Fragility Is Not Reducible to Hardness: Conditional Causal Difficulty for LLMs

Mohammad Reza Rezaei Dingxian Wang Nikhil Verma Sajjad Pakdamansavoji Andrew Rabinovich Zhao Chen

Upwork

Measure how often a meaning-preserving rewrite breaks a task the model already solves: this fragility is predictable from text, and separable from how hard the task is.

Motivation

- LLMs fail when a task is **rewritten without changing its requirements**, e.g. reordering two independent sentences.
- Standard difficulty scores (IRT, pass rate) conflate hardness with sensitivity to phrasing.
- Hard tasks are also more phrasing-sensitive, so complexity is an **effect modifier**, not a confounder.

Two causal quantities

Conditional causal difficulty (CCD)

The chance that a random meaning-preserving twin T **breaks a task the model solves**:

$$\pi(J) = \Pr_{T\sim\mathcal{P}}\left[\, Y(T(J)) = 0 \;\middle|\; Y(J) = 1 \,\right]$$

Complexity-residualized difficulty (CRD)

Subtract what complexity C already predicts:

$$D^{\perp}(J) = \pi(J) - \mathbb{E}\big[\pi \mid C = C(J)\big]$$

- CRD = 0: as fragile as its complexity predicts. CRD > 0: anomalously fragile.

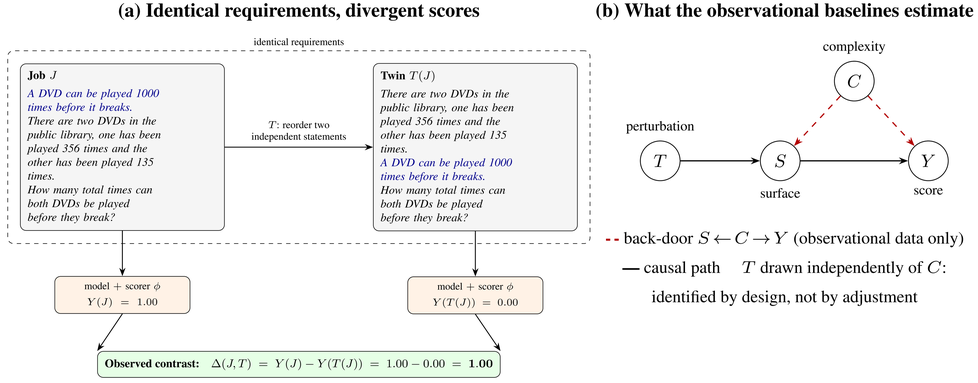
Frail-π: a causally aligned predictor

- One shared encoder f_θ embeds a task and its twin; the **contrast** is the primitive:

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

- The single-tower head g_ω predicts CCD **before** any twin exists.
- Cross-fitted residualization gives CRD: $\widehat{D}^{\perp} = \hat{D} - \hat{\gamma}(C)$.

Same requirements, different scores



(a) A real GlassGSM pair: moving one statement makes the model fail. (b) Comparing J with its own twin avoids the back-door through complexity C.

Benchmarks: GlassGSM and GlassMATH

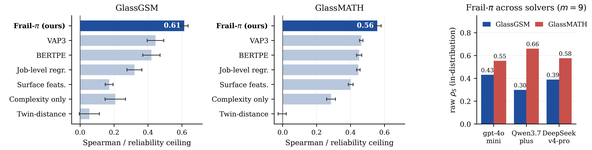
- Augment GSM8K and MATH with **twins** from perturbation families; the **answer key never changes**.
- A usable suite must be: **R1** truly no-op, **R2** fragile enough to rank, **R3** free of an inverse complexity confound.

Suite	Fragile	Twins	R1–R3
GSM-IC	3.7%	20	✗
GSM-Plus	10.5%	2	✗
GSM-Symbolic	25.0%	10	✗
GSM-DC	16.4%	1	✗
GlassGSM	**44.9%**	13.9	✓
GlassMATH	**36.5%**	23.1	✓

Conclusion

- Fragility is a **causal**, per-task property, not hardness in disguise; measure it on the solvable stratum and predict it ahead of time.

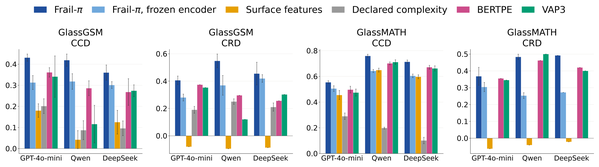
Result 1: CCD is predictable



Spearman / reliability ceiling, 5 seeds.

Method	GSM	MATH
Frail-π	**0.613**	**0.557**
VAP3	0.444	0.463
BERTPE	0.420	0.451
Surface features	0.171	0.401
Complexity only	0.207	0.285

Holds across solvers and twin generators



Result 2: useful and orthogonal

- At a 5% curation budget, Frail-π doubles random recall of fragile items and matches the oracle.
- CRD stays near-uncorrelated with complexity ($|\rho_C| \leq 0.09$) across 3 twin generators.