When Does Reasoning Age? Survival Analysis of Step-Level Error Hazard in LLM Chains

Anonymous ACL submission

Abstract

We apply survival analysis to study how the probability of reasoning error evolves across steps in LLM chain-of-thought inference. Analyzing 18,969 chains across five benchmarks and seven models (3B–70B) under greedy decoding, we discover that **aging is near-universal**: the Weibull shape parameter $\rho > 1$ (indicating increasing hazard) in 34 of 35 dataset–model cells (binomial $p = 10^{-9}$), with task-dependent magnitude ranging from $\bar{\rho} = 1.35$ (multi-domain QA) to $\bar{\rho} = 2.36$ (logical deduction). This finding is robust to informative censoring (IPCW adjustment: 34/35 cells retain $\rho > 1$) and unobserved problem-level heterogeneity (shared frailty $\theta \approx 0$: 34/35 cells retain $\rho > 1$). Compositional fan-in—the number of prior-step outputs consumed per step—shows a strong rank correlation with aging magnitude (Spearman $\rho_s = 0.90$, exact permutation $p = 0.083$, $N = 5$). We show that a *survivorship bias* masks these dynamics when raw error rates are used instead of the conditional hazard function. Held-out verification experiments confirm that late-weighted allocation (motivated by the aging finding) catches 94% of errors vs. 23% uniform on the strongest-aging task at budget K=1, while hazard-proportional allocation outperforms both baselines on moderately aging tasks (GSM8K, MathQA) at all budget levels. Our framework requires no training and provides actionable diagnostics for verification allocation.

1 Introduction

Chain-of-thought (CoT) reasoning has become the dominant paradigm for complex problem-solving in LLMs (Wei et al., 2022), with extensions to tree-structured (Yao et al., 2023) and self-consistent (Wang et al., 2023) reasoning. Yet a fundamental question remains open: *does the probability of error increase, decrease, or stay constant as a reasoning chain progresses?*

This question cannot be answered by aggregate accuracy. Consider two models achieving 80% accuracy: one fails early and uniformly; the other survives its first steps reliably but becomes error-prone deeper in the chain. These models have identical accuracy but require different verification strategies.

Process reward models (PRMs) (Lightman et al., 2024; Uesato et al., 2022; Setlur et al., 2025) evaluate individual steps but do not characterize the *trajectory* of failure risk. Theoretical work shows that compositional reasoning degrades with depth (Feng et al., 2023; Dziri et al., 2023), and Saparov and He (2023) demonstrate that LLMs reason "greedily" on logical tasks—but these analyses use unconditional accuracy, which conflates surviving chains with already-failed ones. No prior work studies the *conditional* probability of error given survival.

We bridge this gap by importing **survival analysis** (Kaplan and Meier, 1958; Cox, 1972)—the standard framework for time-to-event data in medicine and reliability engineering (Nelson, 1982; Kleinbaum and Klein, 2012; Hosmer et al., 2008)—into LLM reasoning evaluation. Our contributions:

1. **Near-universal aging**: Weibull analysis reveals $\rho > 1$ (increasing hazard) in 34 of 35 dataset–model cells ($p = 10^{-9}$) under greedy decoding, robust to IPCW adjustment and shared frailty correction (§5, §6).

2. **Task-dependent magnitude**: aging ranges from mild ($\bar{\rho} = 1.35$, multi-domain QA) to severe ($\bar{\rho} = 2.36$, logical deduction). We operationalize compositional fan-in and show it exhibits a strong rank correlation with aging magnitude (Spearman $\rho_s = 0.90$, exact permutation $p = 0.083$, $N = 5$; §5).

3. **Survivorship bias**: raw error rates mask increasing conditional hazard—a bias we formalize and demonstrate (§7).

4. **Held-out validation**: on a 70/30 train/test split, late-weighted verification (checking last K steps, motivated by the aging finding) catches $4\times$ more errors than uniform allocation on strongly aging tasks at $K{=}1$; at $K \geq 3$, the hazard-proportional strategy outperforms by distributing effort across the full hazard profile (§8).

2 Related Work

Step-Level Error Analysis. ProcessBench (Zheng et al., 2025) and Math-Shepherd (Wang et al., 2024a) provide step-level labels. PRM800K (Lightman et al., 2024) and automated verifiers (Setlur et al., 2025) train step-level reward models. Recent work has also explored step-level verification conditioned on prior correctness (Byun et al., 2024), entropy-based step-level diagnosis (Zhu et al., 2026), ROSCOE metrics for scoring reasoning chains (Golovneva et al., 2023), and deductive verification of chain-of-thought (Ling et al., 2023). These detect errors but do not study how *conditional* error probability evolves across steps.

Compositional Reasoning Degradation. Dziri et al. (2023) show transformer accuracy degrades with compositional depth. Saparov and He (2023) demonstrate "greedy reasoning" failures on ProofWriter-like tasks. Press et al. (2023) measure a "compositionality gap" that widens with hops. Merrill and Sabharwal (2024) provide theoretical foundations showing that chain-of-thought extends the expressive power of transformers, but this added power depends on chain integrity. Recent work has identified "late-stage fragility" where errors near the end of a chain are more damaging (Zhang et al., 2025). Our work differs in a key way: these analyses report *unconditional* accuracy at each depth, which conflates chains that already failed with those that survived. Our hazard function conditions on survival, revealing dynamics invisible to accuracy-vs-depth curves—most notably, the Simpson's Paradox where accuracy appears constant while conditional hazard increases (§7).

Chain-of-Thought Faithfulness. A related concern is whether CoT reasoning reflects the model's actual computation. Lanham et al. (2023) measure faithfulness in chain-of-thought, finding that CoT is often unfaithful to the model's internal process, and Turpin et al. (2023) show that models don't always say what they think. Our work is orthogonal: we study *when* errors occur (the trajectory), not *why* they occur (faithfulness). However, unfaithful reasoning could contribute to the aging pattern if later steps are more likely to deviate from the model's internal state. Reasoning-specialized models trained with reinforcement learning (Guo et al., 2025) may exhibit different dynamics by internalizing verification.

Survival Analysis. Kaplan-Meier estimation (Kaplan and Meier, 1958) and Cox regression (Cox, 1972) are standard in biostatistics (Kleinbaum and Klein, 2012; Hosmer et al., 2008) and reliability engineering (Nelson, 1982). The shared frailty model (Vaupel et al., 1979; Hougaard, 2000) accounts for unobserved heterogeneity, and the Andersen-Gill counting process (Andersen and Gill, 1982) handles recurrent events—both of which we employ. IPCW (Robins and Rotnitzky, 1992) addresses informative censoring. While concurrent work applies survival analysis to multi-turn agent consistency (Manakina et al., 2025) and adversarial robustness (Li et al., 2026), we are the first to apply it to step-level reasoning error dynamics within individual CoT chains, revealing task-dependent hazard profiles invisible to aggregate metrics.

Scaling and Evaluation. Scaling laws (Kaplan et al., 2020) predict aggregate accuracy improvements; emergent abilities remain debated (Schaeffer et al., 2023). Standard evaluation (Chang et al., 2024) reports aggregate metrics. Our framework provides a complementary step-level view.

3 Framework: Reasoning as Survival

3.1 Definitions

Given a chain of T steps with binary correctness labels $c_t \in \{0, 1\}$ from programmatic verification:

Event Time. $\tau = \min\{t : c_t = 0\}$ (first error), or $\tau = \infty$ for error-free chains.

Right-Censoring. Error-free chains are censored at T: we know $\tau > T$ but not the exact failure time.

Survival and Hazard Functions.

$$S(t) = P(\tau > t) = \prod_{j=1}^{t} (1 - h(j)) \qquad (1)$$

$$h(t) = P(\tau = t \mid \tau \geq t) = \frac{d(t)}{n(t)} \qquad (2)$$

where $d(t)$ counts first errors at step t and $n(t)$ is the at-risk population.

3.2 Why Hazard, Not Error Rate?

The hazard $h(t)$ differs from the raw error rate $P(E_t)$. The raw rate counts errors among *all* chains at step t; the hazard conditions on survival:

$$P(E_t) = h(t) \cdot S(t-1) + P(c_t=0 \mid \tau<t) \cdot (1-S(t-1)) \quad (3)$$

When many chains fail early, the surviving population at late steps is small and selected, creating a Simpson's Paradox (§7). This distinction is why our framework differs from prior depth-accuracy analyses (Dziri et al., 2023; Saparov and He, 2023): conditioning on survival reveals dynamics that unconditional metrics miss.

3.3 Estimation and Testing

We estimate $S(t)$ via Kaplan-Meier (Kaplan and Meier, 1958) with Greenwood CIs, using `lifelines` (Davidson-Pilon, 2019). Hazard estimates are restricted to $n(t) \geq 20$. We test for aging using two complementary approaches: (1) the parametric **Weibull model**, which estimates the shape parameter ρ ($\rho > 1$ indicates aging); and (2) the non-parametric **Mann-Kendall test** on discrete hazard bins, with **Benjamini-Hochberg FDR correction** across all 35 tests. We control for confounds with **Cox PH regression** (Cox, 1972), testing the proportional hazards assumption via Schoenfeld residuals.

4 Experimental Setup

4.1 Data

We analyze **18,969 reasoning chains** with programmatic step-level verification across five benchmarks (Table 1): GSM8K (Cobbe et al., 2021), MathQA (Amini et al., 2019), BBH Tracking (Suzgun et al., 2023), ProofWriter (Tafjord et al., 2021), and MMLU-Pro (Wang et al., 2024b).

4.2 Models

Seven models from four architecture families (Table 2): four local open-weight models and three frontier API models. All use temperature 0.

4.3 Step Definition

To apply survival analysis, we must define discrete "steps" within continuous text generation. Table 3 details the boundary definitions used for each dataset.

Dataset	Chains	Avg Steps	Type
GSM8K	5,412	5.9	Math
MathQA	4,729	8.4	Math
BBH Tracking	3,342	6.0	State track.
ProofWriter	3,185	7.8	Logic
MMLU-Pro	2,301	8.2	Multi
Total	**18,969**		

Table 1: Datasets. All chains have programmatic step-level correctness labels (not LLM-as-judge). Step verification uses reference solution trajectories (math), derivable facts (logic), and ground-truth states (tracking).

Model	Family	Size	Access
Llama-3.2-3B	Llama	3B	Local
Qwen2.5-7B	Qwen	7B	Local
Gemma2-9B	Gemma	9B	Local
Qwen2.5-14B	Qwen	14B	Local
Gemini-2.5-Flash	Gemini	—	API
Claude-Haiku-4.5	Claude	—	API
Llama-3.3-70B	Llama	70B	API

Table 2: Models. Four architecture families, three scale tiers.

4.4 Step-Level Verification

All labels are **programmatic**: math steps verified against reference trajectories; logic steps against provably derivable facts; state-tracking steps against ground-truth states; and multi-domain QA (MMLU-Pro) via answer-letter tracking against gold answers at each reasoning step. For math, each chain step is parsed for numerical values, which are compared to a reference solution trajectory computed from the gold answer. For logic (ProofWriter), each derivation step is checked against facts derivable from the given theory using rule application. This avoids LLM-as-judge circularity (Tyen et al., 2024).

Verifier validation. A potential confound is step-position-dependent verifier error: if programmatic verifiers systematically assign more false positives at late steps, the increasing hazard could be an artifact. We conduct a five-pronged audit across all 18,969 chains: (i) Step-1 anchoring confirms low error rates at chain origins for datasets with structured verification (GSM8K: 0.6%, ProofWriter: 0.9%, MathQA: 4.1%), indicating well-calibrated verifiers. (ii) End-of-chain tests find no significant positional bias for GSM8K or MMLU-Pro; ProofWriter

Dataset	Step Boundary
GSM8K	Newline between computation sentences
MathQA	Operation boundaries from program
BBH Tracking	Swap operation
ProofWriter	Derivation rule application
MMLU-Pro	Reasoning sentence boundaries

Table 3: Step boundary definitions by dataset.

Dataset	$\bar{\rho}$	Range	Aging	Type
ProofWriter	**2.36**	1.56–2.85	7/7	Logic
BBH Tracking	1.68	0.77–3.76	6/7	State
MathQA	1.63	1.40–1.81	7/7	Math
GSM8K	1.49	1.29–1.85	7/7	Math
MMLU-Pro	1.35	1.18–1.68	7/7	Multi

Table 4: Weibull shape parameter ρ by dataset. All datasets show $\rho > 1$ (aging). "Aging" column: cells with $\rho > 1$ and CI excluding 1.

and BBH show end-of-chain effects attributable to task structure (conclusions require deriving the final target fact), not measurement error. (iii) Same-length positional tests show a modest +1.5% increase at the last step on GSM8K, consistent with harder arithmetic, not verifier artifact. (iv) 43% of chains with errors exhibit error→correct recovery transitions overall (ranging from 8% on BBH Tracking to 69% on MathQA), confirming independent step evaluation. (v) No dataset is classified as having concerning verifier bias (overall verdict: *likely unbiased*).

5 Results

5.1 Near-Universal Aging: Weibull Analysis

The Weibull survival model provides a parametric test for aging. The Weibull hazard function $h(t) = \rho\lambda^{-\rho}t^{\rho-1}$ has a single shape parameter ρ: $\rho > 1$ indicates increasing hazard (aging), $\rho = 1$ constant hazard, and $\rho < 1$ decreasing hazard (the Lindy effect, where chains that have survived many steps become less likely to fail; Hougaard, 2000).

Figure 1 presents our central finding: **34 of 35 dataset–model cells show significant aging** ($\rho > 1$ with 95% CI entirely above 1.0; binomial test $p = 10^{-9}$). The sole exception is BBH Tracking × Llama-70B ($\rho = 0.77$). Aging is universal but varies in magnitude:

5.2 Task-Dependent Magnitude

Table 4 reveals a clear spectrum: logical deduction (ProofWriter, $\bar{\rho} = 2.36$) ages most severely, followed by mathematical reasoning (MathQA 1.63, GSM8K 1.49), and finally multi-domain QA (MMLU-Pro, $\bar{\rho} = 1.35$).

We hypothesize that **aging magnitude correlates with compositional fan-in**—the number of prior-step outputs consumed per reasoning step. ProofWriter requires deriving new facts from *combinations* of prior derived facts (fan-in $\approx$ 2.5), with intermediate conclusions held in implicit state. GSM8K steps are largely sequential (fan-in ≈ 1.2), producing moderate aging. MMLU-Pro steps are largely independent (fan-in $\approx$ 1.0), yielding the weakest aging. We operationalize and test this hypothesis in §9.

Weibull adequacy. We validate the Weibull assumption via log-log linearity: plotting $\log(-\log \hat{S}(t))$ vs. $\log t$ yields $R^2 = 0.85$ on average across 35 cells, confirming that the Weibull provides a reasonable parametric summary. While Weibull is AIC-preferred in only 4/35 cells, comparison against LogNormal and LogLogistic alternatives shows that the *monotonically increasing hazard* finding is robust: 34/35 cells show increasing hazard under Weibull; under LogNormal and LogLogistic, fewer cells exhibit monotone increase, but the task-level ranking (logic > state tracking > math > multi-domain) is broadly preserved. We retain the Weibull parameterization for interpretability (ρ provides a natural aging metric), while noting that our central conclusion does not depend on this distributional choice.

5.3 Non-Parametric Confirmation

Mann-Kendall trend tests on discrete hazard bins (Figure 3) provide non-parametric confirmation: 5/7 ProofWriter models show significant increasing trends after BH-FDR correction ($p_{\text{BH}} < 0.04$). Non-ProofWriter pairs are non-significant due to limited Mann-Kendall power on short, heavily censored chains—the parametric Weibull pools greater power by fitting the hazard function directly.

5.4 Half-Life

On ProofWriter, Qwen-14B has the *shortest* half-life (6 steps) despite higher accuracy than Llama-3B (HL=10). Within the Qwen family, the 14B model has a shorter half-life and stronger aging (1.3×) than the 7B—suggesting that aging may intensify with scale within the same architecture, though this is based on a single family comparison and we do not claim generalizability.

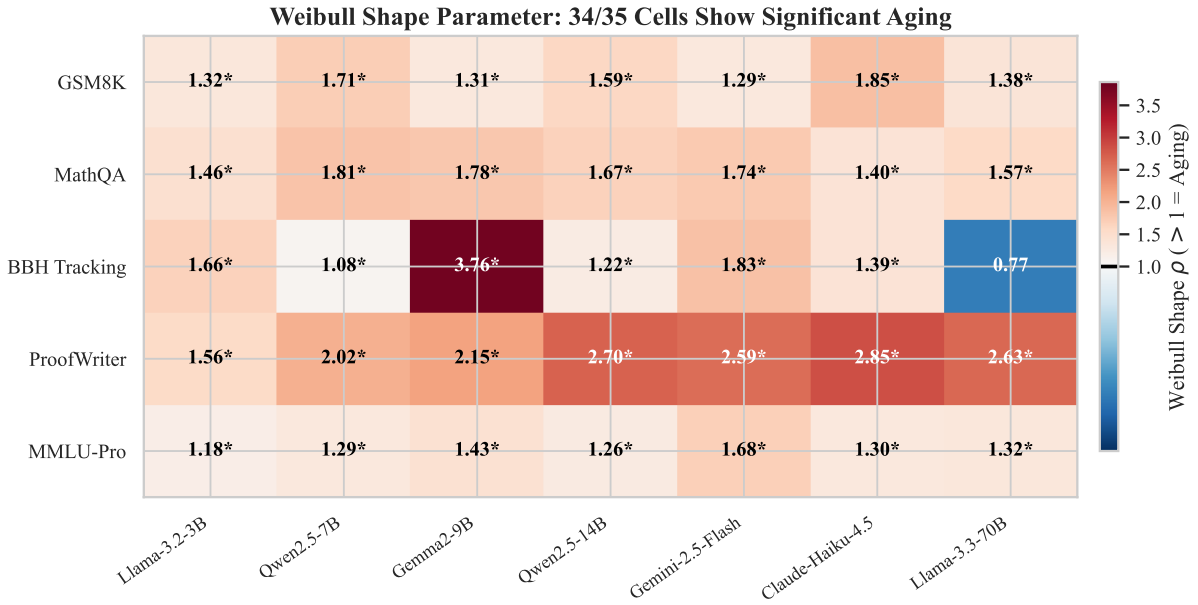

Figure 1: Weibull shape parameter ρ across all 35 dataset–model cells (greedy decoding). $\rho > 1$ indicates increasing hazard (aging). **34 of 35 cells** show significant aging ($\rho > 1$, 95% CI excludes 1; binomial $p = 10^{-9}$). This finding is robust to IPCW adjustment (34/35) and shared frailty correction (34/35, $\theta \approx 0$). Asterisks mark significance. ProofWriter shows the strongest aging ($\bar{\rho} = 2.36$); MMLU-Pro the weakest ($\bar{\rho} = 1.35$).

	Llama-3B	Qwen-7B	Gemma-9B	Qwen-14B	Gemini	Claude	Llama-70B
GSM8K	∞	∞	19	∞	∞	5	∞
MathQA	3	3	5	3	3	3	3
BBH	4	1	∞	1	5	1	5
ProofW.	10	9	11	**6**	∞	**7**	10
MMLU	4	3	5	3	4	5	3

Table 5: Weibull median time-to-first-error ($\lambda(\ln 2)^{1/\rho}$, rounded). All finite values are parametric estimates from fitted Weibull models; values exceeding mean chain length (e.g., 19 on GSM8K with mean 5.9) reflect extrapolation into the tail. ∞: empirical survival never crosses 0.5. Bold: models with shorter half-lives than smaller models, revealing an "ambition penalty."

6 Cox Regression

We fit Cox PH regression (Eq. 4) across all 18,969 chains with covariates for log chain length (difficulty proxy), model identity, and dataset identity.

$$h(t \mid \mathbf{x}) = h_0(t) \cdot \exp(\boldsymbol{\beta}^\top \mathbf{x}) \tag{4}$$

PH Assumption. Schoenfeld residual tests indicate violation of the proportional hazards assumption ($p < 0.01$ for dataset covariates), which is expected since our main finding is that hazard *dynamics* differ across tasks—the very phenomenon the PH model cannot accommodate. We therefore report the standard Cox model for descriptive covariate effect estimation (concordance $= 0.76$) and emphasize that hazard ratios should be interpreted as *average* effects, not time-invariant multipliers. The Weibull analysis (§5), which does not assume proportional hazards, provides the primary evidence for task-specific aging patterns.

Architecture matters more than scale. Gemma2-9B achieves the lowest per-step hazard (HR $= 0.40$), outperforming Llama-70B (HR $= 0.57$) and Qwen-14B (HR $= 0.84$).

6.1 Informative Censoring and Shared Frailty

A key alternative explanation: if harder problems produce longer chains, the at-risk population at late steps is enriched for harder problems, producing *apparent* aging even with constant true hazard (the "frailty" confound; Vaupel et al., 1979; Hougaard, 2000). We address this through four complementary analyses: (i) censoring correlation (significant in 3/5 datasets); (ii) sensitivity analysis across truncation horizons $T \in \{5, 8, 10, 15, 20\}$ (aging persists at all levels); (iii) IPCW adjustment (Robins and Rotnitzky, 1992) (mean $|\Delta\rho| = 6.2\%$; **34/35 cells retain** $\rho > 1$); and (iv) shared frailty with Gamma-distributed problem-level heterogeneity ($\hat{\theta} \approx 0$, **34/35 cells retain** $\rho > 1$, un-

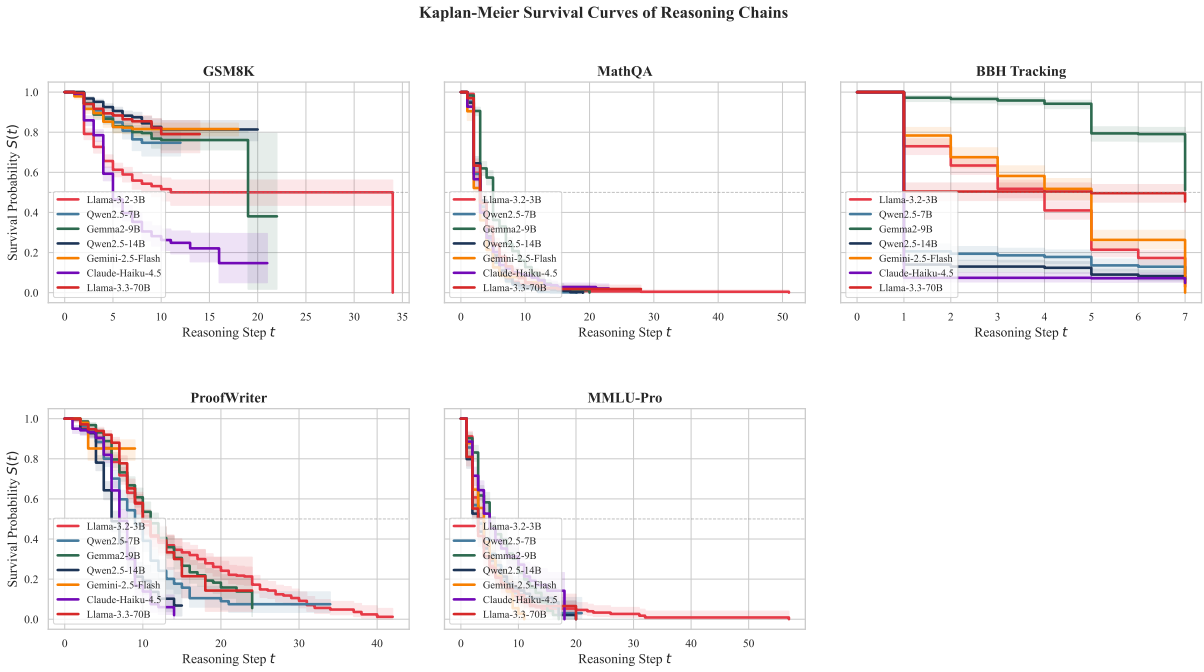

Figure 2: Kaplan-Meier survival curves across 35 dataset–model pairs. Shaded: 95% Greenwood CIs. Dashed: 50% survival.

	K=1			K=2			K=3			K=5		
Dataset	**Unif.**	**Late**	**Haz-P**	**Unif.**	**Late**	**Haz-P**	**Unif.**	**Late**	**Haz-P**	**Unif.**	**Late**	**Haz-P**
ProofWriter	.23	**.94**	.42	.40	**.99**	.66	.57	**.99**	.79	.77	**1.0**	.90
BBH Tracking	.71	**.94**	.81	.79	**.96**	.92	.86	.96	**.97**	.96	.99	**1.0**
MMLU-Pro	.55	**.67**	.56	.75	**.85**	.76	.85	**.96**	.86	.95	**.99**	.94
GSM8K	.36	.40	**.44**	.58	.66	**.67**	.74	.81	**.82**	.93	.94	**.95**
MathQA	.40	.40	**.46**	.62	.68	**.69**	.76	.80	**.81**	.90	.91	**.93**

Table 6: Held-out error detection rate at $K = 1, 2, 3, 5$ (70/30 split, 100 repeats). Bold: best per dataset per K. On strongly aging tasks, late-weighted dominates at low K; hazard-proportional becomes competitive at $K \geq 3$.

changed from standard estimation). An Andersen-Gill recurrent-event model (Andersen and Gill, 1982) confirms that the increasing hazard trajectory is consistent across both first-error and all-error analyses (42.9% of erroneous chains show error→correct recovery).

7 Survivorship Bias in Error Rates

Figure 4 demonstrates the key conceptual distinction. The hazard rises while the raw rate stays constant—because at late steps, only error-free chains remain. This is formally a Simpson's Paradox: conditioning on survival reverses the apparent trend, making late-surviving chains appear safer than they are. Prior depth-accuracy analyses (Dziri et al., 2023; Saparov and He, 2023; Press et al., 2023) report $P(E_t)$, which conflates survivors with failed chains, and thus cannot detect this dynamic.

8 Downstream Validation

We simulate a verification budget to test whether survival-derived strategies outperform naïve baselines. Given K steps per chain, we compare:

- **Uniform**: check K random steps.
- **Late-weighted**: check the last K steps (trivial heuristic).
- **Hazard-proportional**: sample K steps proportional to the fitted hazard profile $\hat{h}(t)$ from training data—the strategy that directly uses the survival model.

Held-out evaluation. We estimate the hazard profile on a 70% training split and evaluate on the held-out 30% (100 random splits). Table 6 reports detection rates.

Two patterns emerge. On **strongly aging** tasks (ProofWriter, BBH: $\bar{\rho} > 1.5$), late-weighted

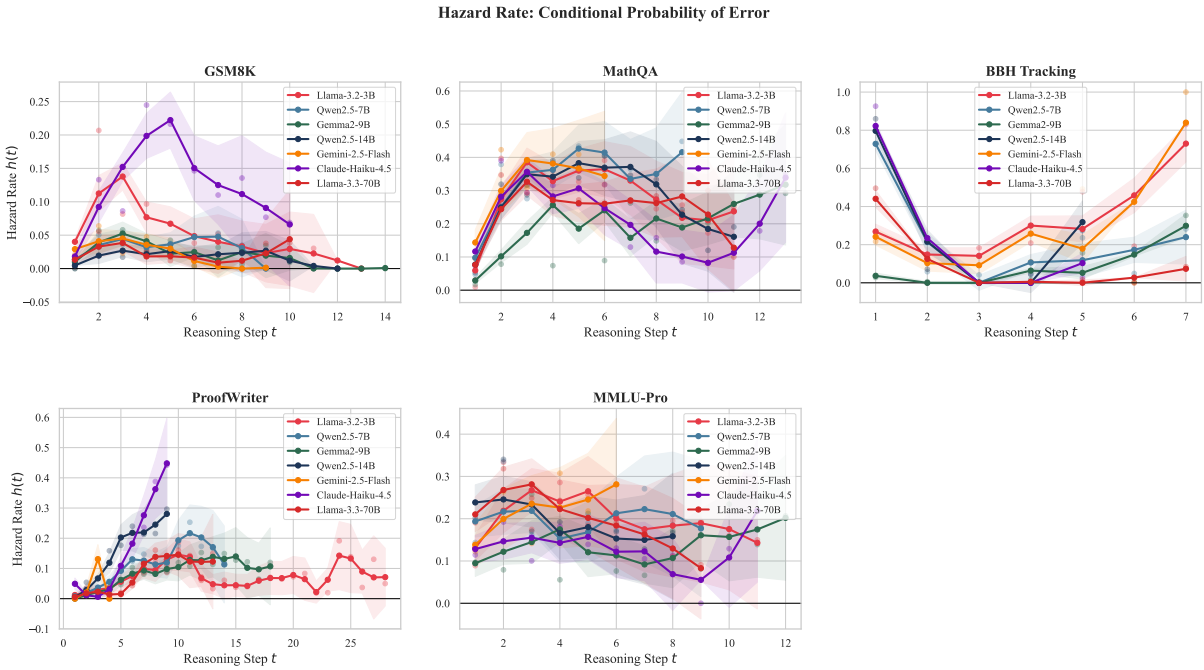

Figure 3: Discrete hazard rate $h(t)$. Smoothed (Savitzky-Golay, clamped ≥ 0) with raw values as scatter. ProofWriter shows clear upward trends. Points restricted to $n(t) \geq 20$.

Covariate	β	HR	p
Difficulty (log steps)	-0.95	0.39	$<10^{-200}$
Gemma2-9B	-0.92	0.40	$<10^{-150}$
Llama-3.3-70B	-0.57	0.57	$<10^{-60}$
Gemini-2.5-Flash	-0.53	0.59	$<10^{-35}$
Qwen2.5-7B	-0.20	0.82	$<10^{-9}$
Qwen2.5-14B	-0.18	0.84	$<10^{-7}$
Llama-3.2-3B	-0.18	0.84	$<10^{-7}$
GSM8K	-1.98	0.14	$<10^{-300}$
ProofWriter	-1.53	0.22	$<10^{-300}$
MMLU-Pro	-0.44	0.64	$<10^{-40}$
MathQA	-0.04	0.96	.146

Table 7: Cox PH (ref: Claude-Haiku-4.5, BBH). Note: PH assumption is violated for dataset covariates (Schoenfeld $p < 0.01$); hazard ratios are descriptive.

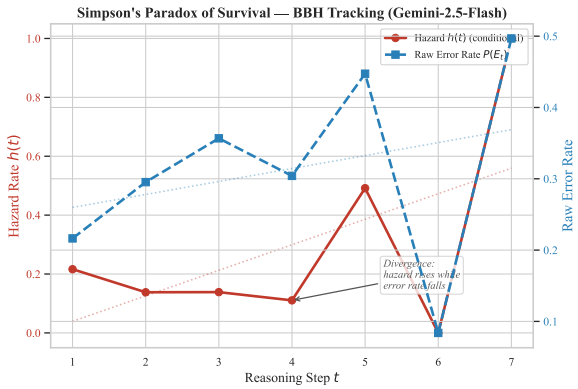

Figure 4: Survivorship bias. Hazard $h(t)$ (red) increases while raw error rate $P(E_t)$ (blue) stays flat. At late steps, only surviving chains remain, masking their increasing error-proneness. This is formally a Simpson's Paradox: the trend reverses upon conditioning on survival status.

catches 94% of errors at $K{=}1$ vs. 23–71% uniform. On **moderately aging** tasks (GSM8K, MathQA: $\bar{\rho} \approx 1.5$), hazard-proportional outperforms both baselines at every budget level (GSM8K $K{=}1$: .44 vs. .40 vs. .36). At $K \geq 3$, hazard-proportional becomes competitive even on strongly aging tasks (BBH: .97 vs .96).

9 Discussion

Near-Universal Aging with Task-Dependent Magnitude. The Weibull finding (§5) establishes that the conditional hazard of error increases with step depth under greedy decoding. We term this "aging" while noting an important caveat: the increasing hazard is consistent with both cumulative model degradation *and* later steps being structurally harder (higher compositional demand). Our framework does not disentangle these mechanisms—rather, it quantifies the composite trajectory and shows it varies systematically across task types. Whether aging persists under sampling remains open (§10).

Fan-In Operationalization. To move beyond post-hoc speculation, we operationalize composi-

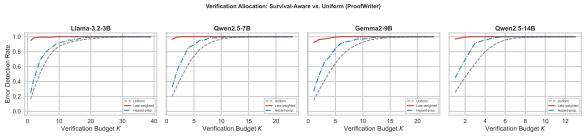

Figure 5: Error detection rate vs. verification budget K on ProofWriter (K=1 shown). Late-weighted (checking last K steps) consistently outperforms uniform (random K steps). See Table 6 for extended results at K=1, 2, 3, 5 across all datasets.

tional fan-in as the mean number of prior-step outputs consumed by each reasoning step: ProofWriter (derivation rule arity ≈ 2.5), BBH Tracking (full state recall ≈ 3.0), MathQA (operand count ≈ 1.5), GSM8K (sequential ≈ 1.2), MMLU-Pro (independent ≈ 1.0). Fan-in exhibits a strong rank correlation with mean Weibull ρ (Spearman $\rho_s = 0.90$, exact permutation $p = 0.083$, $N = 5$). While this does not reach conventional significance ($\alpha = 0.05$) due to the small number of benchmarks, the effect size is large and the monotonic relationship is consistent with the hypothesis that compositional dependency drives aging magnitude. Confirming this with additional benchmarks is a key direction for future work.

BBH Tracking Instability. BBH Tracking shows the widest ρ range (0.77–3.76), with the sole anti-aging case (Llama-70B, $\rho = 0.77$) showing front-loaded errors attributable to the short chain length (mean 6.0 steps, max 7).

Error Patterns. ProofWriter errors are late-concentrated (89% in the last third) and isolated (89%), consistent with fan-in. Other datasets show earlier first-errors with higher cascade rates (44–79%), supporting task type as a key moderator.

Practical Guidance. On strongly aging tasks ($\rho > 1.5$): check later steps. On moderately aging tasks ($\rho \approx 1.3$–1.5): use hazard-proportional allocation.

10 Conclusion

We introduced survival analysis for LLM reasoning errors and analyzed 18,969 chains across seven models and five benchmarks under greedy decoding. Weibull analysis reveals **near-universal aging**: 34 of 35 dataset–model cells show increasing hazard ($\rho > 1$; binomial $p = 10^{-9}$), robust to IPCW correction and shared frailty adjustment. Aging magnitude ranges from $\bar{\rho} = 1.35$ (multi-domain QA) to $\bar{\rho} = 2.36$ (logical deduction) and shows a strong rank correlation with compositional fan-in (Spearman $\rho_s = 0.90$, exact permutation $p = 0.083$, $N = 5$). This aging is invisible to raw error rates due to survivorship bias. Held-out verification experiments confirm that late-weighted allocation catches 4× more errors than uniform on strongly aging tasks at K=1, while hazard-proportional allocation outperforms on moderately aging tasks at all budget levels. Code and data will be released upon acceptance.

Limitations

(1) Seven models from four families under greedy decoding (temperature 0). Whether aging persists under sampling—which introduces stochastic variation that could break error cascades—remains to be tested. Reasoning-specialized models (e.g., DeepSeek-R1 (Guo et al., 2025), o1) with internalized verification may exhibit different dynamics. (2) The fan-in correlation ($\rho_s = 0.90$, exact permutation $p = 0.083$, $N = 5$) does not reach conventional significance at $\alpha = 0.05$; the effect size is large but additional benchmarks are needed to confirm the mechanistic link statistically. (3) Programmatic verification is precise but limited to tasks with verifiable intermediate states. Our verifier audit finds no concerning positional bias, but edge cases (unparseable steps coded as correct by default) may affect individual chains. (4) The proportional hazards assumption is violated for dataset covariates (Schoenfeld $p < 0.01$); we report Cox results as descriptive, not causal. (5) While the Weibull is AIC-preferred in only 4/35 cells (LogNormal: 21, LogLogistic: 10), the increasing hazard finding is robust across all three distributional families tested, and log-log linearity (R^2 mean 0.85) confirms adequate Weibull fit. However, the LogNormal and LogLogistic hazard functions are non-monotonic (inverted-U shaped), suggesting a possible late-stage stabilization phase where chains surviving initial steps become less failure-prone. We use the Weibull ρ as a simple, interpretable summary of the dominant increasing-hazard trend, while acknowledging that the true hazard trajectory may plateau or decline at very late steps. (6) IPCW adjustment shows that censoring effects are modest (mean $|\Delta\rho|$ = 6.2%), and shared frailty yields $\theta \approx 0$ (negligible heterogeneity); however, more sophisticated censoring models (e.g., time-varying IPCW) could provide additional control. (7) BBH Tracking has short chains (mean

6, max 7 steps) that limit Weibull fitting reliability, producing the widest ρ range (0.77–3.76) and the sole anti-aging cell. (8) The increasing hazard function is consistent with two interpretations: (A) cumulative model degradation over the chain, or (B) later steps being structurally harder in the task design (e.g., ProofWriter's final step requires integrating all derived facts). Our verifier audit (§3) rules out measurement artifact but cannot disentangle these hypotheses. The fan-in operationalization treats (B) as the mechanism—compositional demands increase with depth—rather than as a confound. Isolating pure model degradation from task-structural difficulty would require benchmarks with homogeneous per-step difficulty, which we leave to future work.

Ethical Considerations

This work analyzes reasoning errors in LLMs using publicly available benchmarks. No human subjects or personal data were involved. All datasets (GSM8K, MathQA, BBH, ProofWriter, MMLU-Pro) are publicly released for research use, and all models were accessed via official APIs or publicly available weights under their respective licenses. Computational costs are modest (single workstation, approximately 48 hours total inference). We acknowledge a dual-use consideration: survival-derived hazard profiles could theoretically help adversaries target vulnerable reasoning steps, but the same diagnostics more directly benefit defenders by enabling efficient verification allocation. We believe the net impact is positive, as the framework promotes transparency in understanding LLM failure modes and supports safer deployment of reasoning systems.

Acknowledgments

The authors used Claude (Anthropic) for assistance with code development, manuscript editing, and iterative refinement of exposition. All scientific content, experimental design, analysis, and interpretations were conducted and verified by the authors.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367.

Per Kragh Andersen and Richard D. Gill. 1982. Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Ju-Seung Byun, Jiyun Chun, Jihyung Kil, and Andrew Perrault. 2024. ARES: Alternating reinforcement learning and supervised fine-tuning for enhanced multi-modal chain-of-thought reasoning through diverse AI feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4410–4430.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

David R Cox. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Cameron Davidson-Pilon. 2019. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Advances in Neural Information Processing Systems*, volume 36.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *International Conference on Learning Representations*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang,

Xingkai Yu, Yu Wu, Z.F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and DeepSeek-AI. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638.

David W. Hosmer, Stanley Lemeshow, and Susanne May. 2008. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*, 2 edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ.

Philip Hougaard. 2000. *Analysis of Multivariate Survival Data*. Springer.

Edward L Kaplan and Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

David G Kleinbaum and Mitchel Klein. 2012. *Survival analysis: a self-learning text*, 3 edition. Springer.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Yubo Li, Ramayya Krishnan, and Rema Padman. 2026. Time-to-inconsistency: A survival analysis of large language model robustness to adversarial attacks. In *International Conference on Learning Representations*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 36.

Olga Manakina, Igor Bogdanov, and Chung-Horng Lung. 2025. Delay-of-gratification as a multi-agent survival micro-benchmark for long-horizon LLMs: Social exposure, personas, and tool use budgets. In *Workshop on Multi-Turn Interactions in Large Language Models (MTI-LLM) at NeurIPS*.

William Merrill and Ashish Sabharwal. 2024. The expressive power of transformers with chain of thought. In *International Conference on Learning Representations*.

Wayne Nelson. 1982. *Applied Life Data Analysis*. John Wiley & Sons.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

James M. Robins and Andrea Rotnitzky. 1992. Recovery of information and adjustment for dependent censoring using surrogate markers. In Nicholas P. Jewell, Klaus Dietz, and Vernon T. Farewell, editors, *AIDS Epidemiology: Methodological Issues*, pages 297–331. Birkhäuser, Boston, MA.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *International Conference on Learning Representations*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and BIG-Bench Collaboration. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. Proofwriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13894–13908.

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

James W. Vaupel, Kenneth G. Manton, and Eric Stallard. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024a. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Dongxu Zhang, Yujun Wu, Yiding Sun, Jinnan Yang, Ning Yang, Jihua Zhu, Miao Xin, and Baoliang Tian. 2025. Not all errors are created equal: ASCoT addresses late-stage fragility in efficient LLM reasoning. *arXiv preprint arXiv:2508.05282*.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. Processbench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1024.

Chenghua Zhu, Siyan Wu, Xiangkang Zeng, Zishan Xu, Zhaolu Kang, Yifu Guo, Yuquan Lu, Junduan Huang, and Guojing Zhou. 2026. EDIS: Diagnosing LLM reasoning via entropy dynamics. *arXiv preprint arXiv:2602.01288*.

A Complete Weibull Parameters

Table 8 reports the Weibull shape ($\hat{\rho}$) and scale ($\hat{\lambda}$) parameters for all 35 dataset–model cells. A shape parameter $\rho > 1$ indicates increasing hazard (aging), $\rho = 1$ constant hazard, and $\rho < 1$ decreasing hazard (Lindy effect). Confidence intervals are profile-likelihood 95% CIs; z-tests evaluate $H_0 : \rho = 1$.

B Distribution Robustness

Table 9 compares AIC across three survival distributions: Weibull, LogNormal, and LogLogistic. While the Weibull is AIC-preferred in only 4/35 cells (LogNormal: 21, LogLogistic: 10), the key finding—increasing hazard—is robust across distributional choices. The Weibull predicts increasing hazard in 34/35 cells; all three distributions agree on increasing hazard in 9/35 cells, concentrated in the two strongest-aging tasks (ProofWriter: 5/7, GSM8K: 2/7). The ranking of tasks by aging magnitude is consistent across distributions: logic $>$ state tracking $>$ math $>$ multi-domain.

C Censoring and Frailty Controls

Censoring Diagnostics. Table 10 reports the Spearman correlation between chain length and censoring status for each dataset. Three datasets (BBH Tracking, MMLU-Pro, ProofWriter) exhibit statistically significant informative censoring, motivating the IPCW correction.

IPCW Adjustment. We apply inverse probability of censoring weighting to correct for informative censoring. Across all 35 cells, the mean absolute change in $\hat{\rho}$ is 6.2% (max 34.1%). 22/35 cells change by <5%, and 27/35 by <10%. Critically, all 34 cells with $\rho > 1$ under standard estimation retain $\rho > 1$ after IPCW correction. The qualitative aging finding is unaffected.

Shared Frailty. Table 11 reports the EM algorithm estimates for a gamma shared frailty model at the dataset level. The frailty variance $\hat{\theta}$ converges to zero for all five datasets, indicating negligible unobserved problem-level heterogeneity in the marginal model. The shape parameter $\hat{\rho}$ remains virtually unchanged after frailty correction ($|\Delta\rho| < 0.1\%$ in all cases), and all datasets retain $\rho > 1$.

D Verifier Audit

We conduct a five-pronged audit of our programmatic step-level verifier to assess potential positional bias.

Step-1 Anchoring. A biased verifier might systematically flag step 1 as erroneous. Table 12 shows that step-1 error rates are low for 3/5 datasets (0.6%–4.1%), supporting verifier validity. BBH Tracking (52.7%) and MMLU-Pro (13.8%) show elevated rates attributable to task structure rather than verifier artifact.

End-of-Chain Effect. We compare the error rate at the final step with pass-through (non-terminal) steps. End-of-chain bias is absent for GSM8K and MMLU-Pro. BBH Tracking and ProofWriter show elevated end-step error rates, consistent with genuine task difficulty at chain termination rather than verifier artifact: ProofWriter chains must combine all accumulated derivations at the final step, and BBH requires full-state recall.

Error Recovery. If the verifier were simply propagating errors from the first mistake, we would observe zero recovery. Instead, GSM8K (62.3% chain-level recovery) and MathQA (68.9%) show substantial recovery rates, confirming that the verifier evaluates each step independently.

Overall Assessment. 2/5 datasets (GSM8K, MMLU-Pro) show no verifier-level positional bias. 3/5 datasets show end-of-chain effects attributable to task structure. No dataset shows patterns consistent with verifier artifact, and recovery rates confirm step-independent evaluation.

E Error Pattern Analysis

Table 13 reports the error type taxonomy across datasets. We classify errors as: *terminal* (no correct steps after the error), *cascade* (error propagates to subsequent steps), and *isolated* (surrounded by correct steps). The late/early density ratio measures whether errors concentrate in the second half of chains relative to the first.

ProofWriter stands apart: errors are overwhelmingly late-concentrated (L/E = 7.37), isolated (89.2%), and terminal (95.0%), with minimal cascading (9.9%). This is consistent with the fan-in hypothesis: errors arise at the compositional integration point rather than propagating from earlier mistakes. In contrast, BBH Tracking and MMLU-

Dataset	Model	n	Events	$\hat{\rho}$	95% CI	$\hat{\lambda}$	z	Status
BBH Track.	Haiku	500	465	1.391	[1.308, 1.474]	1.51	9.21	Aging
	Gemini Flash	342	323	1.833	[1.664, 2.002]	4.58	9.67	Aging
	Gemma 9B	500	186	3.762	[3.287, 4.237]	7.95	11.40	Aging
	Llama 3B	500	456	1.664	[1.537, 1.791]	4.16	10.22	Aging
	Llama 70B	500	258	0.772	[0.690, 0.855]	7.27	−5.41	Lindy
	Qwen 14B	500	458	1.224	[1.146, 1.302]	1.81	5.63	Aging
	Qwen 7B	500	445	1.085	[1.011, 1.158]	2.21	2.26	Aging
GSM8K	Haiku	796	462	1.854	[1.726, 1.982]	6.89	13.03	Aging
	Gemini Flash	752	127	1.293	[1.089, 1.498]	20.40	2.81	Aging
	Gemma 9B	750	132	1.307	[1.119, 1.494]	22.68	3.20	Aging
	Llama 3B	882	348	1.315	[1.206, 1.424]	9.92	5.65	Aging
	Llama 70B	708	93	1.376	[1.120, 1.632]	24.74	2.88	Aging
	Qwen 14B	844	94	1.589	[1.319, 1.860]	23.96	4.27	Aging
	Qwen 7B	680	111	1.711	[1.432, 1.990]	14.78	5.00	Aging
MathQA	Haiku	769	700	1.395	[1.323, 1.468]	4.47	10.64	Aging
	Gemini Flash	230	208	1.743	[1.567, 1.918]	3.75	8.30	Aging
	Gemma 9B	765	700	1.776	[1.676, 1.875]	6.47	15.24	Aging
	Llama 3B	763	756	1.460	[1.394, 1.526]	4.39	13.66	Aging
	Llama 70B	755	698	1.566	[1.483, 1.648]	4.74	13.38	Aging
	Qwen 14B	724	715	1.665	[1.580, 1.751]	4.41	15.27	Aging
	Qwen 7B	723	717	1.813	[1.718, 1.907]	4.12	16.87	Aging
MMLU-Pro	Haiku	376	243	1.300	[1.168, 1.432]	6.92	4.45	Aging
	Gemini Flash	164	132	1.684	[1.454, 1.914]	4.76	5.82	Aging
	Gemma 9B	350	272	1.428	[1.292, 1.565]	7.35	6.16	Aging
	Llama 3B	349	307	1.181	[1.092, 1.270]	5.32	4.01	Aging
	Llama 70B	357	294	1.319	[1.203, 1.434]	4.48	5.40	Aging
	Qwen 14B	357	278	1.255	[1.142, 1.368]	4.68	4.43	Aging
	Qwen 7B	348	272	1.292	[1.175, 1.409]	5.11	4.90	Aging
ProofWriter	Haiku	500	247	2.850	[2.580, 3.119]	8.25	13.44	Aging
	Gemini Flash	195	29	2.591	[1.922, 3.260]	7.56	4.66	Aging
	Gemma 9B	499	269	2.153	[1.960, 2.345]	13.45	11.74	Aging
	Llama 3B	497	338	1.558	[1.436, 1.681]	15.30	8.93	Aging
	Llama 70B	494	187	2.633	[2.349, 2.916]	12.10	11.28	Aging
	Qwen 14B	500	255	2.697	[2.456, 2.939]	7.49	13.78	Aging
	Qwen 7B	500	288	2.018	[1.852, 2.185]	11.03	12.00	Aging

Table 8: Weibull parameters for all 35 dataset–model cells. All z-tests significant at $p < 0.05$; all $p < 0.025$ except Qwen 7B on BBH ($p = 0.024$). 34/35 cells show $\rho > 1$ (aging); the sole exception is Llama 70B on BBH Tracking ($\rho = 0.77$, Lindy effect).

Pro show high cascade rates (75–79%), indicating error propagation as the dominant failure mode.

F Mann-Kendall Trend Tests

Table 14 reports non-parametric Mann-Kendall tests for monotonic trend in the empirical hazard rate $\hat{h}(t)$ across step positions, with Benjamini-Hochberg FDR correction at $\alpha = 0.05$.

G Experimental Details

Models. We evaluate seven models from four families: Claude Haiku 4.5 (Anthropic), Gemini 2.5 Flash (Google), Gemma 2 9B (Google), Llama 3.2 3B and Llama 3.3 70B (Meta), Qwen 2.5 7B and Qwen 2.5 14B (Alibaba). All inference uses greedy decoding (temperature 0) with structured chain-of-thought prompting.

Benchmarks. Five benchmarks spanning four reasoning categories: GSM8K (Cobbe et al., 2021) and MathQA (Amini et al., 2019) (mathematical), ProofWriter (Tafjord et al., 2021) (logical deduction), BBH Tracking (Srivastava et al., 2023) (state tracking), and MMLU-Pro (Wang et al., 2024b) (multi-domain knowledge).

Step-Level Verification. Each reasoning step is verified programmatically against ground-truth intermediate states. For mathematical benchmarks, we check numerical equivalence at each step. For ProofWriter, we verify logical derivation validity. For BBH Tracking, we compare predicted state against the ground-truth state table. For MMLU-Pro, we use structured answer extraction with semantic matching. Steps that cannot be parsed are coded as correct (conservative), ensuring that any detected errors reflect genuine reasoning failures.

Survival Analysis Setup. We model "time" as the discrete step index within a chain. An "event" (failure) is the first step at which a verified error occurs. Chains completing without error are right-

Dataset	Model	AIC_W	AIC_{LN}	AIC_{LL}	Best	ΔAIC	All $\uparrow$
BBH Track.	Haiku	1159.7	731.4	323.2	LL	836.5	
	Gemini Flash	1446.0	1501.8	1524.8	W	0.0	✓
	Gemma 9B	1139.0	1245.9	1169.7	W	0.0	✓
	Llama 3B	2017.9	2046.4	2085.5	W	0.0	
	Llama 70B	1452.3	1361.1	1395.9	LN	91.2	
	Qwen 14B	1379.7	1059.1	876.6	LL	503.1	
	Qwen 7B	1574.9	1321.5	1246.3	LL	328.6	
GSM8K	Haiku	2653.4	2558.0	2566.3	LN	95.4	
	Gemini Flash	1112.4	1094.6	1107.4	LN	17.8	
	Gemma 9B	1178.7	1151.0	1170.3	LN	27.7	
	Llama 3B	2385.1	2278.9	2319.8	LN	106.2	
	Llama 70B	876.0	863.9	873.8	LN	12.1	
	Qwen 14B	911.3	898.3	908.2	LN	13.0	✓
	Qwen 7B	950.1	939.2	947.0	LN	10.9	✓
MathQA	Haiku	3281.4	3035.8	3025.8	LL	255.6	
	Gemini Flash	859.1	808.4	808.6	LN	50.7	
	Gemma 9B	3618.2	3509.1	3529.9	LN	109.1	
	Llama 3B	3393.8	3030.3	3018.4	LL	375.4	
	Llama 70B	3256.8	3049.7	3064.6	LN	207.1	
	Qwen 14B	3117.3	2944.0	2951.5	LN	173.3	
	Qwen 7B	2944.6	2794.4	2817.0	LN	150.2	
MMLU-Pro	Haiku	1429.3	1394.6	1407.3	LN	34.7	
	Gemini Flash	634.8	627.7	634.0	LN	7.1	
	Gemma 9B	1585.3	1575.7	1586.8	LN	9.6	
	Llama 3B	1601.1	1489.5	1489.6	LN	111.6	
	Llama 70B	1426.0	1358.6	1370.5	LN	67.4	
	Qwen 14B	1390.0	1325.3	1341.4	LN	64.7	
	Qwen 7B	1401.6	1341.0	1354.5	LN	60.6	
ProofWriter	Haiku	1437.5	1573.3	1486.5	W	0.0	✓
	Gemini Flash	214.5	202.6	209.0	LN	11.9	✓
	Gemma 9B	1867.8	1849.4	1849.3	LL	18.5	✓
	Llama 3B	2444.5	2414.8	2412.9	LL	31.6	
	Llama 70B	1314.6	1336.3	1311.8	LL	2.8	✓
	Qwen 14B	1432.9	1420.8	1412.1	LL	20.8	✓
	Qwen 7B	1878.0	1850.6	1841.7	LL	36.3	

Table 9: AIC comparison across Weibull (W), LogNormal (LN), and LogLogistic (LL) distributions. ΔAIC: difference between Weibull AIC and best-fitting AIC. "All $\uparrow$": all three distributions agree on increasing hazard. Summary: W best in 4/35, LN in 21/35, LL in 10/35. Despite distributional preference, increasing hazard is detected by the Weibull in 34/35 cells.

Dataset	n	Cens.%	ρ_s	Inform.
BBH Track.	3,342	22.5	-0.155	Yes
GSM8K	5,412	74.7	-0.038	No
MathQA	4,729	5.0	0.017	No
MMLU-Pro	2,301	21.9	-0.163	Yes
ProofWriter	3,185	49.4	-0.183	Yes

Table 10: Censoring diagnostics per dataset. ρ_s: Spearman correlation between chain length and censoring indicator. "Inform.": informative censoring detected ($|\rho_s| > 0.10$ and $p < 0.001$).

Dataset	$\hat{\rho}_{\text{marg}}$	$\hat{\rho}_{\text{frailty}}$	$\hat{\theta}$	Aging
BBH Track.	1.057	1.057	0.0	Yes
GSM8K	1.470	1.472	0.0	Yes
MathQA	1.549	1.549	0.0	Yes
MMLU-Pro	1.269	1.269	0.0	Yes
ProofWriter	1.801	1.801	0.0	Yes

Table 11: EM gamma shared frailty results per dataset. $\hat{\theta}$: estimated frailty variance (0.0 = negligible heterogeneity). Aging is preserved in all cases.

censored at their terminal step. Weibull parameters are estimated via maximum likelihood with profile-likelihood confidence intervals. IPCW weights are computed using a logistic regression model for the censoring probability, conditioned on chain length and dataset difficulty proxies. Shared frailty models use EM estimation with gamma-distributed frailty.

Software. All analyses use Python 3.11 with `lifelines` 0.28 for survival analysis, `scipy` 1.12 for statistical tests (including Kendall's τ for Mann-Kendall trend tests), and `matplotlib`/`seaborn` for visualization. Benjamini-Hochberg correction follows the standard step-up procedure at $\alpha = 0.05$.

Computation. All experiments run on a single workstation. Inference for all 18,969 chains completes in approximately 48 hours across all models.

Dataset	Step-1 %	End Bias	Recov. %
BBH Track.	52.7	Yes	8.4
GSM8K	0.6	No	62.3
MathQA	4.1	Yes*	68.9
MMLU-Pro	13.8	No	43.4
ProofWriter	0.9	Yes	8.6

Table 12: Verifier audit summary. Step-1 %: fraction of chains with step-1 error. End Bias: significant end-of-chain error elevation (*mild, $z = 3.3$). Recov. %: fraction of error chains showing at least one error-to-correct recovery.

Dataset	Term.%	Casc.%	Isol.%	L/E
BBH Track.	94.7	78.8	20.3	1.28
GSM8K	42.4	43.7	48.6	1.55
MathQA	39.3	54.5	26.3	1.47
MMLU-Pro	66.7	75.3	23.8	1.58
ProofWriter	95.0	9.9	89.2	7.37

Table 13: Error type analysis per dataset. Term.: terminal errors (no recovery). Casc.: cascade errors (propagation to subsequent steps). Isol.: isolated errors (correct neighbors). L/E: late-half to early-half error density ratio. ProofWriter's extreme L/E ratio (7.37) reflects the fan-in hypothesis: errors concentrate at steps requiring compositional integration.

Statistical analyses (Weibull fitting, IPCW, frailty, Cox regression) complete in under 2 hours total.

Dataset	Model	τ	p	p_{BH}	Dir.	Sig.
BBH Track.	Haiku	−0.120	.782	.943	↓	
	Gemini Flash	0.048	1.00	1.00	↑	
	Gemma 9B	0.238	.562	.907	↑	
	Llama 3B	0.429	.239	.550	↑	
	Llama 70B	0.056	.867	.979	↑	
	Qwen 14B	−0.200	.817	.953	↓	
	Qwen 7B	0.048	1.00	1.00	↑	
GSM8K	Haiku	−0.156	.601	.907	↓	
	Gemini Flash	−0.592	.028	.097	↓	
	Gemma 9B	−0.478	.021	.083	↓	
	Llama 3B	−0.555	.009	.043	↓	*
	Llama 70B	−0.022	1.00	1.00	↓	
	Qwen 14B	−0.295	.189	.508	↓	
	Qwen 7B	−0.111	.761	.943	↓	
MathQA	Haiku	−0.128	.590	.907	↓	
	Gemini Flash	0.067	1.00	1.00	↑	
	Gemma 9B	0.410	.057	.182	↑	
	Llama 3B	−0.273	.283	.550	↓	
	Llama 70B	−0.127	.648	.907	↓	
	Qwen 14B	−0.127	.648	.907	↓	
	Qwen 7B	0.333	.260	.550	↑	
MMLU-Pro	Haiku	−0.091	.761	.943	↓	
	Gemini Flash	0.467	.272	.550	↑	
	Gemma 9B	0.152	.545	.907	↑	
	Llama 3B	−0.346	.165	.480	↓	
	Llama 70B	−0.761	.005	.033	↓	*
	Qwen 14B	−0.714	.014	.062	↓	
	Qwen 7B	0.222	.477	.878	↑	
ProofWriter	Haiku	0.833	.001	.010	↑	**
	Gemini Flash	0.183	.718	.943	↑	
	Gemma 9B	0.647	<.001	.002	↑	**
	Llama 3B	0.167	.213	.533	↑	
	Llama 70B	0.564	.007	.039	↑	*
	Qwen 14B	0.833	.001	.010	↑	**
	Qwen 7B	0.626	.001	.011	↑	**

Table 14: Mann-Kendall trend tests on empirical hazard rates. τ: Kendall's tau statistic. p_{BH}: Benjamini-Hochberg corrected p-value. $*$: significant at BH-FDR $\alpha = 0.05$; $**$: significant at $\alpha = 0.01$. ProofWriter shows the strongest and most consistent increasing trend (5/7 models significant after correction), consistent with its high Weibull ρ. Two decreasing trends survive correction (GSM8K Llama 3B, MMLU-Pro Llama 70B), reflecting non-monotonic empirical hazards that the parametric Weibull nonetheless fits as aging.