Fragility is not Reducible to Hardness: Conditional Causal Difficulty for LLMs

Anonymous authors
Paper under double-blind review

Abstract

Large language models show large, seemingly arbitrary performance drops when a task is rewritten in ways that leave its requirements untouched. We recast task difficulty as a *causal* quantity: the treatment effect of meaning-preserving perturbations on a model. A meaning-preserving rewrite can only be observed to hurt a task the model can already solve, so we place the estimand on the principal stratum where the degradation occurs: the *conditional causal difficulty* (CCD), the probability that a randomly drawn twin breaks a solvable task. Because harder tasks tend to be more phrasing-sensitive, complexity enters as an effect modifier rather than a confounder, and we residualize it out to obtain *complexity-residualized difficulty* (CRD); zero for a task exactly as fragile as its complexity predicts, positive for anomalously fragile ones. Current benchmarks do not supply the data CCD requires, so we introduce *GSM-IC-Adv* and *MATH-Adv*, augmenting GSM-IC and MATH with twins built from perturbation families, raising fragility on GSM-IC-Adv by $\approx 13\times$ over the original GSM-IC ($3.7\% \rightarrow 50.2\%$) while keeping the answer key unchanged. Our predictor (Frail-π) learns the per-perturbation break probability and reaches 0.613 ± 0.023 of the suite reliability ceiling on GSM-IC-Adv and 0.557 ± 0.023 on MATH-Adv, against 0.171/0.401 for a surface-feature baseline and 0.444/0.463 for supervised baselines such as VAP3, and doubles random recall of fragile items under a curation budget. Additionally, cross-fitted residualization leaves CRD uncorrelated with the declared proxy in practice.

1 Introduction

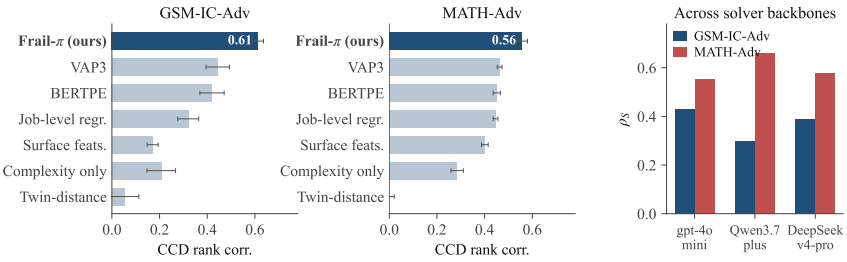


Figure 1: Frail-π predicts CCD better than every baseline on both suites (left, middle; test-fold Spearman normalized by the suite reliability ceiling, mean$\pm$sd over five seeds), and holds up when the solver backbone is changed (right; in-distribution ρ_S). Details in Tables 2 and 4.

When a large language model performs well on a benchmark, that success implicitly promises the same performance on any semantically equivalent form. This promise often fails: two prompts with identical requirements and the same gold answer, differing only by the order of two logically independent statements, can receive very different scores (Figure 2(a)); the reordering adds, removes, and renames nothing, yet the score moves anyway.

Most prior work either catalogues this brittleness descriptively (Mirzadeh et al., 2025; Zhu et al., 2024; Han et al., 2026; Zhuo et al., 2024) or predicts absolute performance of a task J, $Y(J)$, on a fixed phrasing (Chen et al., 2025; Truong et al., 2025; Razavi et al., 2025; Arabzadeh et al., 2020;

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.

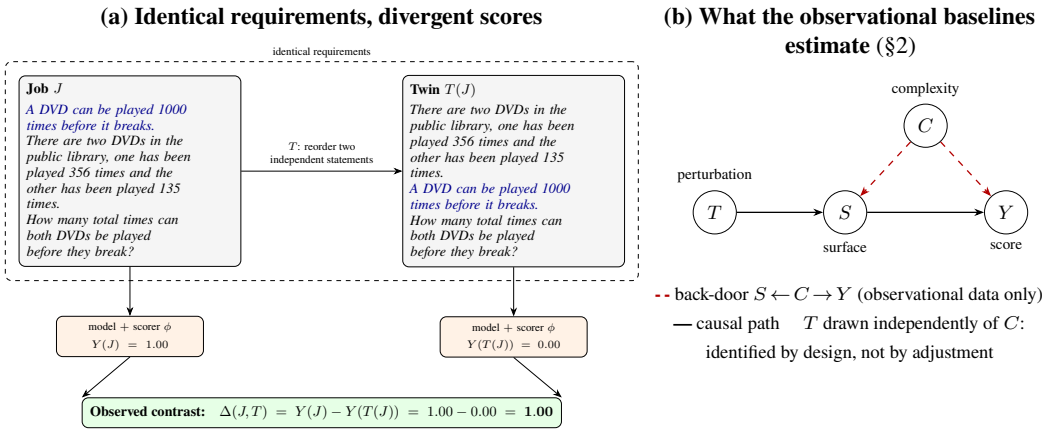


Figure 2: *(a)* A real pair from **GSM-IC-Adv** (m=1): the twin $T(J)$ differs from J only by moving the blue statement past an independent one. The model answers J correctly but this twin incorrectly, so exact-match scoring gives $Y(J) = 1.00$ and $Y(T(J)) = 0.00$, an instance of $\Delta(J,T)$ with no requirement-level basis. One pair is illustrative; unperturbed re-solves break only 1.1% of the time (Appendix D.2). *(b)* The graph shows why C confounds an *observational* comparison across prompts through the back-door $S \leftarrow C \rightarrow Y$ (dashed), so every baseline in Table 2 estimates a complexity-inflated quantity. Our contrast never opens that back-door: T is drawn independently of C, so comparing J against its own twin is identified.

Arabzadeh & Bagheri, 2025; Khodabakhsh et al., 2024); none estimates the *causal magnitude* of a rewrite, forecasts it ahead of time, or factors out the part of sensitivity driven by intrinsic complexity. We therefore treat fragility as a *causal* quantity: the effect of altering surface form while holding requirements constant. Identification here can be achieved by comparing J to its own counterpart in a within-item randomized experiment without adjustment for complexity C. The back-door shown in Figure 2(b), $S \leftarrow C \rightarrow Y$, characterizes what the *alternatives* estimate: observational $Y(J)$ predictors, and prior causal treatments of prompting, which compare across items and so must adjust for a proxy-recoverable confounder (Chen et al., 2026; Pryzant et al., 2021; Veitch et al., 2019). The causal framework instead does its work by fixing the estimand, settling *where* it is defined, and providing the one adjustment the pairing does not make for free: C is an *effect modifier* (harder tasks tend to be more phrasing-sensitive), which is why we residualize it into CRD, zero for a task exactly as fragile as the typical task at its complexity level.

A meaning-preserving rewrite can only be seen to hurt a task the model already solves. On the complementary stratum ($Y(J)$=0) every realized contrast satisfies $\Delta(J,T) \leq 0$, so a task the model cannot solve at all is recorded as maximally robust rather than untested; the stratum is censored, not informative. Averaging over both strata therefore dilutes a real effect with a structural zero, and we find empirically that the pooled average is unlearnable on real suites (see Appendix C.3). We therefore present the conditional causal difficulty (CCD): the treatment effect on the solvable stratum, equivalently the fraction of the twin family that breaks a task. Measuring CCD requires perturbations that are meaning-preserving, fragile often enough to supply signal, and not inversely tied to complexity. Existing suites fail at least one requirement (§4): GSM-Symbolic (Mirzadeh et al., 2025) moves the correct answer, and the original GSM-IC (Shi et al., 2023) leaves under 4% of problems fragile. We therefore build **GSM-IC-Adv**, whose twins draw each of m perturbations independently at random from three operator families.

Contributions. 1): We formalize CCD as a principal-stratum estimand (Frangakis & Rubin, 2002), together with its logistic dose response (§3), and show why the unconditional average τ=D_{avg} is the wrong target (Appendix C.3). **2):** We predict CCD by learning the per-perturbation break probability, which supplies ≈100 supervised observations per task (20 twins at each of five doses) instead of a job-level scalar and aggregates analytically to π, to the dose-response parameters (a_J, b_J), and to the CRD, orthogonal to every function of a *declared* complexity observable via cross-fitted residualization; this predictor outperforms all baselines on both suites and across solver backbones (Figure 1). **3):** We show the standard suites each fail one of three requirements and propose **GSM-IC-Adv** and **MATH-Adv**, which raise fragility while preserving the No-Op property.

2 Background and Problem Setup

A prompt (*task*) is a sequence of tokens $J \in \mathcal{T}$ that an LLM maps to another sequence of output tokens, the response $O(J)$, which a *quality-scoring function* ϕ collapses to a scalar

$$Y(J) \;=\; \phi\big(O(J), J\big) \;\in\; [0,1], \tag{1}$$

an approximately unbiased measurement of real output quality. A *meaning-preserving perturbation operator* (MPPO) $T : \mathcal{T} \to \mathcal{T}$ rewrites a task into a variant with the same requirements but a different surface form (a cosmetic change); we call $T(J)$ the *twin* of J and let $\mathcal{P}$ be a distribution over such operators. Three latent attributes recur below: intrinsic complexity C, requirements R, and surface form S; an MPPO acts on S and leaves R fixed by construction (Assumption 1).

Assumption 1 (Requirement invariance)**.** For every T in the support of $\mathcal{P}$ and every J: The scoring criteria of ϕ for J and $T(J)$ are identical; hence an ideal solver attains $Y(J) = Y(T(J))$, and T acts only on the surface form S; the requirements $R(J) = R(T(J))$ are left unchanged.

Under the potential-outcomes framework (Rubin, 1972; Imbens & Rubin, 2015), each task has two scores, $Y(J)$ (control) and $Y(T(J))$ (treatment), with per-instance effect and conditional average treatment effect (CATE)

$$\Delta(J,T) \;=\; Y(J) - Y\big(T(J)\big), \qquad \tau(J) \;=\; \mathbb{E}_{T\sim\mathcal{P}}\big[\,\Delta(J,T) \mid J\,\big], \tag{2}$$

valid under Assumption 1 (the task-perturbation analogue of the prompt CATE of Chen et al. (2026)).

T is drawn independently of the latent attributes of J, and $\Delta(J,T)$ compares the *same* task under two randomly assigned surface realizations: a randomized experiment within the item, whose contrast is unbiased without any adjustment. The back-door $S \leftarrow C \rightarrow Y$ (Pearl, 2009) enters instead as a description of what the *alternatives* estimate: across two naturally occurring prompts $s \neq s'$, C shapes both description and performance, so $P(Y \mid S{=}s)$ (the object every $Y(J)$ predictor in §6 regresses) is complexity-inflated. Given so permissive a design, the causal framework earns its place downstream, in three steps the pairing does *not* give for free: it fixes the *estimand*, to be predicted ex-ante rather than measured post-hoc (whence $Y(J)$ ranks our target *negatively*, Table 2); it forces the question of *where* that estimand is defined, a stratum problem (Frangakis & Rubin, 2002) rather than a confounding one (§3); and it demands one real adjustment, for effect modification by C (§3.1).

3 Conditional causal difficulty (CCD)

IRT-style difficulty assigns a task a static b_J from average-model performance (Chen et al., 2025), conflating hardness with sensitivity to phrasing. Causal difficulty separates them: zero whenever J and its twin are equally hard, non-zero only when phrasing itself moves the score.

Definition 1 (Unconditional causal difficulty (U-CCD))**.** Fix a finite *operator set* $\mathcal{O} \subset \mathrm{supp}(\mathcal{P})$ with $|\mathcal{O}| = G$, the G twins sampled for a task. The U-CCD of a task J is

$$D_{\mathrm{avg}}(J) = \mathbb{E}_{T\sim\mathcal{P}}\big[\Delta(J,T)\big] = \tau(J), \tag{3}$$

the signed expected degradation of a typical innocuous edit, estimated from the G sampled twins.

Definition 1 averages over all tasks, including those the model fails outright, where $Y(J){=}0$ forces $\Delta(J,T) \leq 0$ for every T (and likewise for the worst case $D_{\mathrm{wc}}(J) = \max_T \Delta(J,T)$): the population average mixes a live effect with a censored one. Appendix C.3 confirms the consequence — scored against U-CCD, every method lands within a seed's worth of noise of the others. We therefore restrict the estimand to the *solvable stratum* ($Y(J){=}1$), using the per-twin *break event* $B(J,T) \;=\; \mathbf{1}\{Y(J){=}1,\; Y(T(J)){=}0\}$: a twin that breaks a task the model solved in that same measurement. On this stratum $\Delta(J,T) = B(J,T)$ exactly, so the conditional fragility below is not a substitute quantity but the treatment effect of equation 2 evaluated where it is non-degenerate.

Definition 2 (Conditional causal difficulty and dose response)**.** For a task J with $Y(J){=}1$ and a family indexed by perturbation *dose* m (here, the total number of perturbation operations used to build a twin),

$$\pi(J) = \mathbb{E}_{T\sim\mathcal{P}}\big[B(J,T) \mid Y(J){=}1\big], \qquad \mathrm{logit}\,\Pr\big[B(J,T){=}1 \mid m\big] = a_J + b_J\,(m - \bar{m}). \tag{4}$$

$\pi(J)$ is the *conditional causal difficulty* (CCD) (A.K.A., the conditional fragility): the fraction of the family that breaks J. a_J is fragility at typical dose and b_J is the *degradation rate*, how fast fragility grows with perturbation intensity.

Conditioning an effect on a potential outcome is the setting of *principal stratification* (Frangakis & Rubin, 2002), which names CCD more precisely than a "conditional CATE" does. Taking the outcome as the stratifying variable, a (task, twin) pair falls in one of four basic strata indexed by $\big(Y(J), Y(T(J))\big)$ (always-solved, breaks, repaired, never-solved) and $\Delta{=}B$ on $\{Y(J){=}1\}$ says that $\pi(J) = \Pr[(1,0) \mid Y(J){=}1]$, the share of the *solvable* union (always-solved plus breaks) that breaks. Our conditioning set is that union, defined by the control potential outcome alone, not a single stratum defined by the joint pair. Two features of the paired design make this the tractable case. Stratum membership is *observed*: canonically only one arm is realized per unit, so strata are latent and principal effects need monotonicity or partial-identification bounds(Zhang & Rubin, 2003), whereas here base and twins are all run, $Y(J)$ is measured for every item, and π is estimated directly. And the conditioning opens no selection path: T is drawn independently of every attribute of J, $Y(J)$ included, so restricting to $\{Y(J){=}1\}$ cannot induce collider bias. What the restriction does cost (that $D_{\text{avg}} \leq \pi$ always, that the partition is solver-relative, and that the stratum label is measured rather than given) is stated in §7. For bookkeeping, we write $\mu_J(s) = \mathbb{E}[Y \mid J, \text{do}(S{=}s)]$ for task J's response under a surface realization s, and $s' = S(T(J))$; the within-item contrast of §2 has conditional mean

$$\mathbb{E}\big[\Delta(J,T) \mid J,\, S{=}s,\, S'{=}s'\big] = \mu_J(s) - \mu_J(s'), \tag{5}$$

with J held fixed by the pairing, so no attribute of J (C included) need be observed or assumed unchanged by T.

3.1 Complexity-residualized difficulty (CRD)

Decomposing $\mu_J(s) = \alpha(C(J)) + \beta(C(J), s)$ in equation 5, the paired contrast equals $\beta(C(J), s) - \beta(C(J), s')$, which still depends on $C(J)$ unless β is additively separable: under effect modification the twin returns a CATE at the task's own complexity, and the same holds verbatim for π. Partialling C out is therefore not an optional refinement but the mechanism that separates anomalous fragility from intrinsic complexity, which makes Definition 3 an important object.

Definition 3 (Complexity-residualized difficulty (CRD))**.** Fix a base quantity $Q \in \{\tau, \pi\}$, a *declared* complexity observable C, and a task population p (restricted to the $Y(J){=}1$, when $Q = \pi$, since π is only defined there). With $m_{Q,C,p}(c) = \mathbb{E}_{J' \sim p}[Q(J') \mid C(J') = c]$ the mean of Q at c,

$$D^{\perp}_{Q,C,p}(J) \;=\; Q(J) - m_{Q,C,p}\big(C(J)\big). \tag{6}$$

Our primary instantiation is $Q = \pi$, the estimand of Definition 2; residualizing it is the version of C2 (§5.2) that isolates anomalous fragility.

CRD, $D^{\perp}$, is zero for a task exactly as fragile as the typical same-C task and positive for *anomalously* fragile tasks. Whereas Q itself inherits the complexity of the task, CRD is the part the declared C cannot account for.

Proposition 1 (Residualization by projection)**.** ***(i)*** *CRD is the L_2 projection residual of Q onto measurable functions of the* declared *C; hence $\mathbb{E}[D^{\perp}] = 0$ and $\mathbb{E}[D^{\perp}\,\eta(C)] = 0$ for every $\eta \in L_2(\sigma(C))$, so CRD is uncorrelated with* every *function of C, for either $Q = \tau$ or $Q = \pi$. Orthogonality to latent intrinsic complexity is* not *implied, and holds only insofar as C is a sufficient summary of it.* ***(ii)*** *CRD is estimated by cross-fitted residualization: regress the measured $\hat{Q}$ on a complexity proxy out-of-fold to obtain $\hat{m}$ and set $\hat{D}^{\perp} = \hat{Q} - \hat{m}$, consistent whenever $\hat{Q} \to Q$ and the complexity regression is consistent (Chernozhukov et al., 2018; Chen et al., 2026) (Proof in Appendix A).*

3.2 A complexity-residualized predictor

Computing the estimand exactly requires running the model on J and every twin from $\mathcal{P}$, which might be prohibitive at scale. We instead learn a predictor that takes the text of J and is *causally aligned*: it must estimate the effect of a surface change, not learn that verbose descriptions score low, so we make the *contrast* the architectural primitive. A single shared encoder f_θ embeds J and $T(J)$ into $\boldsymbol{h} = f_\theta(J)$ and $\boldsymbol{h}' = f_\theta(T(J))$; their difference $\boldsymbol{h} - \boldsymbol{h}'$ feeds a contrast head ψ_ρ, while a single-tower head g_ω predicts difficulty when no twin exists at acquisition time:

$$\widehat{\Delta}(J,T) = \psi_\rho\big(\boldsymbol{h} - \boldsymbol{h}'\big), \qquad \hat{D}(J) = g_\omega\big(\boldsymbol{h}\big). \tag{7}$$

Table 1: Where each real-math perturbation suite fails. *Fragile*: $d_{\text{obs}} > 0.05$; *Zero*: $|d_{\text{obs}}| < 0.001$. **R1** No-Op (meaning preserving); **R2** enough fragile problems to rank; **R3** no inverse complexity–fragility confound. *MATH (original)* has no perturbation suite, so Fragile/Zero are n/a. For GSM-Plus answer preservation verified empirically (zero drift across full test set). GSM-IC-Adv and MATH-Adv numbers are for the audited m=9 corpus (Appendix D.4).

Dataset	Problems	Twins	Fragile	Zero	R1	R2	R3
GSM-IC (original)	1,319	20	3.7%	90.1%	✓	×	✓
GSM-Plus (MPPO) (Li et al., 2024)	1,319	2	10.5%	86.2%	✓	×	✓
GSM-Symbolic p1	100	10	25.0%	56.0%	×	✓	n/a
GSM-DC (Yang et al., 2025)	6,222	1	16.4%	64.0%	✓	✓	×
GSM-IC-Adv (ours)	**1,318**	**avg. 16.3**	**50.2%**	**34.2%**	✓	✓	✓
MATH (original) (Hendrycks et al., 2021)	5,000	1	n/a	n/a	n/a	×	✓
MATH-Adv (ours)	4,994	avg. 17.4	33.7%	39.9%	✓	✓	✓

The residualized prediction subtracts a cross-fitted complexity regression from the contrast output, $\widehat{D}^{\perp}(J) = \widehat{D}(J) - \hat{m}_{\text{pred}}(C(J))$ (Prop. 1(ii)); this shares the additive term $-\hat{m}(C(J))$ with the target, and §5.2 confirms the resulting prediction is near-orthogonal to C in practice, not just the measured target (Table 3), holding across suites and across generator models. Our deployed instantiation is the $Q{=}\pi$ predictor FRAIL-π (§5.1), which applies this recipe to the per-observation break event $B(J,T)$ and so receives $\approx$100 labelled contrasts per task (20 twins at each of the five doses $m \in \{1,3,5,7,9\}$, before audit filtering). The same recipe applied to the unconditional τ (FRAIL-τ) is retained for Synth-CD, where scores are continuous, no stratum is censored, and τ is therefore well posed (§5.2, Appendices B.1 and C.3).

Training objective. We *amortize* the inner expectation of equation 2 so difficulty for an unseen task costs one forward pass, minimizing $\mathbb{E}_{J\sim p_{\mathcal{D}}}[\mathcal{L}(\theta,\rho,\omega;J)]$ with per-task term

$$\mathcal{L}(\cdot;J_i) = \sum_{T\in\mathcal{S}_i} \ell_\delta\big(\widehat{\Delta}(J_i,T), \Delta(J_i,T)\big) + \ell_\delta\big(g_\omega(\boldsymbol{h}_i),\ \widehat{D}^{\,\text{obs}}(J_i)\big) + \lambda\,\mathcal{R}_{\text{inv}}(\theta;J_i), \tag{8}$$

Here, $\mathcal{S}_i$ is the collected twins of J_i, and ℓ_δ is the Huber loss: the first (contrast) term supplies the primary causal supervision and the second term distils it into the deployable single-tower head. The encoder decomposes $\boldsymbol{h} = c(C(J)) + b(R(J),S(J))$ into a complexity part $c(\cdot)$ and a surface part $b(\cdot)$; since T leaves C,R fixed,

$$\boldsymbol{h} - \boldsymbol{h}' \;=\; b\big(R(J),S(J)\big) - b\big(R(J),S(T(J))\big). \tag{9}$$

Since $c(C(J))$ cancels, giving a contrast that depends on S alone, we can encourage the decomposition with an invariance regularizer, $\mathcal{R}_{\text{inv}}(\theta) = \mathbb{E}_{J,T\sim\mathcal{P}}\|\Pi_C\boldsymbol{h} - \Pi_C\boldsymbol{h}'\|^2$, with Π_C a learned complexity-subspace projection.

4 BENCHMARKING CAUSAL DIFFICULTY

To estimate CCD, a benchmark must meet three conditions: (**R1**, No-Op) the perturbation satisfies Assumption 1, so a score drop reflects fragility and not a harder problem; (**R2**, signal) enough problems in the solvable stratum are actually fragile to support a ranking metric; and (**R3**, no inverse confound) fragility is not driven *down* by intrinsic complexity, which would leave the target nearly unlearnable. Table 1 scores widely used public suites against these criteria.

Existing suites fail at least one requirement. *GSM-Symbolic* (Mirzadeh et al., 2025) moves the correct answer, violating **R1**. *GSM-DC* (Yang et al., 2025) passes R1/R2 but fails **R3**: its stripped base leaves easy items (already solved on the natural full twin) zero-fragility while hard items hit a floor on the twin, an inverse complexity–fragility confound. The original *GSM-IC* (Shi et al., 2023) and *MATH* (Hendrycks et al., 2021) both fail **R2**: GSM-IC's inert filler is trivially ignored (3.7% fragile), and MATH ships with no perturbation suite at all. *GSM-Plus* (Li et al., 2024) passes **R1/R3** on its two MPPOs (problem understanding, distraction insertion) but fails **R2**: fragility is 10.5%, too thin for reliable per-instance ranking. We include it in Table 1 to complete the survey.

We therefore introduce **GSM-IC-Adv**, whose twins combine three operator families: entity renaming, sentence reordering, and injecting a *math-domain* distractor sentence from a *different* GSM8K problem; each dose m applies m such operations, drawn independently at random across the three families, so a given twin's fragility reflects the joint effect of whichever operators it happens to draw rather than any single family in isolation. The contamination audit (Appendix D.4) removes twins whose injected distractor sentence accidentally reproduces the target problem's answer key or operands (the dominant leakage mode we screened for); this criterion brings GSM-IC-Adv from 20 nominal twins/problem down to $21{,}534$ retained pairs over its $1{,}318$ problems (avg. 16.3 twins/problem). **MATH-Adv** applies the same three-family protocol and the same audit criterion to Hendrycks MATH (Hendrycks et al., 2021), additionally carrying each item's problem-setter-assigned level as a complexity proxy that is external to our pipeline (set independently of any operator draw or model output, hence non-circular), and retains $87{,}067$ twin pairs over $4{,}994$ problems (avg. 17.4 twins/problem).

5 Experiments

We evaluate our method on Synth-CD (known causal ground truth) and the real math data GSM-IC-Adv/MATH-Adv. **C1** (§5.1): the CCD is predictable from task text ahead of time, while the U-CCD on the same suite is not (Appendix C.3). **C2** (§5.2): the CRD is orthogonal to the declared complexity observable, and Appendix D.3 shows the ranking is actionable under a curation budget. Baselines: surface features, a declared-complexity score, frozen twin-distance, three PromptSET measures (Razavi et al., 2025; Arabzadeh et al., 2020), measured base correctness $Y(J)$, BERTPE (Khodabakhsh et al., 2024), and VAP3 (Arabzadeh & Bagheri, 2025). Solvers decode at temperature 0.2 with 5 samples per prompt, and Y scores the modal answer, so a twin could break a task by re-sampling alone. An identity-twin control (Appendix D.2) measures this floor: re-solving unperturbed stratum bases breaks 1.1% (GSM-IC-Adv) and 2.5% (MATH-Adv) of re-solves, and the modal base answer is reproduced on the full datasets, so stratum membership is stable. Backbone specs and the training recipe are in Appendix C; the datasets in Appendix D.1; U-CCD results on Synth-CD and GSM-IC-Adv in Appendices B.1 and C.3.

5.1 C1: CCD is predictable from text

Our main model, Frail-π, fine-tunes a `bert-base-uncased` encoder end to end with the break-probability head, evaluated against the baselines with 5 different seeds. Table 2 reports CCD and the dose-response parameters, with every Spearman normalized by that suite's reliability ceiling (0.969 and 0.972; Appendix D.2) so 1.0 is the attainable maximum. Frail-π reaches $0.613{\pm}0.023$ of ceiling on GSM-IC-Adv ($\rho_S{=}0.594$) and $0.557{\pm}0.023$ on MATH-Adv ($\rho_S{=}0.541$), well above the surface-feature and declared-complexity-score baselines. Representation choice (TF-IDF+SVD, frozen sentence/OpenAI encoders) is ablated in Appendix C.2: fully fine-tuning the encoder beats every frozen representation on both suites.

Same features, job-level regression (frozen encoder only, since the fine-tuned encoder has no representation independent of its own objective) trails Frail-π on both suites, so neither the per-observation reformulation nor fine-tuning alone explains the full lift; each contributes. Four further training-recipe levers do not close the gap either (Appendix C.2), so the gap remains a property of the observable, not of model capacity or training recipe.

Scored against the U-CCD on the same GSM-IC-Adv holdout, every method, ours included, is weak and within noise of the others (best $\rho_S{=}0.092$ vs. Frail-π's 0.594; Appendix C.3): restricting to the solvable stratum (§3) is what turns a censored, diluted average into a rankable target.

5.2 C2: CRD is recoverable and orthogonal

On the real suites the primary C2 result uses $Q = \pi$ (Definition 2), residualized via Frail-π against C. Table 3 shows orthogonality holds on both suites for the measured target itself ($|\text{Corr(target},C)|{\leq}0.06$, as Prop. 1(i) predicts for a consistent complexity regression) and for the model's prediction (Corr(pred,C) similarly near zero), so residualization is not merely correcting a large raw imbalance – it is confirmed to hold at the level Prop. 1 requires.

Table 2: Test-fold Spearman *normalized by the suite reliability ceiling* (0.969 for GSM-IC-Adv, 0.972 for MATH-Adv; Appendix D.2), mean±sd over five seeds. π: CCD predicted by the single-tower head g_ω; *a/b*: dose-response intercept/slope (Definition 2). *Job-level regression*: same frozen representation, regressed directly, isolating FRAIL-π's per-observation objective. Frozen twin-distance / Prompt Clarity / Embedding Entropy / Embedding Specificity-IEF (from PromptSET, Razavi et al. 2025; Arabzadeh et al. 2020) are retargeted to π with a TF-IDF+SVD embedding on the same seeded splits (Appendix C.1); they have no per-dose grid, so *a/b* are n/a.

Suite	Method	$\pi\uparrow$	$a\uparrow$	$b\uparrow$
GSM-IC-Adv	**FRAIL-π**	**0.613±0.023**	**0.611±0.020**	**0.289±0.021**
	VAP3	0.444±0.049	0.441±0.046	0.245±0.031
	BERTPE	0.420±0.051	0.420±0.049	0.255±0.035
	Same features (frozen encoder), job-level regr.	0.320±0.044	0.320±0.049	0.184±0.019
	Surface features (10, GBT)	0.171±0.023	0.164±0.026	0.146±0.056
	Declared complexity score only	0.207±0.060	0.192±0.058	0.181±0.049
	Frozen twin-distance	0.054±0.059	n/a	n/a
	Prompt Clarity (PromptSET)	−0.060±0.035	n/a	n/a
	Embedding Entropy (PromptSET)	0.023±0.026	n/a	n/a
	Embedding Specificity-IEF (PromptSET)	0.034±0.033	n/a	n/a
	Measured base correctness $Y(J)$	−0.124±0.037	−0.130±0.032	0.011±0.017
MATH-Adv	**FRAIL-π**	**0.557±0.023**	**0.548±0.022**	**0.492±0.015**
	VAP3	0.463±0.010	0.454±0.010	0.172±0.026
	BERTPE	0.451±0.015	0.441±0.017	0.272±0.053
	Same features (frozen encoder), job-level regr.	0.446±0.010	0.431±0.008	0.182±0.021
	Surface features (10, GBT)	0.401±0.014	0.389±0.015	0.180±0.025
	Declared complexity score only	0.285±0.026	0.299±0.028	0.242±0.044
	Frozen twin-distance	−0.003±0.024	n/a	n/a
	Prompt Clarity (PromptSET)	0.003±0.027	n/a	n/a
	Embedding Entropy (PromptSET)	−0.160±0.032	n/a	n/a
	Embedding Specificity-IEF (PromptSET)	0.020±0.025	n/a	n/a
	Measured base correctness $Y(J)$	−0.285±0.040	−0.280±0.042	0.057±0.036

Synth-CD has continuous scores and no censored stratum, so τ is well posed there and the suite isolates a question the real suites cannot: whether the contrast architecture is *intrinsically* complexity-orthogonal, independent of any explicit residualization. Both analytic confounders are calibrated to be strong effect modifiers ($C \perp E$ by construction), so there is a large complexity-explained component to remove; the FRAIL-τ contrast head reaches ρ_S=0.617 at $|\rho_C|$=0.028 (Appendix B.1), roughly twentyfold below the raw D_{avg}-level $|\rho_C|\approx$0.60, and gets there *without* explicit residualization, since differencing exposes only the C-independent expertise signal E (equation 9); cf. Table 3.

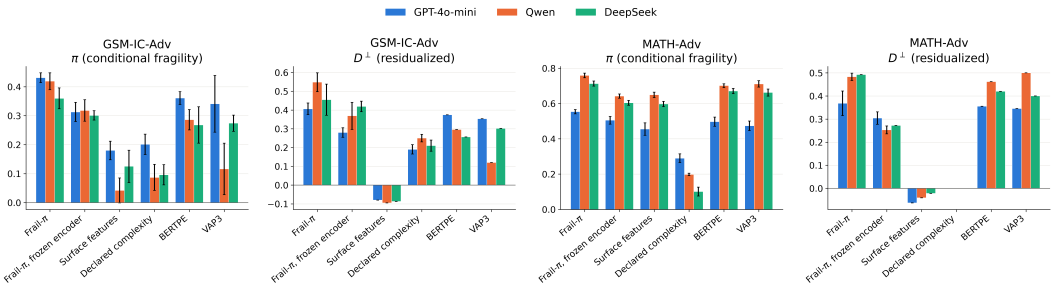


Figure 3: Cross-model generalization, single-dose (m=9) twins, mean±sd over five seeds. The *FRAIL-π* bars are the cross-generator run whose orthogonality diagnostics appear in Table 3; shown alongside two ablations not in that table: a frozen-encoder representation (*FRAIL-π, frozen encoder*) and per-generator retrained BERTPE/VAP3. Other baselines follow Table 2's definitions. This is a lower-data regime than the 5-dose run behind Table 2, so π here is not directly comparable to that table's numbers.

Generalization across generator models. To check whether FRAIL-π's ranking ability is specific to that generator, we reran the fine-tuned baseline on single-dose (m=9) twins generated from the same

base problems by Qwen and DeepSeek, scoring GPT on the matching single-dose slice for a fair comparison (CCD ranking in Figure 3; orthogonality diagnostics in Table 3). CCD replicates at a comparable or higher level for Qwen and for DeepSeek; the residualized CRD stays orthogonal to C on the Qwen slices of both suites and on the DeepSeek GSM-IC-Adv slice, so the lift is not caused by the generator. **Solver transfer.** The checks above fix the solver (`gpt-4o-mini`); since the stratum

Table 3: CRD orthogonality against C; the primary instantiation of Definition 3. The declared proxy C is suite-specific: the problem-setter-assigned competition level for MATH-Adv (external, non-circular; §4), and the number of reasoning steps in each problem's GSM8K gold solution for GSM-IC-Adv, which ships no such label.

			$\mathrm{Corr}(\cdot, C)$		$\mathrm{Corr}(\cdot, C^2)$
Suite / backbone	MAE↓	$\lvert\rho_C\rvert$ ↓	target	pred.	pred.
GSM-IC-Adv, FRAIL-π(hf_finetune)	0.129	0.044	−0.043	−0.061	−0.068
MATH-Adv, FRAIL-π(hf_finetune)	0.129	0.030	0.046	−0.023	−0.022
GSM-IC-Adv, Qwen twins (hf_finetune)	0.074	0.092	−0.058	−0.055	−0.069
MATH-Adv, Qwen twins (hf_finetune)	0.117	0.067	−0.018	−0.016	−0.012
GSM-IC-Adv, DeepSeek twins (hf_finetune)	0.058	0.087	−0.025	−0.060	−0.056
MATH-Adv, DeepSeek twins (hf_finetune)	0.103	0.053	−0.055	−0.040	−0.018

is defined by the solver's own $Y(J)$, varying the *solver* directly tests how far the estimand travels: on the same GSM-IC-Adv and MATH-Adv items and twins (m=9), we independently measure FRAIL-π's target – each solver's own CCD – for `gpt-4o-mini`, Qwen3.7-plus, and DeepSeek-v4-pro, train the fine-tuned on one solver's measured π, and evaluate the held-out ranking against a *different* solver's measured π on the same holdout problems (Table 4). Off-diagonal ρ_S is positive throughout, and several cells are comparable to one of the two solvers' own in-distribution accuracy (e.g. on GSM-IC-Adv, DeepSeek-v4-pro-trained transfer *into* Qwen3.7-plus reaches 0.306, above Qwen3.7-plus's own 0.299): a predictor trained on one solver's fragility labels ranks a different solver's fragility well above chance without ever seeing that solver's labels. Transfer is not uniformly close to in-distribution accuracy, so we read this as partial, solver-dependent transfer rather than solver-invariance; as the solver-relativity of the stratification (§3) predicts. CCD is thus the right target for transfer: CRD's orthogonality to C is a within-solver deconfounding device, and porting it across solvers would lower ρ_S by ≈0.12 on GSM-IC-Adv and ≈0.08 on MATH-Adv with no compensating gain.

Table 4: **Solver transfer.** Held-out ρ_S of the fine-tuned FRAIL-π trained on the row solver's measured π and evaluated against the column solver's measured π, same items/twins throughout (m=9). Diagonal = in-distribution; off-diagonal = cross-solver transfer.

GSM-IC-Adv train \ eval	gpt-4o-mini	Qwen3.7-plus	DeepSeek-v4-pro
gpt-4o-mini	0.431	0.355	0.411
Qwen3.7-plus	0.297	0.299	0.401
DeepSeek-v4-pro	0.237	0.306	0.389
MATH-Adv train \ eval	gpt-4o-mini	Qwen3.7-plus	DeepSeek-v4-pro
gpt-4o-mini	0.554	0.180	0.295
Qwen3.7-plus	0.205	0.660	0.468
DeepSeek-v4-pro	0.270	0.456	0.576

6 RELATED WORK

Non-causal views. Mirzadeh et al. (2025); Zhu et al. (2024) catalogue task fragility descriptively; Han et al. (2026); Zhuo et al. (2024) localize failures post-hoc or via decoding confidence. IRT-style difficulty (Chen et al., 2025; Li et al., 2025; Ding et al., 2024; Truong et al., 2025; Scarlatos et al., 2025) assigns a static score b_J from a single phrasing — orthogonal to CCD: a task can be

hard yet robust or easy yet fragile. Ours is a causal contrast on the solvable stratum, refined into the proxy-orthogonal CRD. **Per-instance prompt performance prediction.** Razavi et al. (2025) predict per-prompt correctness via variation-interaction and embedding-specificity signals (Arabzadeh et al., 2020); Arabzadeh & Bagheri (2025) add VAP3 and BERTPE (Khodabakhsh et al., 2024). We retarget five of these – Prompt Clarity, Embedding Entropy, Embedding Specificity-IEF, VAP3, and BERTPE – to CCD (Definition 2, Table 2; adaptation in Appendix C.1). *How we differ:* all five predict $Y(J)$ on a fixed phrasing, not the contrast on the solvable stratum, so perfect $Y(J)$ prediction carries no information about which solvable problems break; indeed $Y(J)$ itself ranks π *negatively* on both suites (Table 2). None of them measures or removes $Y(J)$'s C-confound; our CRD does. VAP3, Prompt Clarity, and Embedding Entropy also use held-out variation text at test time (an oracle condition); our ex-ante head uses the base text alone and still beats all three (Table 2).

Causal views. Zhang et al. (2022) frame perturbation robustness causally; CPO estimates prompt conditional average treatment (CATE) by double machine learning (Chen et al., 2026; Chernozhukov et al., 2018); other text-causal work adjusts for text-recoverable confounding (Pryzant et al., 2021; Veitch et al., 2019; Yang et al., 2026; Guo et al., 2025; Imai & Nakamura, 2026; Yang et al., 2021); Andreu et al. (2024) identify causal latents via a contrastive structured-treatment representation. All compare *across* items and must recover a confounder from text; we perturb the *task*, so identification is a property of the design, not of an adjustment (equation 5), and we claim no contribution for it. *How we differ* lies downstream: C enters as an effect *modifier*, so CRD (the object of study, not a selection criterion) still demands a real adjustment. Closest in kind is **principal stratification** (Frangakis & Rubin, 2002): our estimand is a principal-stratum quantity (§3) rather than a population average pooling a stratum where the effect is censored. That literature's defining difficulty (latent stratum membership, handled by monotonicity or partial identification (Zhang & Rubin, 2003)) does not arise here, since the paired design realizes both arms on every item; what carries over is its interpretive discipline, that an effect defined on a stratum is a statement about that stratum and not about the population as a whole (§7).

7 Conclusion

We recast task difficulty for LLMs as a causal measure on the principal stratum where a meaning-preserving rewrite can be observed to bite: the CCD, the effect of such a rewrite on tasks a fixed model already solves, predicted per instance and ahead of time. FRAIL-π learns the per-perturbation break probability and aggregates it to CCD, the dose response, and CRD; it beats every baseline and frozen representation we test and roughly doubles random recall of fragile items under a curation budget, while the same sweep scored against the U-CCD collapses to noise. The causal framing fixes the estimand, its stratum, and the residualization separating superficial complexity from model-specific weakness, supporting difficulty-aware dataset design and targeted robustness assessment.

Limitations. The principal stratum (§3) is solver-relative, and cross-solver transfer, though positive, falls below in-distribution accuracy in several cells (Table 4), therefore is not exactly solver-invariant. The stratum label is a measured $Y(J)$, so π carries a re-sampling floor, measured with identity twins (1.1% and 2.5% of re-solves; Appendix D.2), and since $D_{\text{avg}} \leq \pi$, π describes solvable tasks, not the whole population. CRD is orthogonal only to the *declared* proxy; a different C would yield a different CRD (Prop. 1). The No-Op audit (Appendix D.4) is a best-effort filter, not a proof of zero contamination, though a second LLM adjudicator over all $\approx$10K flagged-but-kept MATH-Adv pairs and human review of a $1{,}000$-pair sample reproduced its verdicts.

Use of Large Language Models

We used generative AI for three purposes. **Dataset construction (required).** The twins defining GSM-IC-Adv and MATH-Adv are LLM-generated: an LLM added each meaning-preserving perturbation (§4, Appendix D), and the cross-generator ablations vary which model does so. **LLM-as-judge for data validation.** The No-Op audit filtering those twins uses an LLM judge (`gpt-5`) to decide whether a twin changes or invalidates the gold answer, with a second, independently queried adjudicator (Claude) for the cross-checks in Appendix D.4; verdicts are combined with deterministic structural collision checks rather than trusted alone. **Writing.** We used LLM tools to edit and polish wording and formatting of text we wrote. LLMs were not used to generate ideas or hypothe-

ses, develop the theory, formulate or prove the mathematical claims, design the methodology or experiments, or interpret results; the remaining disclosure categories do not apply.

We have reviewed all AI-assisted work: every generated pair passes the automated filter, every unit-collision-flagged pair was additionally re-adjudicated by a second LLM judge over the full ≈10K-pair pool, a 1,000-pair stratified sample was verified by human review against the base text, twin text, and gold answer (agreeing with the LLM verdicts on every pair), and all LLM-edited manuscript text was read and approved by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Oriol Corcoll Andreu, Athanasios Vlontzos, Michael O'Riordan, and Ciaran M. Gilligan-Lee. Contrastive representations of high-dimensional, structured treatments, 2024. URL https://arxiv.org/abs/2411.19245.

Negar Arabzadeh and Ebrahim Bagheri. Vap3: Variation-aware prompt performance prediction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 2794–2799, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3730264. URL https://doi.org/10.1145/3726302.3730264.

Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, and Ebrahim Bagheri. Neural embedding-based metrics for pre-retrieval query performance prediction. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, pp. 78–85, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-45441-8. doi: 10.1007/978-3-030-45442-5_10. URL https://doi.org/10.1007/978-3-030-45442-5_10.

Jianhao Chen, Chenxu Wang, Gengrui Zhang, Peng Ye, Lei Bai, Wei Hu, Yuzhong Qu, and Shuyue Hu. Learning compact representations of llm abilities via item response theory. *arXiv preprint arXiv:2510.00844*, 2025.

Wei Chen, Yanbin Fang, Shuran Fu, Fasheng Xu, and Xuan Wei. Optimizing prompts for large language models: A causal approach, 2026. URL https://arxiv.org/abs/2602.01711.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 02 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Mucong Ding, Chenghao Deng, Jocelyn Choo, Zichu Wu, Aakriti Agrawal, Avi Schwarzschild, Tianyi Zhou, Tom Goldstein, John Langford, Anima Anandkumar, and Furong Huang. Easy2hard-bench: standardized difficulty labels for profiling llm performance and generalization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002. doi: https://doi.org/10.1111/j.0006-341X.2002.00021.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2002.00021.x.

Siyi Guo, Myrl G Marmarelis, Fred Morstatter, and Kristina Lerman. Estimating causal effects of text interventions leveraging LLMs. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 170–190, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-303-6. doi: 10.18653/v1/2025.findings-ijcnlp.10. URL https://aclanthology.org/2025.findings-ijcnlp.10/.

Shou-Tzu Han, Rodrigue Rizk, and KC Santosh. Fragile reasoning: A mechanistic analysis of llm sensitivity to meaning-preserving perturbations. *arXiv preprint arXiv:2604.01639*, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *CoRR*, abs/2103.03874, 2021. URL https://arxiv.org/abs/2103.03874.

Kosuke Imai and Kentaro Nakamura. Causal inference with generative artificial intelligence: Application to texts as treatments. *Journal of the American Statistical Association*, 0(0):1–14, 2026. doi: 10.1080/01621459.2026.2689629. URL https://doi.org/10.1080/01621459.2026.2689629.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Maryam Khodabakhsh, Fattane Zarrinkalam, and Negar Arabzadeh. Bertpe: A bert-based pre-retrieval estimator for query performance prediction. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part III*, pp. 354–363, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-56062-0. doi: 10.1007/978-3-031-56063-7_27. URL https://doi.org/10.1007/978-3-031-56063-7_27.

Peiyu Li, Xiuxiu Tang, Si Chen, Ying Cheng, Ronald Metoyer, Ting Hua, and Nitesh V Chawla. Adaptive testing for llm evaluation: A psychometric alternative to static benchmarks. *arXiv preprint arXiv:2511.04689*, 2025.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2961–2984, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.163. URL https://aclanthology.org/2024.acl-long.163/.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. Causal effects of linguistic properties. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4095–4109, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.323. URL https://aclanthology.org/2021.naacl-main.323/.

Amirhossein Razavi, Mina Soltangheis, Negar Arabzadeh, Sara Salamat, Morteza Zihayat, and Ebrahim Bagheri. Benchmarking prompt sensitivity in large language models. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part III*, pp. 303–313, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-031-88713-0. doi: 10.1007/978-3-031-88714-7_29. URL https://doi.org/10.1007/978-3-031-88714-7_29.

Donald Rubin. Estimating causal effects of treatments in experimental and observational studies. *ETS Research Bulletin Series*, 1972(2):i–31, 1972. doi: https://doi.org/10.1002/j.2333-8504.1972.tb00631.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1972.tb00631.x.

Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. SMART: Simulated students aligned with item response theory for question difficulty prediction. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25071–25094, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1274. URL https://aclanthology.org/2025.emnlp-main.1274/.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Sang Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*, 2025.

Victor Veitch, Dhanya Sridhar, and David M. Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:220793347.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. How is LLM reasoning distracted by irrelevant context? an analysis using a controlled benchmark. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13329–13347, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.674. URL https://aclanthology.org/2025.emnlp-main.674/.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9842–9852, 2021. doi: 10.1109/CVPR46437.2021.00972.

Zihao Yang, Mosh Levy, Yoav Goldberg, and Byron C Wallace. Compared to what? baselines and metrics for counterfactual prompting. *arXiv preprint arXiv:2605.01048*, 2026.

Junni L. Zhang and Donald B. Rubin. Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28 (4):353–368, 2003. ISSN 10769986, 19351054. URL http://www.jstor.org/stable/3701340.

Yunxiang Zhang, Liangming Pan, Samson Tan, and Min-Yen Kan. Interpreting the robustness of neural NLP models to textual perturbations. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3993–4007, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.315. URL https://aclanthology.org/2022.findings-acl.315/.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and Xing Xie. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, LAMPS '24, pp. 57–68, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400712098. doi: 10.1145/3689217.3690621. URL https://doi.org/10.1145/3689217.3690621.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1950–1976, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.108. URL https://aclanthology.org/2024.findings-emnlp.108/.

A Proof of Proposition 1

Holds for either $Q \in \{\tau, \pi\}$; write Q throughout. (i) $m(C) = \mathbb{E}[Q \mid C]$ is the L_2 projection of Q onto $L_2(\sigma(C))$, so the residual $D^{\perp} = Q - m(C)$ is orthogonal to that subspace: taking $\eta \equiv 1$ gives $\mathbb{E}[D^{\perp}] = 0$, and for general $\eta \in L_2(\sigma(C))$, $\mathrm{Cov}(D^{\perp}, \eta(C)) = \mathbb{E}[D^{\perp}\eta(C)] = 0$. Note that the projection is onto functions of the *declared* C. (ii) For the plug-in estimator, $\hat{D}^{\perp} - D^{\perp} = (\hat{Q} - Q) - (\hat{m} - m)$; both terms vanish in L_2 as $\hat{Q} \to Q$ ($\hat{\pi} \to \pi$ by the binomial rate's consistency in the number of observations, §3; $\hat{\tau} \to \tau$ by per-job concentration as $G \to \infty$) and the cross-fitted complexity regression is consistent, removing own-observation leakage.

B Synthetic data (Synth-CD)

A synthetic suite with *two* independent ground-truth confounders: structural complexity $C \sim \mathrm{Beta}(2, 2)$ and domain expertise $E \sim \mathrm{Beta}(2, 2)$ ($C \perp E$ by construction). Eight domains (web development, data engineering, ML operations, scientific writing, legal review, medical coding, financial modeling, creative content), each with 12 operator types spanning four semantic tiers from cosmetic-only through dual-sensitive. Text length is controlled to be independent of C and E, preventing the surface-length heuristic from proxying either confounder. 1,000 jobs, G=12 twins. Scores here are continuous and no stratum is censored, so the U-CCD is well posed on this suite (unlike the real suites, §3); we use it accordingly as a mechanism check rather than as a headline setting.

B.1 Synth-CD: dual-confounder robustness check

Generating process. The per-twin effect is $\delta(J, \mathrm{op}) = \max\big(0,\ \mu_{\mathrm{op}}(1{+}C(J)) + \zeta_{\mathrm{op}}\, E(J) + \varepsilon\big)$, calibrated so that $\mathrm{Spearman}(D_{\mathrm{avg}}, C) \approx \mathrm{Spearman}(D_{\mathrm{avg}}, E) \approx 0.60$ and $\mathrm{Spearman}(C, E) \approx 0.03$ (near-zero) in the holdout; the oracle $D^{\perp}(J) = D_{\mathrm{avg}}(J) - m(C(J))$ concentrates on the expertise signal ($\mathrm{Spearman}(\text{oracle } D^{\perp}, E) \approx 0.75$).

Table 5: Synth-CD: D_{avg} prediction (1,000 jobs, G=12 twins, 300 held-out, MiniLM backbone). Text length is controlled to be independent of C and E. ρ_S, NDCG@10, AUROC ↑; MAE, $|\rho_C|$ shown for context.

Method	ρ_S ↑	MAE↓	NDCG@10↑	AUROC↑	$\lvert\rho_C\rvert$
Surface length	0.100	0.021	0.804	0.594	0.052
Score-predictor	0.496	0.014	0.829	0.727	0.052
Frozen emb. distance	−0.112	0.023	0.686	0.446	0.182
Frail-τ ex-ante (ours)	0.420	0.015	0.799	0.712	0.067
Frail-τ contrast (ours)	**0.617**	**0.012**	**0.863**	**0.828**	**0.028**

Table 5 shows the Frail-τ contrast head reaching ρ_S=0.617 and AUROC 0.828, beating the score-predictor by 0.12 Spearman points and the ex-ante head by 0.20; surface length scores near-random (ρ_S=0.100), confirming the advantage is semantic rather than a token-count artifact. The contrast head also reaches the lowest $|\rho_C|$ of any predictor (0.028) *without* residualization.

C Implementation details

Parameter-efficient fine-tuning. For the frozen-backbone experiments, we employ an off-the-shelf small foundation model as the encoder, freeze the backbone, and learn only the heads via a two-stage curriculum: a label-free representation stage (triplet loss pulling twins together, pushing requirement-changing edits apart) and a small score-labeled calibration stage (10^3–10^4 pairs) via equation 8.

Frozen-backbone runs (Synth-CD and the frozen ablations). Predictor heads: hidden dim 256, complexity dim 64, dropout 0.15; contrast MLP 3 layers, ex-ante MLP 4 layers. Stage 1 (representation): 150 epochs, triplet margin 0.5, lr 10^{-3}, batch 128. Stage 2 (calibration): 400 epochs, lr $3{\times}10^{-4}$, AdamW, weight decay 10^{-4}, λ_{inv}=1.5, δ_{Huber}=0.05, cosine LR decay, batch 128. Synth-CD uses a frozen `all-MiniLM-L6-v2` (384-dim) with pre-computed,

cached embeddings (≈2 min on an M-series CPU). The GSM-IC-Adv/MATH-Adv frozen ablations of Table 6 use a frozen `text-embedding-3-small` (GPT-embed, 1536-dim) and a frozen `BAAI/bge-small-en-v1.5` (HF-frozen, 384-dim); heads use hidden dim 64, complexity dim 32/64, dropout 0.25, a 2-layer ex-ante head (avoiding overfitting the modest per-suite job count).

Main model (FRAIL-π, fine-tuned encoder). The headline real-suite model of §5.1 instead fine-tunes `bert-base-uncased` (768-dim matching BERTPE's own default so the comparison is like-for-like) end to end, with a token-level cross-attention. Heads: hidden dim 256, complexity dim 32, dropout 0.25, 2-layer ex-ante head. Stage 1 (representation): 10 epochs, triplet margin 0.5, head lr 10^{-3}, encoder lr 2×10^{-5}, batch 128. Stage 2 (calibration): 20 epochs, head lr 5×10^{-4}, encoder lr 2×10^{-5}, AdamW, weight decay 10^{-2}, δ_{Huber}=0.1, batch 128. The much lower encoder lr keeps full fine-tuning from washing out the pretrained representation in the few epochs the labelled pair count supports.

C.1 EXTENDED BASELINES

Here we discuss in more details how five published sensitivity predictors, plus our own frozen-geometry baseline, are adapted to this task: BERTPE (Khodabakhsh et al., 2024) and VAP3 (Arabzadeh & Bagheri, 2025) require retraining; Frozen twin-distance, Prompt Clarity, and Embedding Entropy (Razavi et al., 2025) need a job's twins at evaluation time; Embedding Specificity-IEF (Razavi et al., 2025; Arabzadeh et al., 2020) is ex-ante (base text only, scored against the training pool). Unlike BERTPE/VAP3, none of the last four fine-tunes a representation: each pairs a frozen TF-IDF+SVD embedding (the same representation Table 6 ablates for FRAIL-π) with the method's own unsupervised signal (twin-embedding distance, twin cosine similarity, pairwise-similarity entropy, or corpus specificity), and only a final linear calibration is fit to π – retargeting changes what the calibration maps to, not what the signal computes. All four score near zero on both suites (Table 2: ≤ 0.060 of ceiling in absolute value on GSM-IC-Adv, ≤ 0.160 on MATH-Adv): their unsupervised signals carry little to no rank information about which solvable jobs a model's twins will break.

Adaptation of supervised baselines. BERTPE and VAP3 were originally trained for binary per-prompt correctness classification via a binary cross-entropy (BCE) loss. We preserve their published architectures and hyperparameters verbatim – BERTPE's single-tower BERT→CLS→MLP head and VAP3's concatenated pair-encoder (base⊕[SEP]⊕twin) with FGSM adversarial perturbations on token embeddings (equations 5–8 of Arabzadeh & Bagheri 2025, ε=0.01, λ=0.5) – and replace only the loss: BERTPE regresses π directly (MSE, one sample per base job, prediction is the calibrated MLP output); VAP3 regresses the per-observation break event $B(J,T) = \mathbf{1}\{Y(J)=1, Y(T(J))=0\}$ (MSE, one sample per pair), with job-level difficulty the mean-over-twins aggregation of calibrated pair predictions, since π is a rate, not a worst case. Both models are fine-tuned from `bert-base-uncased` with AdamW.

C.2 C1 ABLATIONS: REPRESENTATION, TRAINING RECIPE, AND CROSS-GENERATOR VARIANTS

Other variants of FRAIL-π: the text-representation comparison, four further training-recipe levers, and the frozen-encoder representation ablation of the cross-generator check (§5.2). Table 6 compares the main model's fine-tuned `bert-base-uncased` encoder (to isolate the representation) against three frozen/non-adaptive representations. Neither frozen encoder beats plain TF-IDF, so text representation alone is not the bottleneck; letting the encoder *adapt* to the break-probability objective is what yields a real gain over every frozen representation, on GSM-IC-Adv (0.613 vs. a best frozen 0.398) and on MATH-Adv (0.557 vs. 0.435) alike.

Table 7 tries four further levers against the main model; none beats it meaningfully. A stronger backbone (RoBERTa-large in place of `bert-base-uncased`) trails on both suites. LR/patience/batch retuning is worse on GSM-IC-Adv and flat on MATH-Adv. Cross-representation ensembling and dataset pooling are worse on GSM-IC-Adv and marginally ($\leq$0.009) below the main model on MATH-Adv, within seed noise.

C.3 REAL-SUITE U-CCD BASELINES: WHY THE UNCONDITIONAL TARGET FAILS

This is the empirical basis for the estimand choice of §3. Table 8 complements the Synth-CD D_{avg} comparison (Table 5) with the same baseline sweep on a real suite, 5 seeds, scored directly against

Table 6: Representation ablation for FRAIL-π: test-fold Spearman *normalized by the suite reliability ceiling* (0.969 for GSM-IC-Adv, 0.972 for MATH-Adv; Appendix D.2), mean±sd over five seeds; 1.0 is the attainable maximum. Same protocol and normalization as Table 2, so these rows are directly comparable to the main model's 0.613/0.557.

Suite	Representation	$\pi\uparrow$	$a\uparrow$	$b\uparrow$
GSM-IC-Adv	TF-IDF+SVD	0.398±0.042	0.395±0.037	0.039±0.070
	Frozen sentence encoder	0.381±0.026	0.381±0.032	−0.126±0.112
	Frozen OpenAI embedding	0.326±0.040	0.324±0.039	−0.158±0.046
MATH-Adv	TF-IDF+SVD	0.435±0.021	0.425±0.021	0.334±0.095
	Frozen sentence encoder	0.386±0.013	0.394±0.014	0.314±0.037
	Frozen OpenAI embedding	0.431±0.016	0.417±0.015	0.369±0.022

Table 7: Further training-recipe levers vs. the 0.613/0.557 main model (break rate ρ_S *normalized by the suite reliability ceiling*, 0.969/0.972; Appendix D.2; same normalization as Table 2, 1.0 is the attainable maximum; 5 seeds unless noted). None beats the main model meaningfully.

Lever	GSM-IC-Adv	MATH-Adv
Main model (fine-tuned encoder)	0.613	0.557
Stronger backbone (RoBERTa-large)	0.543	0.548
Lower LR / higher patience	0.567	0.552
Larger job batch	0.606	—
Ensemble (mean-rank / mean-z)	0.559 / 0.537	0.549 / 0.551
Pooled training data (per-origin)[§]	0.536	0.548

[§]Trained on GSM-IC-Adv and MATH-Adv pooled, then scored separately on each suite's own holdout. "—": lever not run on that suite.

the unconditional U-CCD rather than the CCD. Raw D_{avg} is itself confounded with the declared complexity proxy C (Spearman$(D_{\text{avg}}, C) = 0.278\pm0.040$), so, unlike the CCD results in Table 2, no method here should be read as complexity-orthogonal. Every predictor is weak and within noise of the others – FRAIL-τ reaches the lowest MAE (0.108 ± 0.016) and is tied with VAP3 for the best ρ_S (0.092), but neither beats the alternatives outside one seed's worth of noise. Against FRAIL-π's 0.594 on the same corpus, this is the difference the conditioning makes: on real, binary-scored suites the unconditional average pools a live effect with a structurally censored stratum and is a noisy target for *every* method, which is why the paper's claims are stated for the per-observation CCD/CRD estimands.

Table 8: GSM-IC-Adv (real suite): U-CCD prediction on the audited, multi-operator-mix, pooled holdout, mean±sd over 5 seeds. ρ_S, NDCG@10, AUROC ↑; MAE, $|\rho_C|$ shown for context.

Method	$\rho_S\uparrow$	MAE↓	NDCG@10↑	AUROC↑	$\lvert\rho_C\rvert$
Surface length	0.162±0.030	0.245±0.009	0.453±0.065	0.727±0.041	0.086±0.040
Correlational score-predictor	0.028±0.036	0.254±0.010	0.030±0.036	0.544±0.045	0.037±0.019
Frozen emb. distance	−0.101±0.054	0.238±0.012	0.107±0.075	0.366±0.044	0.065±0.051
BERTPE	0.073±0.043	0.192±0.010	0.175±0.140	0.611±0.052	0.044±0.029
VAP3	0.092±0.089	0.189±0.034	0.186±0.168	0.621±0.095	0.059±0.042
Embedding Specificity-IEF (PromptSET)	0.022±0.005	6.102±3.501	−0.011±0.134	0.508±0.002	0.030±0.005
Embedding Entropy (PromptSET)	0.011±0.026	0.214±0.011	−0.013±0.027	0.529±0.036	0.091±0.039
Embedding linear probe (RidgeCV)	0.053±0.064	0.170±0.006	0.079±0.104	0.580±0.061	0.024±0.011
Prompt Clarity (PromptSET)	−0.106±0.051	0.188±0.010	0.118±0.088	0.355±0.045	0.066±0.042
FRAIL-τ (ours)	**0.092±0.059**	**0.108±0.016**	0.197±0.122	0.614±0.051	0.040±0.021

D BENCHMARKS

Construction, statistics, and audits behind every real-data benchmark used in the paper: the datasets themselves, why the public alternatives fall short of R1–R3 (§4), the measurement audit motivating the estimand change of §3, the empirical No-Op audit, the C2 controls on the real suites, and the downstream-utility ranking check.

D.1 Datasets

GSM-Symbolic (not used in main experiments). The GSM-Symbolic p1 set (Mirzadeh et al., 2025) changes numbers and names, which changes the correct answer and violates the No-Op requirement (Assumption 1); we exclude it (see Appendix D.2).

GSM-IC (irrelevant context). GSM8K problems with a single injected irrelevant clause; trivially-ignorable distractors produce near-zero fragility on current LLMs.

MATH (original). The Hendrycks MATH benchmark (Hendrycks et al., 2021) (`EleutherAI/hendrycks_math`, test split): $5{,}000$ competition-level problems across seven subjects, each with a human-assigned level (1–5) mapped to `complexity_score`. No distractor perturbation suite is provided (Twins $= 1$). External level supplies a non-circular C proxy (**R3**), but causal-difficulty ranking requires an injected twin suite (**R2**).

GSM-IC-Adv (ours). An upgraded GSM-IC whose twins are built from three declared MPPO families (entity renaming, sentence reordering, and injecting a math-domain setup sentence from a *different* GSM8K problem) raising fragility from 3.7% to 50.2% at $m{=}9$ while preserving the answer key. Dose m is the total number of perturbation operations applied to build a twin, drawn independently at random from the three families. We read $m{=}1$ as the conceptually clean case (Figure 2(a) shows a reordering draw), and $m \in \{1, 3, 5, 7, 9\}$ as a deliberate dose-response probe: higher doses stack more operations. The audited corpus holds 1,318 unique problems; 20 twins per problem before filtering.

MATH-Adv (ours). The GSM-IC-Adv protocol applied to Hendrycks MATH: base $=$ original problem, twin $=$ the same problem with the same three-family perturbation while preserving the gold answer. Each problem carries a human-assigned competition level and subject tag; level maps to an external `complexity_score` in $[0.05, 0.95]$, a non-circular complexity proxy. MATH-Adv carries 7 MATH subject categories (Algebra, Prealgebra, Number Theory, Intermediate Algebra, Geometry, Counting & Probability, Precalculus).

D.2 Measurement audit: what the targets actually are

A parametric bootstrap – with a cross-validated empirical-Bayes prior (fit on held-out job folds, so no job's own rate informs the prior it is shrunk toward) and replicates drawn from each fold's posterior rather than a fixed shrunk point estimate, to avoid overstating agreement from shared shrinkage – agrees closely with the classical split-half estimator (0.939 vs. 0.941 on GSM-IC-Adv, 0.945 vs. 0.951 on MATH-Adv), giving an implied ceiling on any predictor's Spearman of $\sqrt{r}{=}0.969$ and 0.972 respectively. This ceiling also bounds the reliability of stratum membership itself: $Y(J)$ is a measured, not given, quantity (§7).

Identity-twin floor. Because solver outputs are sampled, $B(J, T)$ can fire even when T has no effect. We measure this floor with identity twins, $T(J) = J$: the unperturbed base prompts of solvable-stratum items that the reference ($m{=}9$) run also solves were re-solved in independent runs of the same pipeline, solver, and decoding (temperature 0.2, 5 samples, modal answer; four further runs per item on GSM-IC-Adv, three on MATH-Adv), and a re-solve counts as a break when its extracted answer differs from the reference run's correct answer. Selecting on one run and scoring the others mirrors how B is computed, so the floor also absorbs any regression to the mean from conditioning on a measured $Y(J)$; it is the average CCD that a perturbation family with no effect would receive on the same stratum, and on both suites it lies below the 0.1 cutoff at which we call an item fragile. Table 9 reports it. The floor is small on GSM-IC-Adv and larger on MATH-Adv; on both suites the modal base answer almost never moves, so an item's stratum membership is reproducible rather than the outcome of one lucky draw.

D.3 Downstream utility: top-k curation

C1–C2 are measured against the estimand; a practitioner cares whether the ranking is actionable. We run the curation experiment the method is motivated by: on the audited GSM-IC-Adv holdout, select the top-$k\%$ of jobs by each predictor ($k \in \{5, 10, 20, 50\}$) and report recall of truly fragile jobs ($\pi > 0.05$) against Surface features (GBT) and Declared complexity – the two simple baselines

Table 9: Identity-twin floor: break rate of re-solved, unperturbed base prompts on the solvable stratum (the average CCD of a no-effect perturbation).

Suite	Re-runs/item	Floor	Modal answer $\neq$ ref.
GSM-IC-Adv	4	1.1%	4 (0.3%)
MATH-Adv	3	2.5%	0 (0.0%)

of Table 2 – and an oracle ranking by the measured π itself. FRAIL-π here is the real-suite headline predictor (§5.1), ranked by its actual held-out predictions from the audited, pooled 5-dose run behind Table 2 (same 5 seeds and splits, no retraining).

Table 10: Recall of truly fragile jobs ($\pi > 0.1$) in the top-k% of the GSM-IC-Adv holdout ranked by each predictor, mean±sd over the same 5 seeds as Table 2 (373 held-out jobs/seed, 50.8% fragile on average). Random is the expected recall k; oracle ranks by the measured π itself. The practically meaningful quantity is the lift over random at small k.

Ranking	Recall@5%	Recall@10%	Recall@20%	Recall@50%
Random (expected)	0.05	0.10	0.20	0.50
Surface features (GBT)	0.066±0.005	0.135±0.012	0.251±0.026	0.555±0.044
Declared complexity	0.064±0.007	0.122±0.006	0.244±0.007	0.567±0.018
FRAIL-π (ours)	**0.099±0.002**	**0.189±0.005**	**0.359±0.015**	**0.725±0.020**
$\widehat{D}^{\perp}$ (ours)	0.099±0.002	0.186±0.006	0.352±0.021	0.676±0.020
Oracle (π)	0.100±0.001	0.195±0.002	0.396±0.004	0.981±0.011

On the audited GSM-IC-Adv holdout (Table 10), FRAIL-π roughly doubles random recall at small k (Recall@5%=0.099 vs. 0.05, matching the oracle's 0.100 almost exactly) and stays ahead of both simple baselines at every budget; Surface features and Declared complexity each beat random by a modest, similar margin, well below FRAIL-π. The CRD ranking trails raw π slightly at every budget (e.g. 0.676 vs. 0.725 at Recall@50%) – the expected cost of trading some raw ranking power for orthogonality to the declared complexity proxy. Relative to random, surfacing ≈36% of fragile items requires evaluating only 20% of the corpus, against an oracle upper bound of 0.396 at the same cut.

D.4 EMPIRICAL NO-OP AND COLLISION AUDIT OF THE GENERATED TWINS

Here we check whether a twin's perturbations preserve the gold answer (the No-Op requirement, R1) rather than mislabeling a scoring artifact as causal difficulty. Twins are LLM-authored, applying operators from three families (entity renaming, sentence reordering, injection); since an LLM writes every twin, any family, not only injection, can alter answer-relevant text, so the audit covers all three families. Both real-math suites (`GSM-IC-Adv`, `MATH-Adv`) were passed through a full-corpus audit: every (base, twin) pair is adjudicated, and a pair is dropped before it enters any benchmark if **any** of the following holds.

LLM No-Op adjudication. An LLM adjudicator receives the base text, the twin text, and the gold answer, and judges whether the twin's perturbations change or invalidate the correct answer. A "contradiction" verdict, or an unparsable response after retries, drops the pair.

Renaming and reordering checks. The two non-injection families are checked directly: a renamed entity must keep its gender and pronoun references (he/she/it), so renaming cannot introduce a referential error that changes the answer, and a reordered twin must preserve the base problem's gold answer.

Structural collision checks. Free, local re-implementations of the pipeline's No-Op structural audit catch the contamination modes that an LLM alone can miss deterministically:

- In *entity-name collision*, the injected distractor reuses a proper noun from the target problem, so referential ambiguity is introduced (calendar words such as weekday/month names are excluded);

- In *gold-answer-number collision*, the injected text contains the numeric value of the correct answer, so a model quoting the distractor number is scored correct for the wrong reason;
- *multi-question collision*; an injected distractor containing its own "?" competes for attention, violating single-question No-Op;
- *unit-compatible quantity collision*; a shared numeric quantity with a recognized unit. This heuristic flag on MATH-Adv alone would flag $10K$ of the $100K$ pairs, yet re-adjudicating *all* ≈10K such flagged-but-kept pairs with an independent LLM adjudicator (Claude, same prompt as the contradiction check above) returned no actual contradiction – the only borderline flags traced to a pre-existing data gap (the stored problem text omitted its own multiple-choice answer options, making the lettered gold answer unverifiable independent of the injected text), not a semantic effect of the distractor. So it is recorded but, conservatively, not used to drop pairs.

Any entity-, gold-answer-, or multi-question collision, and any LLM contradiction verdict, removes the pair; the surviving pairs constitute the benchmarks used throughout the paper. Table 11 reports the resulting counts for both suites.

Human-in-the-loop audit. To complement the automated filter, we audited a $1{,}000$-pair sample spanning every audit outcome above (LLM-flagged contradictions, unparsed adjudicator responses, unit-collision pairs flagged but kept, and unflagged pairs) with two independent LLM adjudicators –`gpt-5`, the pipeline's own adjudicator, and Claude, queried separately – followed by human review of the full sample against the base and twin text, the gold answer, and both models' rationales. The audit confirms the filter errs conservatively rather than laxly: the sampled `llm_contradiction` drops trace to the adjudicator re-deriving the *original* problem's answer and making its own arithmetic slip, not to any semantic effect of the perturbation, so borderline pairs are excluded rather than risking contamination reaching the benchmark. The unit-collision-flagged-but-kept pairs were independently confirmed clean by the second adjudicator over the full ≈10K flagged pool, and by both LLM judges and the human reviewer on the sampled subset, with the human and LLM verdicts agreeing on every audited pair. The LLM cross-check is therefore full-corpus over the flagged pool; the human pass remains a $1{,}000$-pair spot-check rather than a full-corpus human audit, but it reproduces the LLM-judge conclusion exactly at that scale.

Table 11: Contamination audit of the two real-math suites. "Kept" is the number of pairs entering the benchmark; "before" is the raw un-audited count. The reason rows are not mutually exclusive – 29 GSM-IC-Adv pairs and 59 MATH-Adv pairs are dropped for two reasons at once, so each reason row counts them again and the reasons sum to slightly more than *Dropped*.

	`MATH-Adv`	`GSM-IC-Adv`
Pairs before audit	99,854	26,374
Kept (benchmark)	**87,067**	**21,534**
Dropped	12,787	4,840
gold-answer collision	12,080	4,738
entity collision	608	130
LLM contradiction	112	0
unparsed response	46	1
Unique problems	4,994	1,318

The drop rates are comparable, and GSM-IC-Adv actually drops a slightly higher fraction of pairs than MATH-Adv (18.4% vs. 12.8%); MATH-Adv shows more collisions in absolute terms because it is $\approx 4\times$ larger (4,994 vs. 1,318 base problems), with gold-answer collisions dominating both. Both suites draw distractor sentences from the same GSM8K, non-question setup source (neither contains question-bearing filler), so the heavier absolute footprint on MATH-Adv reflects its greater scale and the numeric density of GSM8K filler colliding with the target's own gold value; not a distinct filler type. Both suites use LLM-authored twins built from all three operator families (entity renaming, sentence reordering, and injection), so this audit (which adjudicates every pair across all three families, since an LLM-authored rewrite of any family can alter answer-relevant text) is the sole gate before the targets of §3 are measured.