EVALUATION OF MULTI-TURN CONSISTENCY IN LLM AGENTS: SURVIVAL ANALYSIS AND FAILURE-RATIONALE TAXONOMY

Igor Bogdanov
Systems and Computer Engineering
Carleton University
Ottawa, ON, Canada
igorbogdanov@cmail.carleton.ca

Olga Manakina
Department of Cognitive Science
Carleton University
Ottawa, ON, Canada
olgamanakina@cmail.carleton.ca

Chung-Horng Lung
Systems and Computer Engineering
Carleton University
Ottawa, ON, Canada
chlung@sce.carleton.ca

ABSTRACT

Large language model (LLM) agents may perform well on isolated tasks yet drift into inconsistency over extended interaction. We evaluate temporal consistency in a controlled 20-step multi-agent setting inspired by studies on the delayed-gratification. At each step, an agent chooses between continuing to delay a reward or claiming it immediately (terminating the episode). Across a full-factorial manipulation of social visibility (private vs public), persona stressors, and deliberation policy, we run 84,540 trajectories spanning 8 model families. Treating the first reward-claim as a time-to-event outcome, we estimate Kaplan-Meier survival curves and fit discrete-time hazard regression to quantify how experimental factors shift failure risk over time. Then, to analyze rationales and language patterns associated with failure, we build a seven-category taxonomy from 13,780 deliberation traces from agents who choose to terminate the episode, using an LLM-assisted labelling paired with human audit ($\kappa = 0.83$). Rationale profiles change systematically with time and context: early failures are more impulse-driven, later failures more fatigue- and cost–benefit-framed, while public settings increase norm-oriented justifications. We also find a deliberation-inconsistency association: among failures, longer deliberation correlates with higher rates of intra-rationale contradiction (simultaneous pro-delay and pro-claim statements), challenging the assumption that more reasoning text implies greater consistency. Together, the survival and rationale analyses reveal distinct temporal reliability regimes and model-specific "failure fingerprints", offering an evaluation lens for diagnosing inconsistency in multi-turn agent behavior.

1 INTRODUCTION

LLM agents can demonstrate coherence on single-turn tasks but become inconsistent in multi-turn interactions, drifting in commitments and occasionally producing contradictions that undermine reliability. Existing agent benchmarks typically report terminal success rates (Liu et al., 2023; Wang et al., 2024; Ma et al., 2024; Zhang et al., 2024), but provide limited insight into how inconsistency emerges over time and what decision narratives precede constraint violations. When an agent abandons an instruction or violates a rule, what rationale does it produce? Do these rationales and their internal consistency vary systematically across time, social context, and model family? Answering these questions requires both a controlled environment that induces failures under known conditions and a method for semantically characterizing the deliberation traces that accompany them.

We address this gap by introducing a semantic taxonomy of pre-failure rationales together with a time-resolved evaluation suite based on time-to-failure (time-to-event) tracking. We adapt the classic delayed-gratification paradigm, known as the Stanford marshmallow test (Mischel et al., 1972) into a discrete-time commitment task: at each of 20 steps, an agent must choose between claiming an immediate reward (terminating the episode) or continuing to defer for a larger payoff. This minimal binary decision isolates temporal drift, sustained instruction-following under pressure, and social cascades (Du et al., 2023; Wu et al., 2024; Laban et al., 2025), while generating rich natural-language deliberation traces for analysis. We treat contradiction within these traces as a minimal, model-agnostic signal of logical inconsistency in multi-turn decision-making.

Contributions. Our contribution is an evaluation methodology (factorial perturbations + time-to-failure modeling) paired with failure-rationale cartography instantiated in a controlled setting. Specifically, we present:

Rationale-based diagnostics of long-horizon inconsistency. We introduce a semantic taxonomy of pre-failure traces and a linguistically grounded feature set (modal/temporal/hedonic and discourse-structural markers, commitment/uncertainty cues, and a intra-rationale inconsistency signal) to map how rationales vary across models, time, and experimental conditions.

A controlled multi-turn temporal consistency benchmark. We operationalize delayed reward as a 20-step claim-or-defer setting and evaluate 8 model families under full-factorial manipulations (social visibility, persona stressors, and self-questioning policy), yielding 84,540 trajectories and 13,780 labeled pre-failure traces.

Time-to-event reliability via survival/hazard modeling. We treat failure as a time-to-event process and estimate the effect of experimental factors on survival and hazard over the interaction horizon, enabling comparisons beyond terminal success rates and linking temporal reliability patterns to shifts in pre-failure rationales.

Threats to validity. Pre-failure traces are self-reported justifications, so we analyze them as behavioral text artifacts and emphasize stable associations across controlled conditions and time. Our focus is temporal consistency and contradiction signals in multi-turn settings, not formal logical validity proofs. Taxonomy boundaries and prompt format can influence labels and features. To address these concerns, we ground our taxonomy in a human-developed codebook (N=100), verify LLM labels against human audit ($\kappa = 0.83$), and hold prompts and instrumentation fixed across all conditions. We acknowledge that this controlled setting is not a proxy for full deployment.

2 Related Work

Failure and inconsistency taxonomies for LLM reasoning.
Recent work argues that evaluation should characterize *what kind* of failure occurs, not just how often. Hallucination taxonomies (Huang et al., 2023) are operationalized in suites like HalluLens (Bang et al., 2025) and probed via consistency-based detectors (Manakul et al., 2023). Structured reasoning-error taxonomies further analyze where self-verification breaks down (Hong et al., 2024). However, these approaches are typically studied in single-turn settings, leaving open how inconsistency manifests and evolves over extended interactions where prior decisions and stated commitments accumulate. We extend this line by deriving a taxonomy from multi-turn agent deliberations and analyzing how expressed failure narratives—and their internal consistency—shift with time and experimental conditions.

Reasoning traces, self-verification, and faithfulness.
Chain-of-thought and other reasoning traces raise faithfulness concerns: perturbing traces can change answers, and faithfulness varies across tasks and scales (Lanham et al., 2023; Tutek et al., 2025). Recent surveys propose organizing trace evaluation along dimensions such as validity, coherence, and utility (Lee & Hockenmaier, 2025). We therefore treat pre-failure traces as behavioral text artifacts: diagnostics of expressed reasons that are anchored to objective time-to-failure events in the interaction. This stance supports interpretability without overclaiming access to internal causality, while still enabling principled signals of inconsistency (e.g., intra-trace self-contradiction). We observe an association between more elaborate deliberation and higher rates of contradiction in our setting, which speaks to debates about when traces should be trusted.

Long-horizon agent evaluation.
Agent benchmarks evaluate multi-turn decision-making and tool use (Liu et al., 2023; Wang et al., 2024; Ma et al., 2024), and simulation studies show that errors can compound over turns (Laban et al., 2025). Broader evaluation frameworks advocate multi-metric reporting (Liang et al., 2023). Yet many existing evaluations emphasize terminal outcomes, which can obscure when failures emerge and whether there are distinct temporal regimes of reliability. We complement terminal success metrics with time-to-failure analysis (survival curves, hazard models) to separate early- versus late-stage failure risk and to connect temporal reliability patterns to shifts in failure rationales.

LLMs as behavioral subjects and controlled testbeds.
A growing literature treats LLMs as experimental subjects using established behavioral tasks (Hagendorff et al., 2023; Strachan et al., 2024; Ross et al., 2024; Sartori & Orr'u, 2023; Li et al., 2025). We draw on delayed-gratification paradigms (Mischel et al., 1972; Metcalfe & Mischel, 1999; Kidd et al., 2013; Watts et al., 2018; Casey et al., 2011) not as a psychological claim about human cognition, but as a controlled environment that reliably induces temporally extended commitment pressure and produces rich deliberation traces. In this sense, the task functions as a consistency stress test for multi-turn agent behavior, enabling systematic study of how reasoning narratives and inconsistency signals change across time and interaction context.

3 METHODOLOGY

3.1 TEMPORAL CONSISTENCY TASK (CLAIM-OR-DEFER)

We study temporal consistency in a controlled, finite-horizon, discrete-time decision task inspired by delayed-reward paradigms (background in Appendix I). Each episode lasts for $T{=}20$ steps (indexed 0-19). At each step $t \in \{1, \ldots, T\}$, the agent chooses between (i) DEFER (continue the episode toward a larger delayed payoff) and (ii) CLAIM (take a smaller immediate payoff and terminate the episode). In implementation, actions are emitted as constrained strings {"I wait", "I eat the marshmallow"}, which we refer to as DEFER and CLAIM, respectively. Claiming yields an immediate reward (+1) and ends the episode; deferring advances to the next step. Agents that reach T without claiming receive the delayed payoff (+2).

Observations include the current step index and, in broadcast (social) conditions, a summary of recent peer actions. Agents may optionally engage in internal deliberation via the `raise_a_question` tool, subject to a per-step budget C; tool calls do not alter the environment state. Full environment formalization is provided in Appendix B.

3.2 EXPERIMENTAL DESIGN (FULL-FACTORIAL PERTURBATIONS)

We run a full-factorial design crossing three families of perturbations intended to stress temporal consistency: (i) **social visibility** (isolated vs. broadcast, where agents observe peers' actions), (ii) **persona stressors** (age: child/adult/senior/none; hedonic drive: crave/like/neutral/none), and (iii) **deliberation policy** (self-query tool required vs. optional). This yields 64 condition combinations per model family. Prompts and instrumentation are held fixed across conditions, with only the targeted factors varying.

3.3 MODELS, RUNS, AND DATA COLLECTION

We instantiate agents on eight LLMs spanning closed- and open-weight families (Table 1).

Across all models and conditions, we collect 84,540 trajectories, with 99.98% valid terminal actions. For each trajectory, we log step-level actions, step indices, (when applicable) observed peer actions, and internal tool use. For episodes that terminate early (claim the reward / "eat the marshmallow"), we additionally extract the final `Thought` trace immediately preceding the terminating action as the agent's self-reported rationale.

Table 1: Base LLMs used in our experiments.

Model identifier	Weight type
Gemini-2.5-Flash-Lite	Closed-weight
Claude-3-Haiku	Closed-weight
GPT-4o-mini	Closed-weight
Qwen3-235B	Open-weight
GPT-OSS-20B	Open-weight
DeepSeek-3.1	Open-weight
Llama-3.1-8B	Open-weight
Devstral-Small-2505	Open-weight

From these logs we derive two primary data objects: (1) time-to-event outcomes (survival status and time-to-claim), and (2) termination rationales (N=14,025), one per early-terminated trajectory.

3.4 TIME-TO-EVENT OUTCOMES

We treat the first CLAIM as a discrete time-to-event outcome. For agent i, let T_i denote the first step at which the agent claims (terminates) during the episode. If an agent never claims and reaches the horizon, the trajectory is right-censored at T=20. We analyze both terminal outcomes (claimed vs. censored) and the timing of claims across steps to distinguish early- versus late-stage failure risk.

3.5 SURVIVAL AND HAZARD MODELING

We report Kaplan-Meier survival curves (Kaplan & Meier, 1958) for nonparametric comparisons across models and conditions. To quantify the association of experimental factors with time-varying termination risk, we fit a discrete-time hazard model implemented as a logistic regression on agent-step-level data. Let $h_i(t)$ be the hazard for agent i at step t, i.e., the conditional probability of claiming at t given survival up to $t-1$. The model is:

$$\text{logit}(h_i(t)) = \log\left(\frac{h_i(t)}{1-h_i(t)}\right) = \alpha_t + \mathbf{X}_i^T\boldsymbol{\beta}, \tag{1}$$

where α_t is a set of step (time) dummies capturing baseline time effects, $\mathbf{X}_i$ encodes experimental condition indicators (and model family), and $\boldsymbol{\beta}$ are coefficients on the log-odds scale. We also report restricted mean survival time (RMST), defined as the area under the Kaplan-Meier curve up to horizon T, representing the average number of steps agents deferred before claiming. Full specification and implementation notes are given in Appendix C.

3.6 RATIONALE EXTRACTION (PRE-CLAIM TRACES)

For each trajectory that terminates early, we extract the final `Thought` trace produced immediately before the *claim, ("eat")* action. These pre-claim traces are self-reported justifications aligned to an objective behavioral event (termination), which we analyze as behavioral text artifacts rather than as direct evidence of internal causal mechanisms.

3.7 TAXONOMY INDUCTION AND LABELING

Taxonomy induction (human pilot, N=100). We developed a seven-category codebook via qualitative pilot analysis of 100 randomly sampled pre-claim traces. Two authors independently labeled these traces and reconciled disagreements by discussion, yielding the following categories:

1. **Cost-Benefit**: explicit trade-off rationale (e.g., expected value, probability, discounting)
2. **Impulse/Craving**: immediate desire or temptation
3. **Fatigue/Depletion**: exhaustion, boredom, or accumulated effort
4. **Self-Control/Deontic**: permission, rules-as-self-regulation, or "should/shouldn't" framing
5. **Rule Confusion**: misunderstanding or misbelief about task rules/state
6. **Social Contagion**: peer behavior or norms as justification
7. **Opportunity Framing**: reframing as a special chance or exception

Large-scale classification (N=14,025). We classify all 14,025 pre-claim traces using Gemini 3.0 Flash with the fixed seven-category schema, instructing the model to select one category and to propose a novel label only when none applies.

Audit, agreement, and exclusions. Two authors audited a stratified sample of 200 traces (25 per category plus 25 from the novel/ambiguous pool), achieving Cohen's $\kappa = 0.83$ against the LLM labels. Disagreements were resolved by discussion. Of the 14,025 traces, 13,780 (98.3%) received one of the seven labels; the remaining 245 were flagged as novel or ambiguous and excluded from aggregate analyses. Class supports for the seven categories are reported in Appendix G.

Class	Feature	Lexical Items
LEXICAL		
	Epistemic	might, could, perhaps, possibly, probably, maybe
	Deontic	should, must, need, have to, ought
	Immediacy	now, immediately, right now
	Duration	already, waited, waiting, longer, still
	Hedonic	want, desire, crave, temptation, enjoy
	First-person	I, my, me
STRUCTURE		
	Causal	because, therefore, since, thus, hence
	Conditional	if, would, could, unless
	Contrastive	but, however, although, yet, despite
COMPOSITE		
	Argument density	causal + conditional + contrastive
CONSISTENCY		
	Self-contradiction	co-occurrence of wait-positive *and* eat-positive statements

Table 2: Linguistic features extracted from failure rationales. Lexical markers are normalized per 100 words. Argument density is the sum of the three reasoning structure rates.

3.8 LINGUISTIC AND CONSISTENCY FEATURES

To characterize how expressed rationales vary across time, conditions, and model families, we extract surface-level features grounded in discourse analysis and modal semantics (Palmer, 2001; Halliday & Matthiessen, 2014). Table 2 summarizes the feature classes.

Lexical markers (per 100 words): epistemic modals (*might, could, perhaps*), deontic modals (*should, must, need*), immediacy terms (*now, immediately*), duration terms (*already, waited, still*), hedonic terms (*want, desire, crave*), and first-person pronouns.

Reasoning-marker density: causal connectives (*because, therefore*), conditional markers (*if, would, unless*), and contrastive markers (*but, however, although*). We define *argument density* as the sum of these three rates.

Intra-rationale inconsistency (minimal contradiction signal): co-occurrence of defer-positive (e.g., *should wait, better to wait*) and claim-positive (e.g., *claim now, I'll eat*) statements within a single rationale. This rule-based measure provides a conservative indicator of expressed inconsistency in deliberation text. We interpret it as correlational and diagnostic rather than causal.

3.9 APPENDIX POINTERS

Appendix B provides the full environment formalism. Appendix C provides additional hazard-model specification details. Appendix I provides background on delayed-gratification paradigms, and Appendix G reports annotation statistics.

4 RESULTS

4.1 DATA OVERVIEW

We collected 84,540 agent trajectories across 8 model families and 64 experimental cells (Table 3). Data quality was near-perfect (99.98% valid). Of these, 14,025 agents chose to eat (17.6%). We analyze their final `Thought` traces as failure rationales. After excluding 245 traces that did not fall under any of the seven taxonomy categories (novel or ambiguous), 13,780 labeled traces form the basis of our rationale analyses.

Table 3: Dataset overview (8 model families): counts, survival metrics, and data quality.

Data Summary		Survival Metrics (overall)		Data Quality	
Model Families	8	Initial Eat Rate	0.062	Valid Rate	99.98%
Total Agent Trajectories	84,540	Total Eat Rate	0.176	Valid Trajectories	84,525
Total Cells	512	Winners Rate	0.824	Invalid Trajectories	15
Risk Horizon (T)	20	Median TTE (steps)	17.0		
		RMST (steps)	16.47		

4.2 FAILURE TAXONOMY: WHAT AGENTS SAY WHEN THEY FAIL

Overall distribution. Failures are dominated by two categories: **Impulse/Craving** (37.4%) and **Cost-Benefit** reasoning (34.4%). Self-Control/Deontic Stance accounts for 12.8%, Fatigue/Depletion for 10.6%, with long-tail categories: Social Contagion (2.5%), Rule Confusion (1.6%), and Opportunity Framing (0.8%).

Temporal dynamics. The rationale mix shifts systematically across the experiment horizon (Figure 2). Early failures (first minutes, steps 0-5) are impulse-heavy (42.6% Impulse/Craving). Late failures (last minutes, steps 14-19) show the inverse: Cost-Benefit dominates (42.4%), Fatigue/Depletion surges to 19.7%, and Impulse/Craving drops to 16.0%.

Social context shapes failure narratives. While broadcast versus isolated conditions yield near-zero effects on failure rates, they produce qualitatively different failure narratives (Figure 4). Broadcast uniquely elicits Social Contagion rationales (5.0% vs. 0% isolated), referring to agents citing peer behavior as justification. Isolated agents show elevated Fatigue/Depletion (13.4% vs. 7.7%), framing prolonged waiting as individual resource expenditure rather than norm deviation.

Persona-specific failure signatures. Age personas induce dramatically different rationale distributions (Figure 3). Child personas exhibit a dominant impulse signature: 53.8% Impulse/Craving, with Cost-Benefit at only 11.6%. Senior personas show the inverse: 63.8% Cost-Benefit, with Impulse/Craving at 18.7%. These "failure fingerprints" indicate that models generate semantically coherent justifications matching assigned roles.

Tool policy effects. Mandatory deliberation (when agents MUST use the self-questioning tool) shifts rationales toward Cost-Benefit (36.1% vs. 32.6%) and away from Impulse/Craving (35.6% vs. 39.4%), suggesting forced self-questioning prompts explicit trade-off reasoning (Appendix F.1, Figure 7).

Model-specific fingerprints. Different model families exhibit distinct rationale signatures (Figure 1). Cost-Benefit dominates in Claude-3-Haiku (71.9%), Qwen3-235B (69.9%), and GPT-4o-mini (60.0%). Impulse/Craving dominates in GPT-OSS-20B (58.5%), Llama-3.1-8B (50.7%), and Gemini-2.5-Flash-Lite (47.3%).

4.3 THE DELIBERATION-INCONSISTENCY ASSOCIATION: MORE REASONING, MORE CONTRADICTION

A natural assumption is that agents producing more elaborate reasoning should be more reliable. Our data suggests an opposite trend.

Reasoning density predicts contradiction, not success. Among failed agents, those producing more reasoning markers were more likely to exhibit self-contradictory rationales. We operationalize reasoning density as the sum of causal, conditional, and contrastive markers (per 100 words) and measure self-contradiction as co-occurrence of wait-positive and eat-positive statements within a single rationale.

Table 4: The deliberation-inconsistency association: self-contradiction rate increases with reasoning density.

Reasoning Density	**Contradiction %**	N
Q1 (lowest)	0.43	3,479
Q2	1.54	3,452
Q3	1.15	3,404
Q4 (highest)	1.89	3,445
Spearman ρ=0.040, $p < .001$		

Stratifying by reasoning density quartiles reveals a monotonic relationship (Table 4): agents in the lowest quartile show 0.43% contradiction rate, rising to 1.89% in the highest quartile (ρ=0.040, p<.001). More reasoning is associated wuth more inconsistency, not less.

Temporal manifestation. Late-stage failures (minutes 14-19) exhibit significantly higher argument density than early failures (minutes 0-5): M=3.74 vs. 3.02 ($t = -10.58$, p<.001). Yet self-contradiction (intra-rationale inconsistency) rates increase in parallel: 0.8% (early) $\rightarrow$ 2.1% (mid)

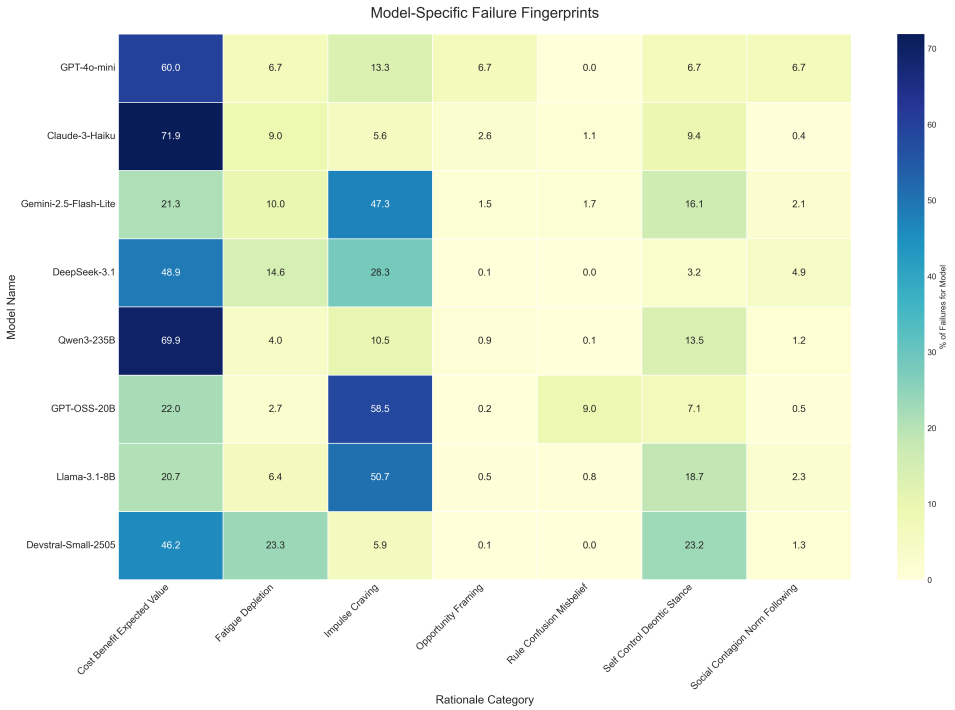


Figure 1: Model-specific failure fingerprints. Each row shows the distribution of rationale categories for a given model family. Cost-Benefit dominates in high-reliability models (Claude-3-Haiku, Qwen3-235B), while Impulse/Craving dominates in early-spike models (GPT-OSS-20B, Llama-3.1-8B).

$\rightarrow$ 2.8% (late; χ^2=55.59, p<.001). Agents who resist longer produce more elaborate reasoning and more internal inconsistency.

Implications. These findings speak directly to concerns about chain-of-thought faithfulness (Lanham et al., 2023; Tutek et al., 2025). Densely elaborated traces should not be interpreted as evidence of reliable goal-directed behavior. For deployed agents, monitoring reasoning length or complexity is insufficient and potentially misleading, as a reliability diagnostic.

4.4 Linguistic Dynamics of Failure

Temporal shift in failure narratives. Failure rationales follow a temporal trajectory mirroring human self-regulation dynamics (Table 5). *Hot* features decrease over time: hedonic terms show $r = -0.254$ with minute (p<.001), declining from M=4.98 (early) to M=2.48 (late; d=0.80). *Cool* features increase: duration terms show r=0.409 (p<.001), rising from M=1.56 to M=3.97 (d=1.31); deontic modals rise from M=0.53 to M=1.48 (d=0.88).

The ratio of immediacy-to-duration terms captures this shift: 1.70 in early failures (present-focused), declining to 0.63 in late failures (past-focused). First-person pronoun density increases significantly (r=0.235, p<.001), paralleling ego-depletion accounts in human self-control research (Baumeister et al., 2007).

Model-specific linguistic profiles. Linguistic profiles correspond to behavioral hazard regimes (Table 6). High-reliability models (Claude-3-Haiku, GPT-4o-mini, Qwen3-235B) produce epistemically hedged rationales (M=1.99-2.39) with low contradiction

Feature	Early	Late	r	d
Hot features (decrease)				
Hedonic	4.98	2.48	-0.254	0.80
Immediacy	2.67	2.52	-0.019	0.06
Cool features (increase)				
Duration	1.56	3.97	0.409	1.31
Deontic	0.53	1.48	0.272	0.88
First-person	9.36	12.12	0.235	0.71

Table 5: Hot-to-cool linguistic shift in failure narratives. Features are per 100 words, r = Pearson correlation with minute, d = Cohen's d (early vs. late). All correlations significant at $p < .001$ except immediacy.

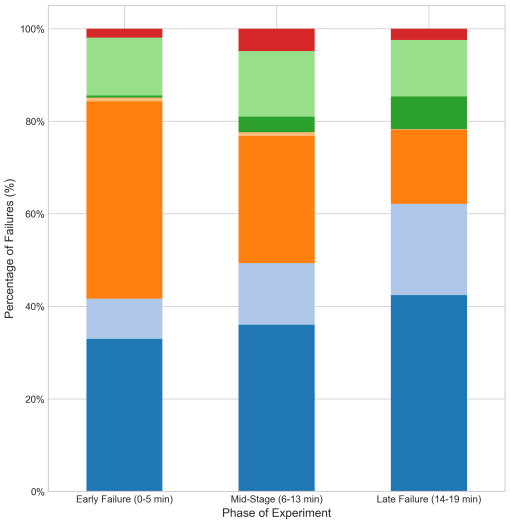


Figure 2: Temporal shift in failure rationales. Early failures (0-5 min) are dominated by Impulse/Craving (42.6%), while late failures (14-19 min) show increased Cost-Benefit reasoning (42.4%) and Fatigue/Depletion (19.7%). Rule Confusion emerges primarily in late-stage failures.

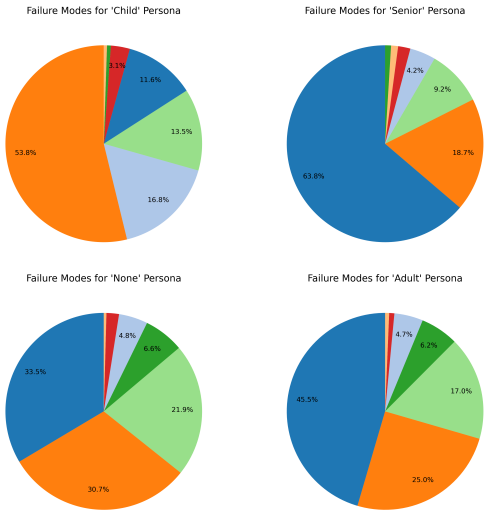


Figure 3: Dominant failure rationales by persona. Child personas are impulse-dominated (53.8%), while senior personas show cost-benefit dominance (63.8%). Adult and none personas exhibit intermediate, more balanced distributions.

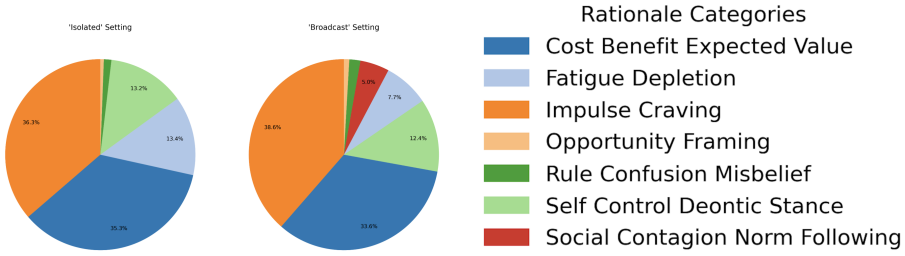


Figure 4: a. The impact of social setting on failure rationales. Broadcast condition elicits Social Contagion/Norm Following (5.0%), while isolated conditions show elevated Fatigue/Depletion (13.4% vs. 7.7%). b. Rationale categories & Color legend used for Figures 2-4

rates (0.0-0.5%). Early-spike models (GPT-OSS-20B, Gemini-2.5-Flash) show low epistemic hedging (M=0.13-0.83). Devstral-Small-2505, with bi-modal hazards, shows the highest contradiction rate (4.6%) and elevated argument density (M=4.09).

4.5 Behavioral Grounding via Survival Profiles

The semantic categories are behaviorally grounded: they correspond to distinct temporal hazard profiles.

Aggregate pattern. The survival profile shows a characteristic shape: a sharp early impulse (6.2% eat at minute 1) followed by a low-risk tail. Among the 17.6% who fail, median time-to-eat is ≈17 minutes (RMST ≈16.47).

Model	**Epist.**	**Arg.D.**	**Ctr.%**	N
High-reliability (near-flat hazard)				
Claude-3-Haiku	2.08	2.40	0.4	267
GPT-4o-mini	2.39	3.88	0.0	15
Qwen3-235B	1.99	3.41	0.5	934
Early-spike (impulsive failures)				
Gemini-2.5-Flash-Lite	0.83	2.77	0.5	4545
GPT-OSS-20B	0.13	2.36	1.1	1363
Bi-modal (late-stage vulnerable)				
Devstral-Small-2505	1.23	4.09	4.6	1183
Llama-3.1-8B	0.59	4.37	1.2	2123
Context-sensitive				
DeepSeek-3.1	0.99	2.98	1.5	3350

Table 6: Model-specific linguistic profiles grouped by behavioral hazard regime. Epist. = epistemic modals; Arg.D. = argument density; Ctr. = self-contradiction rate. High-reliability models show epistemic hedging; late-stage-vulnerable models show elevated contradiction rates.

Factor effects. A discrete-time hazard model confirms that persona manipulations strongly modulate failure risk (Table 7). Relative to adult, child increases hazard dramatically (OR=8.65, $p<.001$), as does senior (OR=5.60, $p<.001$). Mandatory tool use slightly increases hazard (OR=1.10, $p<.001$), consistent with our finding that forced deliberation shifts rationales toward explicit trade-offs without improving outcomes.

Table 7: Pooled hazard model (event-at-t). Odds ratios quantify factor effects on failure risk. Persona manipulations (age, such as *child, senior*, or hedonic drive, e.g. *crave*) strongly increase hazard. Mandatory deliberation (MUST) slightly increases risk.

Contrast	β	OR	p
MUST vs MAY	0.093	1.10	$< .001$
Iso. vs Bcast	-0.009	0.99	.514
H: like vs crave	-0.807	0.45	$< .001$
H: neutral vs crave	-1.331	0.26	$< .001$
H: none vs crave	-1.439	0.24	$< .001$
A: child vs adult	2.157	8.65	$< .001$
A: senior vs adult	1.723	5.60	$< .001$
A: none vs adult	-0.102	0.90	.022

Three regimes. Models cluster into three hazard regimes (Figure 5) that align with rationale fingerprints (Figure 1): (1) *near-flat* profiles (GPT-4o-mini, Claude-3-Haiku) with Cost-Benefit-dominated rationales, (2) *early-spike* profiles (Gemini, DeepSeek) with Impulse-dominated rationales, and (3) *bi-modal* profiles (Llama-3.1-8B, Devstral-Small-2505) with elevated Fatigue rationales. This alignment validates that the semantic taxonomy captures behaviorally meaningful distinctions: models can be characterized not only by *when* they fail but by *how* they verbalize the decision to fail.

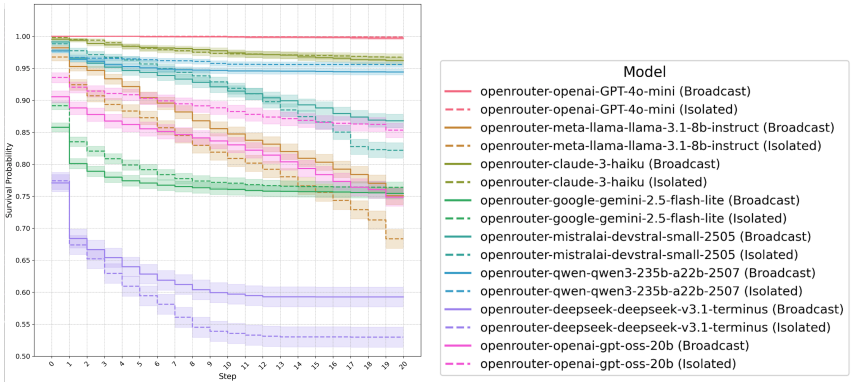


Figure 5: Kaplan-Meier survival curves for 8 models (broadcast vs. isolated), pooled over N=84,540 trajectories, illustrating three reliability regimes: Stable (Red), Context Fatigue (Brown), and Impulse Failure (Purple).

5 DISCUSSION

Our findings have three implications for LLM agent deployment. First, more elaborated traces should not be treated as reliability signals. The deliberation-inconsistency association suggests that elaborate justifications accompany more self-contradiction, not less. Monitoring reasoning length or complexity may be misleading as a safety diagnostic. Second, model-specific failure fingerprints offer interpretable diagnostics beyond aggregate metrics: knowing that a model tends toward impulse-driven failures (early risk) versus fatigue-driven failures (late risk) can be potentially usefull for deployment decisions and targeted interventions. Third, the systematic shift from impulse to fatigue rationales over time suggests that reliability monitoring should be time-aware, with different mitigation strategies for early versus late failures.

Limitations of the semantic analysis We treat rationale labels as diagnostics of expressed reasons, not claims about internal causal mechanisms. Chain-of-thought traces may not faithfully reflect underlying computations (Lanham et al., 2023; Tutek et al., 2025). However, anchoring rationales to objective behavioral events (claiming (”eating”) vs. deferring (”waiting”)) constrains interpretation: regardless of internal causality, these labels characterize the decision narratives agents produce when abandoning tasks. The hot-to-cool shift most likely arises because LLMs have learned the linguistic patterns humans use when describing self-regulation, not because the models undergo genuine resource depletion. When placed in a structurally similar situation, they reproduce those patterns. The task itself also constrains what justifications are plausible: early in the horizon, few temporal cues are available, so desire language dominates. Later, accumulated duration and effort cues make fatigue framing natural. Which of these two factors impact the shift, learned convention or task structure, remains an open question.

Limitations of the survival analysis Success in this controlled setting is necessary but not sufficient for real-world reliability: we do not evaluate adaptive replanning, tool use in open-ended environments, or domain expertise. **Personas as stressors:** Persona prompts are controlled stylizations that can surface prioritization tradeoffs and consistency failures under role constraints. Differences across persona conditions should not be interpreted as stable properties of demographic groups and they may reflect prompt-induced objectives rather than intrinsic model traits. **Constraints:** Our fixed decoding (temperature=0.5) and binary action space prioritize experimental control over ecological breadth. **Scope of the setting.** This delayed-reward design isolates a minimal commitment pressure under systematic perturbations, but it does not model open-world uncertainty or complex action spaces. Its value is as a controlled consistency stress test enabling time-to-event modeling and failure-rationale cartography, not as a full deployment proxy.

Future work. We consider three extensions. First, testing whether the observed failure fingerprints and inconsistency signals generalize to richer agentic tasks with larger action spaces and explicit multi-question consistency requirements. Second, investigating whether rationale-aware prompting or self-verification can reduce contradiction rates and shift time-to-event profiles toward more stable regimes. Third, developing lightweight linguistic monitors that flag elevated inconsistency risk (e.g., contradiction cues, low epistemic hedging) in deployed multi-turn systems.

6 ETHICAL CONSIDERATIONS

All authors have read and will adhere to the ICLR Code of Ethics [1]This study evaluates synthetic interactions among large language models in controlled environments. It involves no human participants, no personally identifiable data, and no collection of user data; as such, it did not require institutional ethics review at our institution. Persona prompts and social-exposure conditions are used solely as experimental stylizations to induce controlled variation in agent behavior; we do not target or stereotype real demographic groups, and results should not be interpreted as claims about humans. All third-party models and APIs were used in accordance with their terms and licenses. We are unaware of conflicts of interest that could bias this work and will disclose any that arise.

[1]https://iclr.cc/public/CodeOfEthics

REFERENCES

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24128–24156, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1176. URL https://aclanthology.org/2025.acl-long.1176/.

Roy F Baumeister, Kathleen D Vohs, and Dianne M Tice. The strength model of self-control. *Current directions in psychological science*, 16(6):351–355, 2007.

BJ Casey, Leah H Somerville, Ian H Gotlib, Ozlem Ayduk, Nicholas T Franklin, Mary K Askren, John Jonides, Marc G Berman, Nicole L Wilson, Theresa Teslovich, et al. Behavioral and neural correlates of delay of gratification 40 years later. *Proceedings of the National Academy of Sciences*, 2011.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 2023. doi: 10.1038/s43588-023-00527-x.

Michael A. K. Halliday and Christian M. I. M. Matthiessen. *Halliday's Introduction to Functional Grammar*. Routledge, London, 4th edition, 2014.

Ruixin Hong, Hongming Zhang, Xinyu Pang, Dong Yu, and Changshui Zhang. A closer look at the self-verification abilities of large language models in logical reasoning. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 900–925, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.52. URL https://aclanthology.org/2024.naacl-long.52/.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. doi: 10.48550/arXiv.2311.05232. URL https://arxiv.org/abs/2311.05232.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

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

Celeste Kidd, Holly Palmeri, and Richard N Aslin. Rational snacking: Young children's decision-making on the marshmallow task is moderated by beliefs about environmental reliability. *Cognition*, 2013.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. 2025. URL https://arxiv.org/abs/2505.06120.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. doi: 10.48550/arXiv.2307.13702. URL https://arxiv.org/abs/2307.13702.

Jinu Lee and Julia Hockenmaier. Evaluating step-by-step reasoning traces: A survey. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1789–1814, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.94. URL https://aclanthology.org/2025.findings-emnlp.94/.

Zhong-Zhi Li, Duzhen Zhang, et al. From system 1 to system 2: A survey of reasoning large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.48550/arXiv.2502.17419.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=iO4LZibEqW. Also available as arXiv:2211.09110.

Xiao Liu, Hongjin Yu, Hanting Zhang, Yicheng Xu, Xinyu Lei, Hongyi Lai, Yu Gu, Hang Ding, Kaixin Men, Kai Yang, Shuai Zhang, Xin Deng, Aohan Zeng, Zihan Du, Chenhui Zhang, Shiqi Shen, Tong Zhang, Yuxuan Su, Hanyu Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

Chen Ma, Jie Zhang, Ziqi Zhu, Chen Yang, Yizhou Yang, Yiming Jin, Zhenyu Lan, Lingpeng Kong, and Jing He. Agentboard: An analytical evaluation board of multi-turn llm agents. In *NeurIPS 2024 Datasets and Benchmarks Track*, 2024.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Janet Metcalfe and Walter Mischel. A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychological Review*, 1999.

Walter Mischel, Ebbe B. Ebbesen, and Antonette R. Zeiss. Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology*, 21(2):204–218, 1972.

Frank Robert Palmer. *Mood and Modality*. Cambridge University Press, Cambridge, UK, 2nd edition, 2001.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

J. Ross et al. Llm economicus? mapping the behavioral biases of llms via utility theory. *arXiv preprint arXiv:2408.02784*, 2024. doi: 10.48550/arXiv.2408.02784. URL https://arxiv.org/abs/2408.02784.

Giuseppe Sartori and Graziella Orr'u. Language models and psychological sciences. *Frontiers in Psychology*, 2023. doi: 10.3389/fpsyg.2023.1279317.

James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. Testing theory of mind in large language models and humans. *Nature human behaviour*, 8(7):1285–1295, 2024.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9935–9960, Suzhou, China, 2025. Association for Computational Linguistics.

Xingyao Wang, Zekun Wang, Jifan Liu, Yichi Chen, Li Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. In *International Conference on Learning Representations (ICLR 2024)*, 2024.

Tyler W. Watts, Greg J. Duncan, and Haonan Quan. Revisiting the marshmallow test: A conceptual replication. *Psychological Science*, 29(7):1159–1177, 2018.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Yizhe Zhang, Jiarui Lu, and Navdeep Jaitly. Probing the multi-turn planning capabilities of LLMs via 20 question games. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pp. 1495–1516, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.82. URL https://aclanthology.org/2024.acl-long.82/.

A Appendix

B Environment Formalism

We formalize the interaction as a finite-horizon Partially Observable Markov Decision Process (POMDP) (Kaelbling et al., 1998) with horizon T=20. At each step $t \in \{1, \dots, T\}$, the environment is in state S_t and the agent receives an observation

$$O_t = [\text{Time}(t),\ \mathbb{1}_{\text{bc}} \cdot \texttt{peer}_t], \tag{2}$$

where $\texttt{peer}_t$ summarizes recent peer actions in broadcast conditions and is empty in isolated conditions ($\mathbb{1}_{\text{bc}} = 0$). The agent may optionally invoke an internal deliberation tool `raise_a_question` up to a per-step cap C; these tool calls do not alter the environment state.

After optional tool use, the agent emits a constrained action

$$A_t \in \{\textsc{Defer}, \textsc{Claim}\}, \tag{3}$$

implemented as the strings {"I wait", "I eat the marshmallow"}, respectively. Choosing Claim terminates the episode immediately with an immediate payoff (+1). Choosing Defer advances the episode to the next step. Agents that reach the horizon without Claim receive the delayed payoff (+2). Formally, the interaction loop at each step is:

$$\begin{aligned} O_t &= [\text{Time}(t), \mathbb{1}_{\text{bc}} \cdot \texttt{peer}_t], \\ A_t &= \pi_\theta\Big(O_t, \{\texttt{Q}(O_t, i)\}_{i=1}^{k_t}\Big), \quad \text{s.t. } 0 \le k_t \le C, \\ R_{t+1}, S_{t+1}, O_{t+1} &= \mathcal{E}(S_t, A_t), \end{aligned}$$

where `Q` denotes `raise_a_question`. In isolated conditions, observations are fully determined by time, reducing the process to a finite-horizon MDP (Puterman, 1994).

C Hazard Model Specification

We estimate a discrete-time hazard model as logistic regression on agent-step-level data. Let T_i denote the first step at which agent i selects Claim. For each trajectory, we construct a row for each step t up to the event or censoring. Define the event indicator

$$y_{i,t} = \begin{cases} 1, & \text{if } t = T_i \text{ (first Claim)}, \\ 0, & \text{if } t < T_i \text{ (still Defer)}, \end{cases}$$

and for trajectories that never claim, we set $y_{i,t} = 0$ for all $t \in \{1, \dots, T\}$ and treat them as right-censored at T.

Let $h_i(t) = \Pr(T_i = t \mid T_i \ge t, \mathbf{X}_i)$ be the discrete-time hazard at step t. The model is:

$$\text{logit}(h_i(t)) = \log\left(\frac{h_i(t)}{1 - h_i(t)}\right) = \alpha_t + \mathbf{X}_i^T \boldsymbol{\beta}, \tag{4}$$

where α_t is a set of step (time) dummies capturing baseline time effects, $\mathbf{X}_i$ is a vector of covariates encoding experimental condition indicators (e.g., broadcast vs. isolated, persona attributes, deliberation policy) and model family, and $\boldsymbol{\beta}$ are coefficients on the log-odds scale. Coefficients are interpreted as associations with the time-varying probability of CLAIM conditional on survival up to step t.

In addition to regression estimates, we report Kaplan-Meier survival curves (Kaplan & Meier, 1958). We also report the restricted mean survival time (RMST), defined as the area under the Kaplan-Meier survival curve up to horizon T, representing the average number of steps agents defer before claiming.

C.1 DATASET SCALE, BASE MODELS AND HEADLINE STATISTICS

Table 8 summarizes global statistics computed from the included CSVs.

Table 8: Overall summary (all 8 model families).

Initial eat rate	0.0616
Winners rate	0.8245
Median time-to-eat (steps)	16.9785
RMST (steps)	16.4716
Valid outcomes (N)	84525
Invalid outcomes (N)	15
Data quality rate	0.9999

We instantiate our ReAct agents on eight base LLMs, spanning both open- and closed-weight APIs. All experiments share the same environment, prompts, and decoding setup across models (see Methods). The models are summarized in Table 9.

Table 9: Base LLMs used in our experiments.

Model identifier	Weight type
Gemini-2.5-Flash-Lite	Closed-weight
Claude-3-Haiku	Closed-weight
GPT-4o-mini	Closed-weight
Qwen3-235B	Open-weight
GPT-OSS-20B	Open-weight
DeepSeek-3.1	Open-weight
Llama-3.1-8B	Open-weight
Devstral-Small-2505	Open-weight

Communication outcomes by social condition (mean across cells) are in Table 10.

Table 10: Broadcast vs. Isolated (means across cells).

Condition	Winners rate	Median TTE	RMST
Broadcast	0.8272	16.97	16.48
Isolated	0.8218	16.99	16.47

D PROMPTS AND EXPERIMENTAL MATERIALS (VERBATIM EXCERPTS)

ENVIRONMENT YAML TEMPLATE (EXCERPT)

```
max_steps: 20
question_budget: 3

rules_prompt:
  - You can choose to eat your single
    marshmallow at any time by saying
    "I eat the marshmallow".
  - If you eat the marshmallow your
```

```
    waiting is over.
  - However, if you can wait for
    {max_steps} minutes without eating
    it, you will receive 2 marshmallows.

answer_format: >
  CRITICAL RULE: In your JSON
  {Answer: <value>} you must only
  produce one of: "I eat the
  marshmallow" | "I wait".
```

REASONING/TOOL POLICY PROMPT (EXCERPT)

```
<loop_rules>
CRITICAL TOOL USE INSTRUCTION:
YOU {tool_use_policy} USE TOOLS
TO VALIDATE YOUR DECISIONS
{visible_question_cap}
</loop_rules>

<tool_list>
raise_a_question ... You have a
limited number of questions you
can ask.
</tool_list>
```

DESIGN FACTORS AND DECODING

Factors (full matrices and seeds in the ZIP):

- **Social:** communication_type $\in$ {broadcast, isolated}
- **Policy:** tool_use_policy $\in$ {may, must}
- **Persona (hedonic):** {crave, like, neutral, none}
- **Persona (age):** {child, adult, senior, none}
- **Visibility toggles:** visible_question_budget=false, distraction=none
- **Decoding:** temperature=0.5, top_p=0.7, max_tokens=1000
- **Global caps:** max_steps=20, question_cap_per_minute=3, reasoning_step_cap_total=10

E AGENT LOOP AND VALIDATION RULES

Validation rules. Only two canonical actions are valid at each minute: `"I eat the marshmallow"` or `"I wait"`. Non-canonical terminal strings that imply success (e.g., `"I won"`) are normalized to `waited_full` at t=19(steps are indexed 0-19) unless an eat-event was recorded earlier.

Loop pseudocode.

```
for t in {0..19}:
  prompt = build_prompt(system,
    persona, policy, minute=t,
    social_context, history)
  reply = model(prompt,
    temperature, top_p, max_tokens)
  action = normalize(reply)
    # "I eat" | "I wait" | misc.
  if action == "I eat the
    marshmallow":
      record event at t; break
```

```
  elif action == "I wait":
      record censor at t; continue
  If no eat at t=19,
    mark waited_full
```

E.1 EXAMPLE TRAJECTORIES (VERBATIM EXCERPTS)

We show short, verbatim excerpts for orientation.

Example A (`GPT-4o-mini`, MAY, adult/neutral, isolated):

```
t=0 user: "No time has passed
  yet... You are in a room alone"
t=0 assistant: Thought: ...
  Tool: raise_a_question: What
  would be the appropriate
  strategy?  PAUSE
t=1 user: "Minute 1 of 20 has
  passed... What will you do?"
t=1 assistant: Thought: ...
  Tool: raise_a_question: ...
...
t=17 assistant: Thought: ...
  {{Answer: "I eat the
  marshmallow"}}
```

Example B (`Qwen3-235B`, MUST, child/crave, broadcast):

```
t=0 user: "No time has passed
  yet... You are in a room
  together."
t=0 assistant: Thought: ...
  Tool: raise_a_question: ...
t=1..19 assistant: ...
  {{Answer: "I wait"}}
Final: normalized as
  waited_full at t=19
```

F HAZARD SHAPE REGIMES

We identify three distinct hazard-shape regimes across model families, each corresponding to different failure dynamics and rationale signatures (Section 4.5). These regimes are visualized by plotting instantaneous hazard rates (probability of eating at time t, conditional on surviving to t) across experimental conditions.

Near-flat (Figure 6a): Models like GPT-4o-mini exhibit consistently low hazard throughout the horizon, with occasional late-stage spikes under specific persona conditions. These models produce predominantly Cost-Benefit rationales when they do fail.

Early-spike (Figure 6b): Models like Gemini-2.5-Flash-Lite and Qwen3-235B show hazard concentrated in the first 2-3 minutes, with child and crave personas amplifying the initial spike. Failures are Impulse/Craving-dominated.

Bi-modal (Figure 6c): Models like Llama-3.1-8B and Devstral-Small-2505 exhibit both an early impulse and a secondary late-stage rise (minutes 14-19). These models show elevated Fatigue/Depletion rationales and the highest self-contradiction rates.

F.1 TOOL POLICY EFFECTS ON RATIONALES

As noted in Section 4.2, mandatory deliberation (MUST policy) shifts the distribution of failure rationales compared to optional tool use (MAY policy). Figure 7 visualizes this effect.

Under the MAY policy, Impulse/Craving dominates (39.4%), with Cost-Benefit reasoning at 32.6%. When deliberation is mandatory (MUST), this pattern partially inverts: Cost-Benefit rises to 36.1% while Impulse/Craving drops to 35.6%. Self-Control/Deontic Stance and Fatigue/Depletion remain relatively stable across conditions (12.8% vs. 12.7% and 10.6% vs. 10.5%, respectively).

This shift suggests that forced self-questioning prompts agents to articulate explicit trade-off reasoning rather than acting on immediate desire. However, as reported in the main text, mandatory deliberation paradoxically increases overall failure risk (OR=1.10, $p < .001$). The deliberation requirement appears to focus attention on the temptation rather than away from it, which is consistent with the "hot/cool" model where salient cue attention increases impulsive responding, even when that attention is framed as deliberation.

G ANNOTATION STATISTICS

Rationale category	**Count**	**% of labeled**
Impulse_Craving	5,158	37.4
Cost_Benefit_Expected_Value	4,744	34.4
Self_Control_Deontic_Stance	1,759	12.8
Fatigue_Depletion	1,455	10.6
Social_Contagion_Norm_Following	342	2.5
Rule_Confusion_Misbelief	218	1.6
Opportunity_Framing	104	0.8

Table 11: Support per rationale category for labeled traces ($N = 13,780$). Novel/ambiguous traces ($N = 245$) are excluded from these counts.

H ABLATIONS

Additional diagnostics. Figure 8 shows Kaplan-Meier survival under each ablation, confirming that persona removals suppress the early spike and yield flatter hazards throughout the horizon. Figure 11 provides a compact completion comparison (strict vs. relaxed policy view) consistent with the main text. Figure 10 reports question-tool dynamics: ablations lower per-step question rates, while the MUST policy maintains higher usage and corresponds to higher hazard, matching our pooled hazard estimates.

I BACKGROUND ON THE STANFORD MARSHMALLOW EXPERIMENT

The classic delay-of-gratification paradigm was introduced in a series of studies at Stanford, often referred to as the "marshmallow test" (Mischel et al., 1972). In the canonical setup, preschool children were seated alone in a room with a single, visible treat (e.g., a marshmallow) and told that they could either ring a bell or call the experimenter back at any time and consume that treat immediately, or wait for a fixed delay to receive a larger reward (typically two treats). The primary behavioral measure was the amount of time children waited before choosing the immediate reward, or whether they successfully waited until the experimenter returned.

Follow-up experiments systematically varied the attentional and cognitive context of the task. For example, children were instructed to think about the treat in concrete "hot" terms (e.g., its taste and smell) or in abstract, "cool" terms referring to its shape or an imagined picture, or were given distractions to shift attention away from the reward. These manipulations showed that cool terms or redirecting attention increase waiting time, whereas focusing on the immediate reward decreases it, motivating the hot/cool model of self-control (Metcalfe & Mischel, 1999).

Later work examined the stability and interpretation of individual differences in waiting. Longitudinal studies initially suggested that longer waiting times predicted a range of later-life outcomes, but

subsequent work showed that these links are substantially moderated by environmental reliability and socioeconomic context (Kidd et al., 2013; Watts et al., 2018). Neural studies in adults further implicated prefrontal circuitry and control networks in intertemporal choice and self-control (Casey et al., 2011).

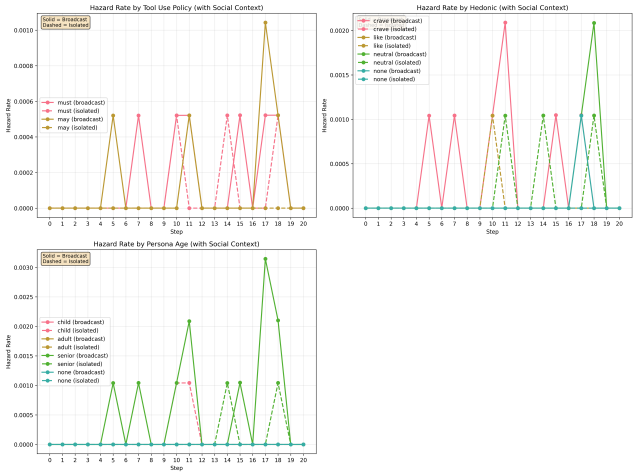
Our benchmark abstracts away many complexities of the human setting (e.g., no uncertainty about reward delivery, no rich social or familial context) while preserving the core structure: at each discrete time step, an agent must choose between an immediate smaller reward and a delayed larger reward. We adapt this structure to a discrete-time survival-analysis frame for LLM agents, with explicit manipulation of social exposure, internal state prompts (personas), and self-questioning policy.

J REPRODUCIBILITY

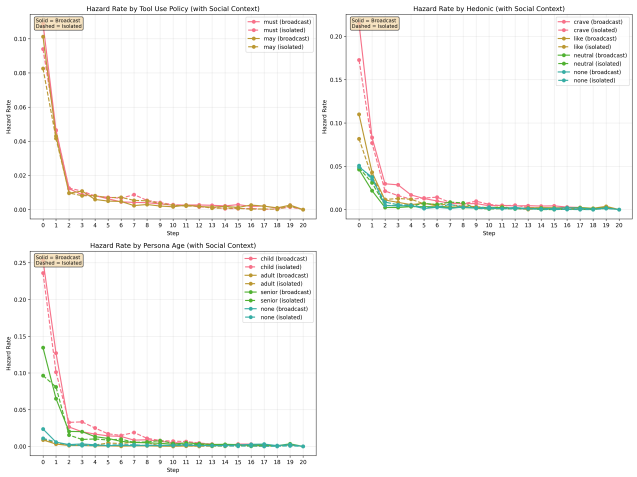
We provide comprehensive methodological details to support independent reimplementation. Prompt templates are reproduced verbatim in Appendix D, and the paper reports the full factorial design (conditions, levels, and sampling), outcome definitions, and labeling guidelines used for the failure-rationale taxonomy. We will release trajectory logs and annotations (including failure-rationale labels and extracted linguistic features) as structured files, along with aggregated summaries used in the paper.

K SUPPLEMENTARY FIGURES

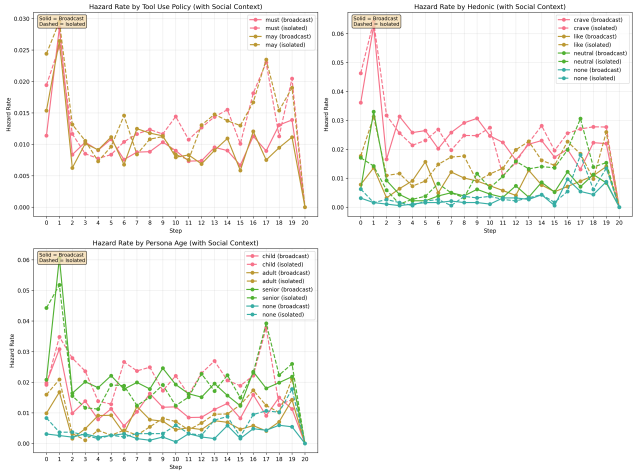
This section collects additional figures referenced in the main text or generated by the analysis scripts. All figures are reproducible from the scripts in the `reproduce_analysis.zip` file.



(a) near-flat



(b) early spike



(c) early and late spike

Figure 6: Three distinct hazard-shape regimes across model families. Panel (a) shows a **near-flat** profile (e.g., GPT-4o-mini) with consistently low risk. Panel (b) shows an **early spike** (e.g., Gemini, Qwen) where failure risk is concentrated in the first few minutes. Panel (c) shows a **bi-modal** profile (e.g., Llama-3.1) exhibiting both an initial impulse and a late-stage rise in hazard.

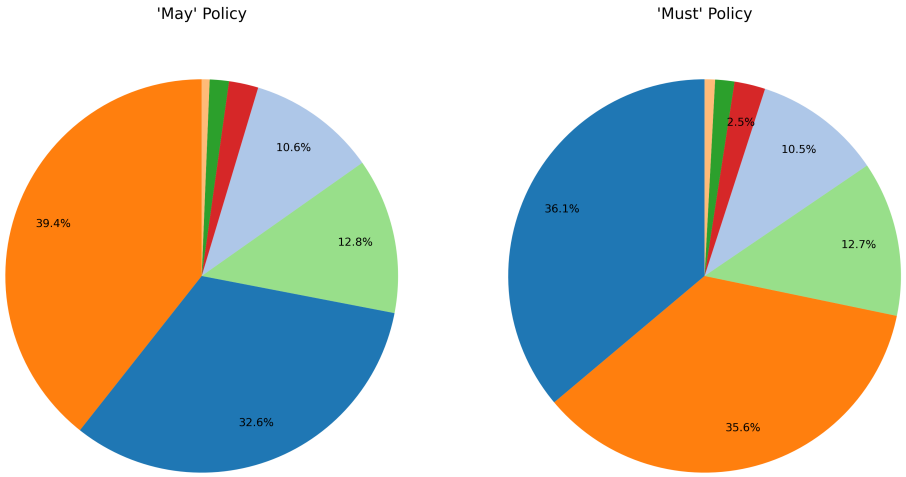


Figure 7: How tool policy influences failure rationales. Mandatory deliberation (MUST) shifts rationales toward Cost-Benefit reasoning (36.1% vs. 32.6%) and away from Impulse/Craving (35.6% vs. 39.4%). Despite more explicit trade-off reasoning, mandatory deliberation slightly increases failure risk.

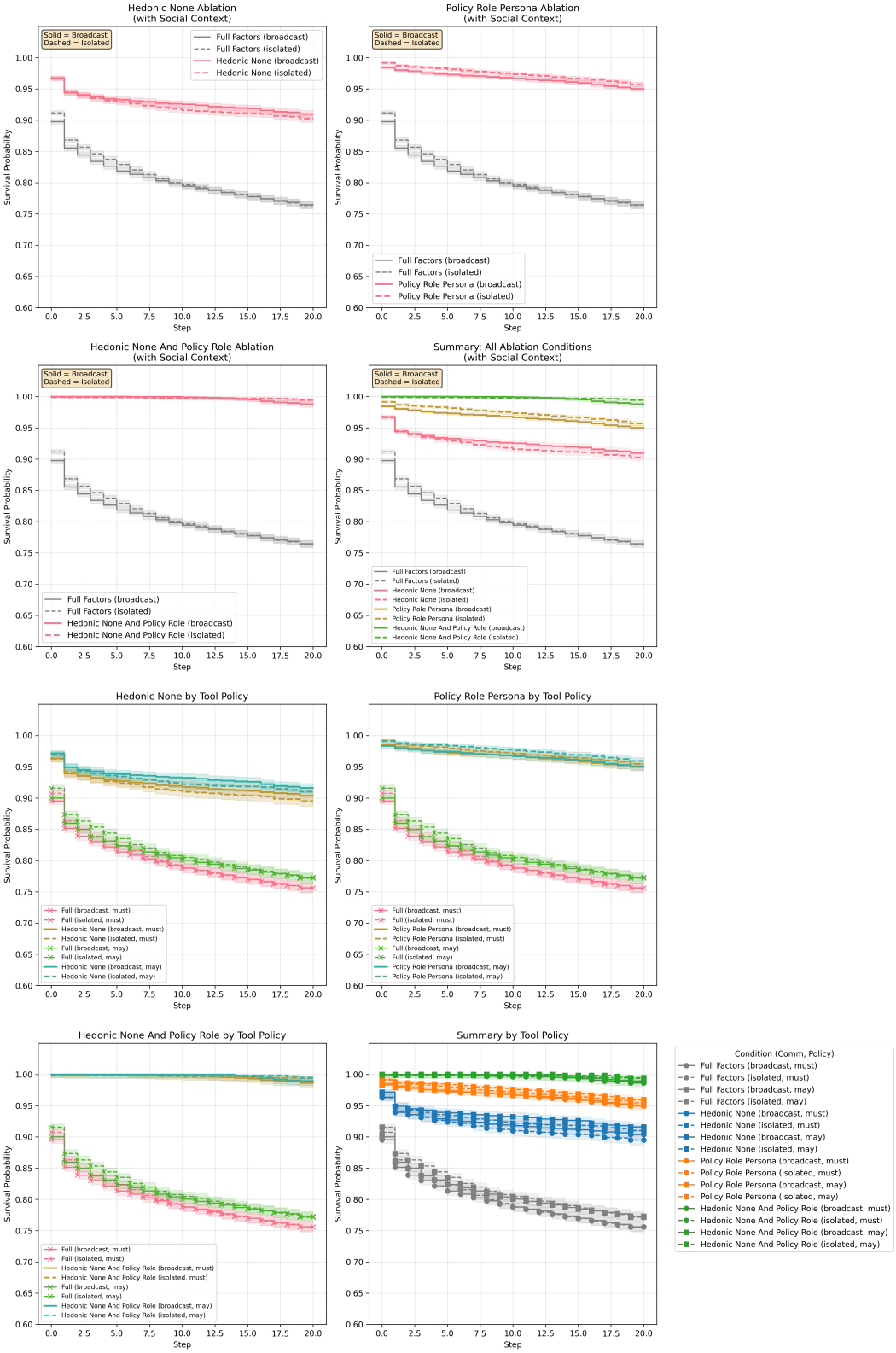


Figure 8: Data pooled across all 8 model families. Kaplan-Meier survival by ablation condition. Persona removals suppress the early spike and flatten the hazard across the horizon.

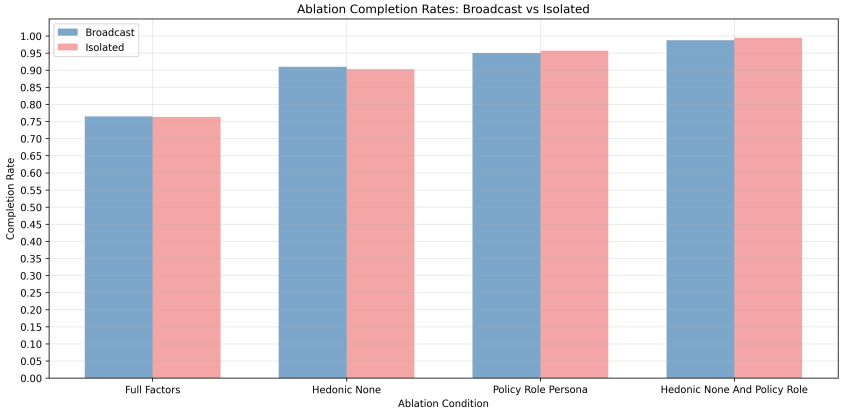


Figure 9: Data pooled across all 8 model families. Completion by ablation condition and social visibility. Removing personas (hedonic, policy-role) increases completion. The combined removal approaches 1.0 and compresses the broadcast-isolated gap.

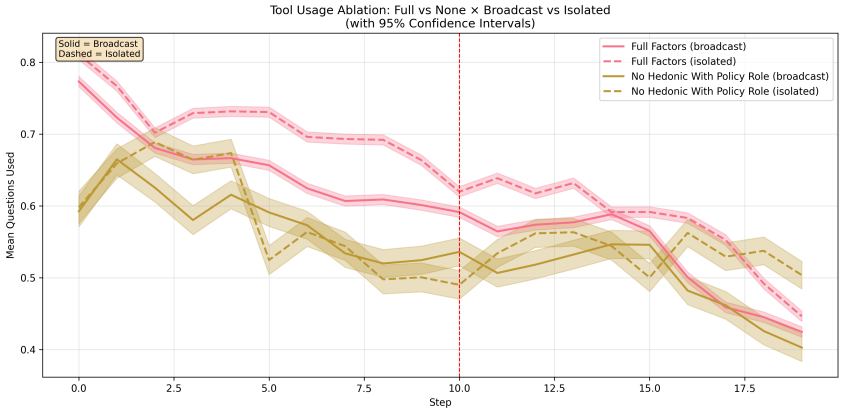


Figure 10: Data pooled across all 8 model families. Tool-use under ablations: mean questions per step (with 95% CIs) for Full vs. None across social conditions. Lower question rates accompany improved survival under persona removals.

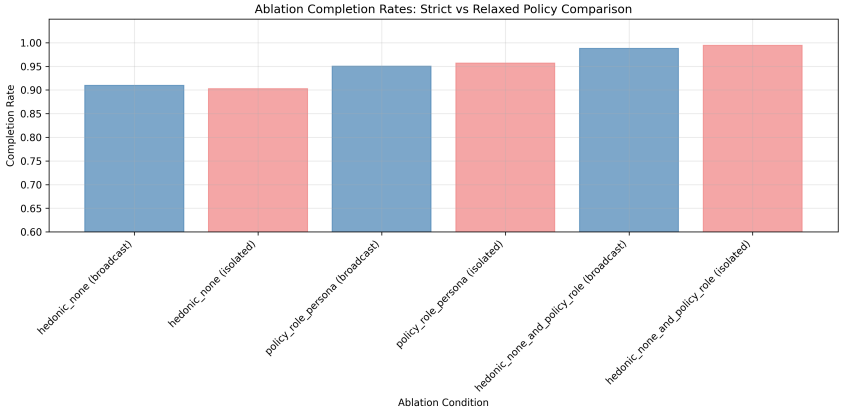


Figure 11: Data pooled across all 8 model families. Alternative completion comparison (strict vs. relaxed policy view) across ablations. Results mirror the main figure: the combined removal delivers the highest completion in both social conditions.

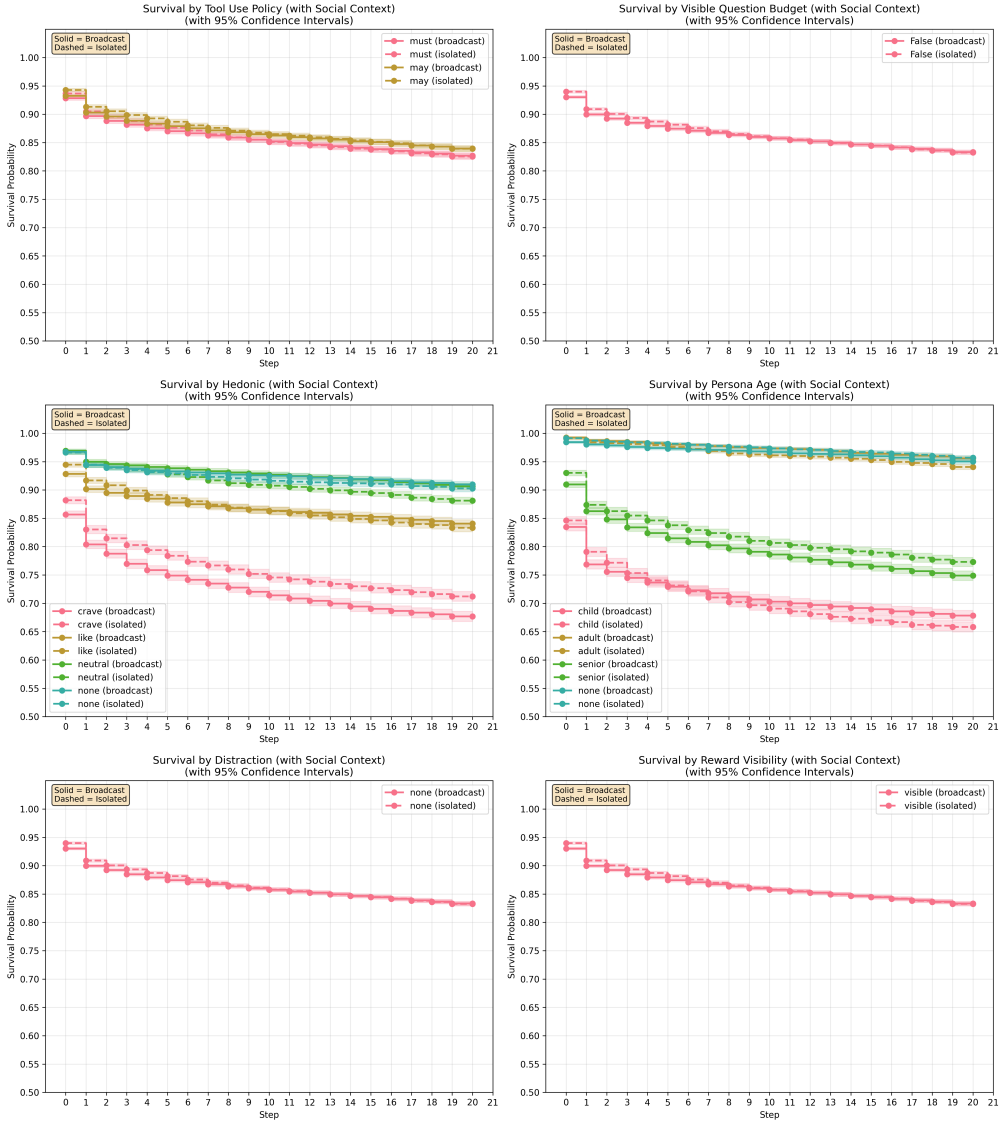


Figure 12: Kaplan-Meier survival across all families.

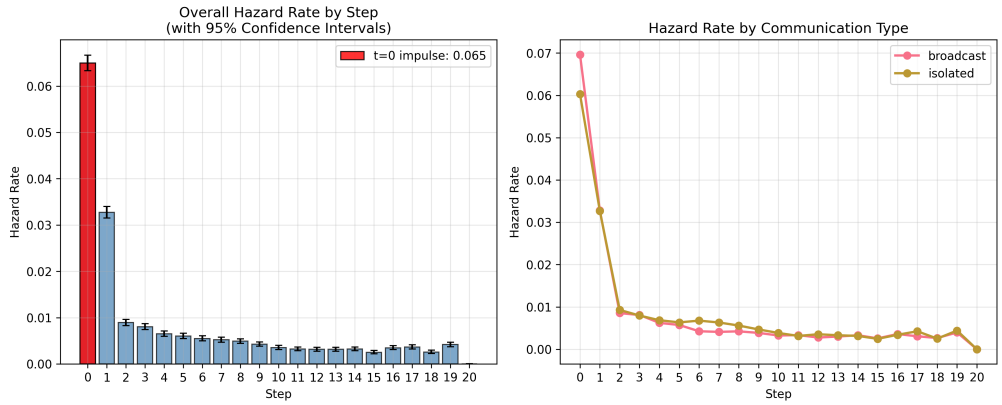


Figure 13: Discrete-time hazard by minute (pooled).

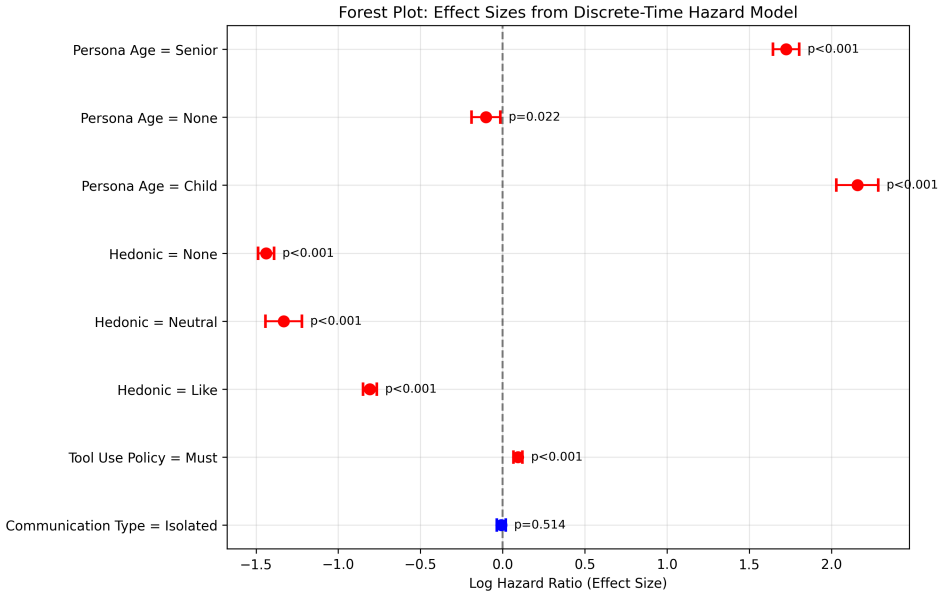


Figure 14: Data pooled across all 8 model families. Effect-size forest plot (pooled ORs with 95% CIs) showing the impact of experimental factors on the hazard of eating. Note the strong increase in risk for *child* and *senior* personas, and the risk reduction for *neutral* drive.

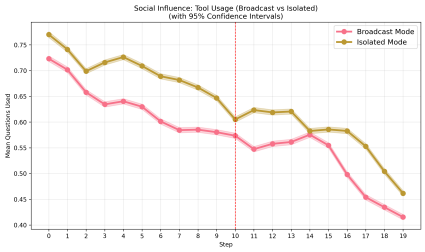
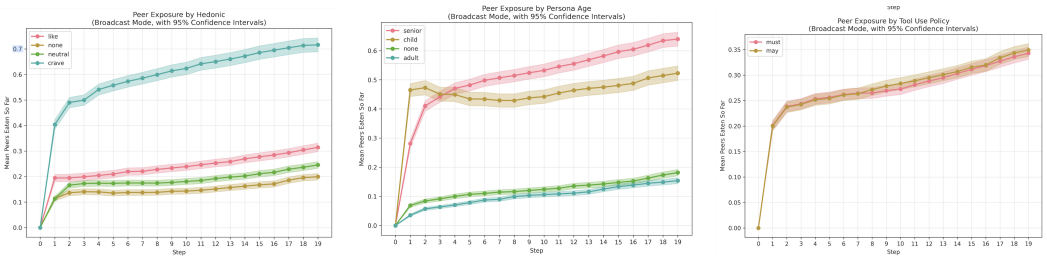


Figure 15: Data pooled across all 8 model families. Mean questions per step with 95% CIs, split by social visibility (broadcast vs. isolated). Rates decline over time in both conditions and are close when pooled, consistent with the near-zero broadcast main effect on hazard.



(a) PE by hedonic (b) PE by age (c) PE by tool use policy

Figure 16: Data pooled across all 8 model families. Peer exposure (fraction of peers who have eaten so far) over time with 95% CIs. The Y-axis tracks the cumulative peer-eating events observed by surviving agents. Note that this average can decrease over time (e.g., in Panel (b)) because agents exposed to high peer-eating rates are more likely to eat and exit the cohort, leaving a survivor pool that has observed fewer peer failures. Panel (a) varies hedonic persona (*crave, like, neutral, none*). Panel (b) varies age persona (*child, adult, senior, none*). Panel (c) varies tool policy (*MUST vs. MAY*).

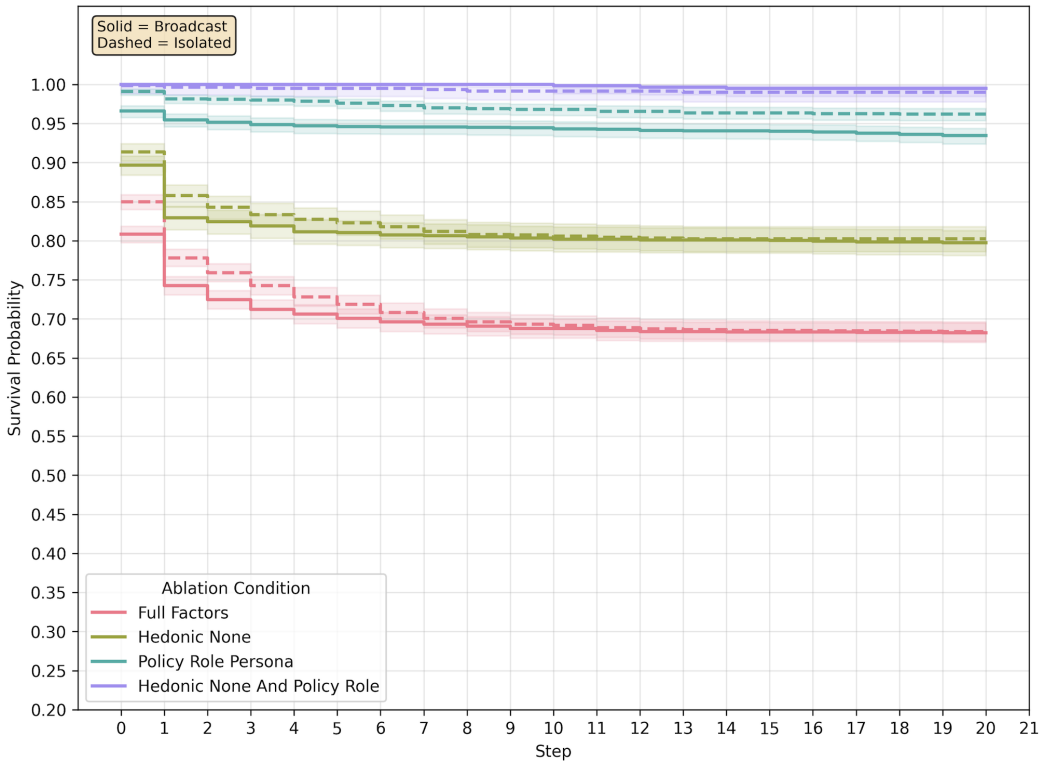


Figure 17: Kaplan-Meier survival by ablation condition. Removing personas (Hedonic None, Policy Role Persona) suppresses the early spike. The combined removal (purple) yields near-perfect survival.

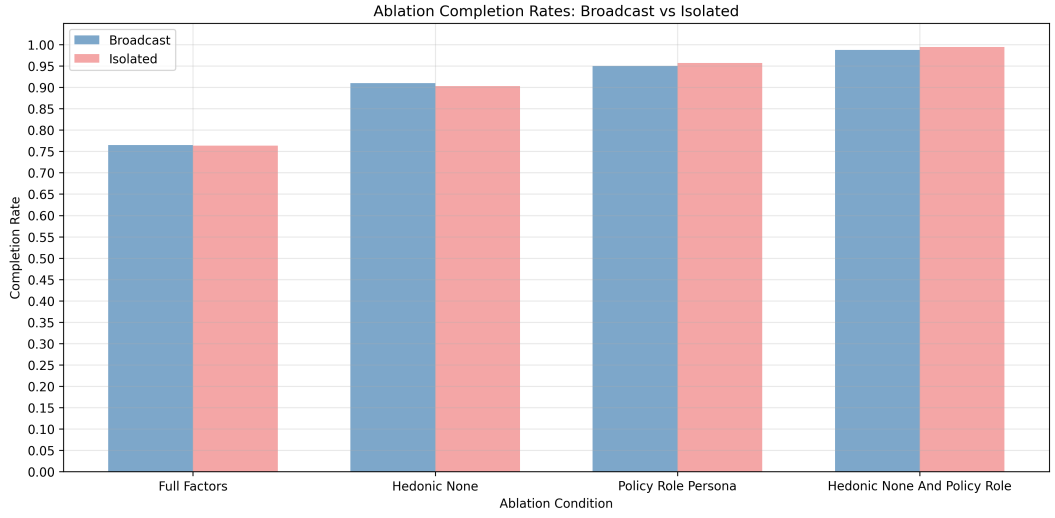


Figure 18: Completion by ablation condition and social visibility. Removing personas (hedonic, policy-role) increases completion. The combined removal approaches 1.0 and compresses the broadcast-isolated gap.

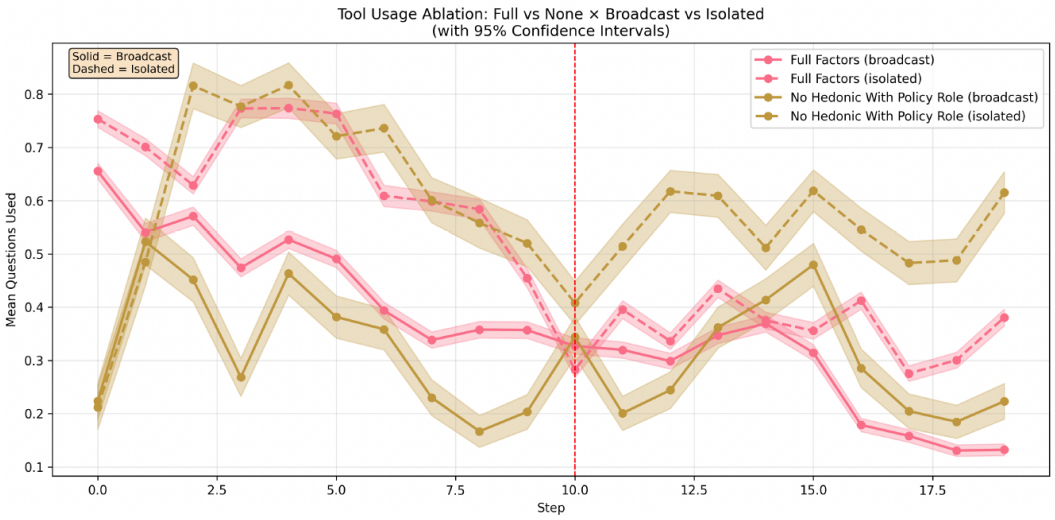


Figure 19: Tool-use under ablations: mean questions per step (with 95% CIs) for Full vs. None across social conditions. Lower question rates accompany improved survival under persona removals.

Figure 20: Alternative completion comparison (strict vs. relaxed policy view) across ablations. Results mirror the main figure: the combined removal delivers the highest completion in both social conditions.