLR, 2023. https://openreview.net/forum?id=nA5AZ8CEyow.

Renos Zabounidis et al. Benchmarking and Enhancing Disentanglement in Concept-Residual Models. arXiv:2312.00192, 2023.

Margaux Zaffran, Aymeric Dieuleveut, Julie Josse, and Yaniv Romano. Conformal Prediction with Missing Values. In ICML, PMLR 202, 2023. https://proceedings.mlr.press/v202/zaffran23a.html.