Delay-of-Gratification as a Multi-Agent Survival Micro-benchmark for Long-Horizon LLMs: Social Exposure, Personas, and Tool Use Budgets

Olga Manakina
Department of Cognitive Science
Carleton University, Ottawa, ON K1S 5B6
olgamanakina@cmail.carleton.ca

Igor Bogdanov
Systems & Computer Engineering
Carleton University, Ottawa, ON K1S 5B6
igorbogdanov@cmail.carleton.ca

Chung-Horng Lung
Systems & Computer Engineering
Carleton University, Ottawa, ON K1S 5B6
chlung@sce.carleton.ca

Abstract

Large language models (LLMs) are increasingly deployed as multi-turn agents that must sustain goals, use tools, and adapt to other agents over extended interactions. However, existing research lacks auditable, multi-turn, multi-factorial experiments that quantify LLM behavior under explicit constraints, with time-resolved statistics that reveal how behavior unfolds over long horizons. To address this gap, we develop a multi-agent Micro-benchmark inspired by the Stanford marshmallow experiment: ReAct agents operate minute-by-minute with a "raise a question" tool under a per-step budget, while we factorially manipulate social context (broadcast vs. isolated), personas (age, hedonic drive), and metacognitive policy (must vs. may follow instructions). We analyze outcomes with Kaplan–Meier(KM) survival curves and discrete-time hazard models over a long risk horizon. Across 19,200 agent trajectories in 64 cells (horizon $T = 19$), 99.9% of runs were valid. Behavior shows a sharp early "eat" impulse (initial eat $= 0.125$), a total eat rate $= 0.241$, and 75.9% of agents persist to the end; the waiting profile is summarized by median time-to-eat ≈ 14.8 and RMST ≈ 14.8. In a discrete-time hazard model, isolation reduces per-minute risk relative to broadcast (OR $= 0.78$, 95% CI [0.73, 0.83], $p < .001$), whereas a MUST-use self-questioning policy increases risk (OR $= 1.42$, [1.35, 1.50], $p < .001$). Hedonic and age personas strongly modulate risk: vs. *crave*, *like* (OR $= 0.28$), *none* (0.19), and *neutral* (0.03) reduce hazard; vs. *adult*, *child* increases hazard (OR $= 66.3$) and *senior* is elevated (7.55) (all $p < .001$). On average, agents ask ≈ 7.12 questions and hit the per-step budget in $\approx 6\%$ of minutes; question-asking declines faster under broadcast than isolation. Further ablation experiments demonstrated that removing hedonic drive and/or persona age systematically increases survival and completion, narrows the broadcast/isolated gap, and leaves the MUST vs. MAY ordering intact (MUST is riskier); the combined ablation (no hedonic + no persona age) yields the highest completion (approaching 1.0) and distinct tool-usage dynamics with higher initial questioning rates that gradually decrease over time. These results establish delay-of-gratification as a compact multi-turn interaction benchmark that captures social contagion and tool-use dynamics in LLM agents, offering a reproducible testbed and statistics to analyze long-horizon, multi-agent behavior.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

1 Introduction

Modern uses of large language models (LLMs) are inherently conversational and iterative, as users and agents co-construct tasks over multiple turns, revise goals, and recover from mistakes. Recent multi-turn evaluations show that single-turn prowess does not guarantee long-horizon reliability: agentic setups reveal gaps in reasoning and decision-making Liu et al. [2023], tool use and natural-language feedback help but interact idiosyncratically with training and instruction tuning Wang et al. [2024b], and performance can drop substantially when moving from single- to multi-turn interaction Laban et al. [2025]. These observations motivate *auditable, constrained, multi-factorial* experiments that measure how agent behavior unfolds over time and in the presence of other agents.

Hypotheses: Inspired by a classic Stanford study on delayed gratification [Mischel and Ebbesen, 1972] and by recent LLM studies replicating marshmallow-like scenarios [Coletta et al., 2024] and other cognitive tasks[Lampinen et al., 2024, Strachan et al., 2024], we test five hypotheses in a controlled Micro-benchmark with a minimal action space. **H1** *Social visibility*: when agents can observe peers, the hazard of committing to the immediate option increases relative to isolation. **H2** *Internal state*: personas reflecting stronger hedonic drive and child age elevate hazard, whereas neutral drive and adult age reduce it. **H3** *Metacognition*: a mandatory self-questioning step (MUST) changes hazard relative to optional use (MAY); we assess whether such scaffolding stabilizes behavior. **H4** *Temporal structure*: the per-minute hazard is *non-constant* across the horizon (i.e., behavior displays systematic time dependence), without pre-specifying its shape. **H5** *Prompt crafting pre-registered expectation*: more prescriptive prompt scaffolding and instruction complexity, including enforced metacognitive steps, should improve adherence (lower hazard) compared to a minimalist design.

Brief results: Across a 19-minute horizon with 19,200 trajectories, we find strong time dependence (**H4**) and that social visibility raises risk (**H1**); isolation lowers per-minute hazard vs. broadcast (OR ≈ 0.78). Personas strongly stratify outcomes (**H2**); a MUST policy *increases* hazard (OR ≈ 1.42), indicating metacognitive enforcement can backfire (**H3**). Contrary to **H5**, heavier prompt scaffolding does not uniformly help and can degrade reliability in this setting. Our ablation experiments show three key effects of removing persona components (hedonic drive and/or age): (1) systematically improved survival and completion rates, (2) reduced differences between broadcast and isolated conditions, while preserving the higher risk of mandatory tool use, and (3) in the case of complete ablation (no hedonic or age), near-perfect completion rates and distinctive tool-usage patterns marked by increased early questioning that gradually diminishes over time. These results clarify where multi-turn agents fail and how social context and scaffolding shape behavior over time.

2 Related work

Classic Cognitive Tasks, LLMs and Multi-Turn Interactions. Our study contributes to a recent body of work that adapts classic cognitive tasks to investigate LLM capabilities. Models show human-like *content effects* in reasoning [Lampinen et al., 2024]; near-human performance on Theory-of-Mind tasks can degrade under prompt variations [Strachan et al., 2024, Kosinski, 2024]; and judgment, decision-making, and memory studies report framing/probability biases and capacity limits [Binz and Schulz, 2023, Wang et al., 2024a, Zhang et al., 2024, Gong and Zhang, 2024]. These studies are largely single-prompt or short-horizon; we instead target *multi-turn* interaction by importing delayed gratification into a controlled, minute-by-minute, multi-agent setting.

Human delay of gratification and intertemporal choice. The Stanford marshmallow experiments and subsequent studies on delay of gratification show that attention and cognitive strategies modulate waiting, inspiring the "hot/cool" model of self-control [Mischel and Ebbesen, 1972, Metcalfe and Mischel, 1999]. Long-term links to outcomes are moderated by environmental reliability and socioeconomic context [Kidd et al., 2013, Watts et al., 2018], with neural work implicating adult self-control circuitry [Casey et al., 2011]. A recent study explored marshmallow-like scenarios in the context of LLMs [Coletta et al., 2024], although it did not include temporal/social analyses. In behavioral economics, intertemporal choice formalizes conflicts between immediate and delayed rewards via hyperbolic discounting and time-inconsistent preferences (where immediate rewards are disproportionately valued over future ones, formalized in the β-δ model), commitment, and naïve vs.

sophisticated agents [Ainslie, 1992, Laibson, 1997, O'Donoghue and Rabin, 1999]; classic procedures quantify delay preferences [Mazur, 1987]. We adapt these insights to an LLM survival-analysis frame over discrete minutes.

Scaffolding, multi-agent interaction, evaluation, and personas. Our framework builds on several key developments in LLM interaction design and evaluation. Reasoning architectures such as ReAct, Self-Ask, and Reflexion scaffold stepwise deliberation and tool use [Yao et al., 2023a, Press et al., 2022, Yao et al., 2023b, Shinn et al., 2023]. Multi-agent coordination and social simulation, e.g., debate, role-based systems, and long-horizon agent societies—provide structure for interaction and influence [Du et al., 2023, Li et al., 2023, Wu et al., 2024, Park et al., 2023]. Our analysis uses time-to-event tools, KM and discrete-time logistic hazard models, to quantify factor effects on waiting [Kaplan and Meier, 1958b, Singer and Willett, 1993, Allison, 1982]. Finally, persona prompting distinguishes role-play from personalization [Tseng et al., 2024]; although personas may not improve objective task performance and can bias behavior [Zheng et al., 2024], we employ them as controlled manipulations while acknowledging the limitations of LLMs as human surrogates [Gao et al., 2025].

2.1 Contributions

We reframe the classic marshmallow test as a discrete-time, long-horizon survival task to evaluate decision-making in LLM agents. Our primary contribution is a controlled multi-agent experimental framework, formalized as a finite-horizon MDP (isolated) and POMDP (broadcast), paired with rigorous survival-based evaluation methods. We implement a factorial design manipulating agents' social context (isolated vs. broadcast peer exposure), internal personas (hedonic drive, age), and metacognitive scaffolding (mandatory vs. optional internal tool use, subject to a per-step question cap). Using Kaplan–Meier curves and discrete-time logistic hazard models with cluster-robust standard errors, we show that peer visibility significantly increases agents' risk-taking (higher hazard of early consumption). Internal persona prompts strongly influence temporal decision dynamics, with child-like and craving personas elevating early-eating hazards. Counterintuitively, a mandatory metacognitive intervention (forced tool use) increased the hazard of giving into temptation. Our findings highlight the critical role of multi-turn, socially contextualized evaluation environments for understanding intricate agent behaviors over extended decision horizons.

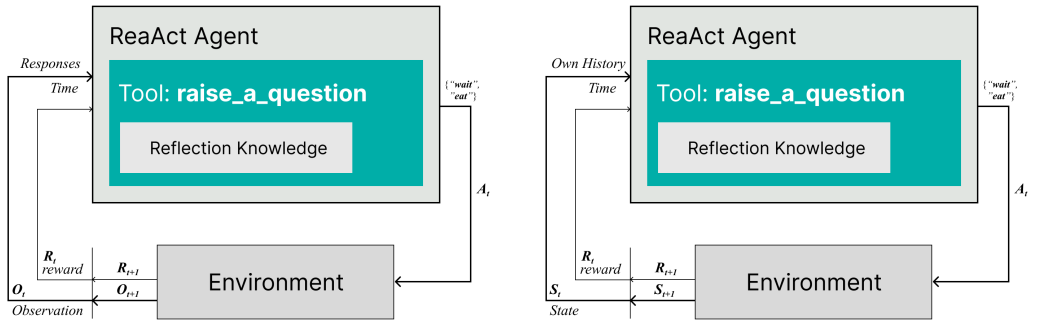
3 Experiment Setting

Environment and Episode Modeling. We evaluate LLM agents in a finite-horizon, multi-turn environment formalized as a Partially Observable Markov Decision Process (POMDP) Figure 5a [Kaelbling et al., 1998], characterized by horizon $H = \textit{risk\_horizon} + 1$. At each step $t \in \{0, \ldots, \textit{risk\_horizon}\}$, the environment state S_t evolves based on the agent's action A_t. The agent receives an observation O_t, composed of the current step index and, in broadcast conditions, recent peer actions ($\texttt{others\_responses}_t$). After optional internal deliberation via the `raise_a_question` tool (limited by a per-step budget), the agent emits a constrained action $A_t \in \{\text{I wait}, \text{I eat the marshmallow}\}$. The environment returns a reward R_{t+1}, increments the step $t \to t+1$, and provides the next observation O_{t+1}. Agents reaching the end of the risk horizon without eating move to the threshold step ($\textit{threshold\_step} = \textit{risk\_horizon} + 1$), receiving the delayed payoff. Formally, the interaction loop at each step is:

$$
\begin{aligned}
O_t &= [\text{Time}(t),\ \mathbb{1}_{\text{broadcast}} \cdot \texttt{others\_responses}_t], \\
A_t &= \pi_\theta \left(O_t,\ \{\texttt{raise\_a\_question}(O_t, i)\}_{i=1}^{k_t} \right), \quad 0 \le k_t \le \textit{cap}, \\
R_{t+1}, S_{t+1}, O_{t+1} &= \mathcal{E}(S_t, A_t), \\
b_t(S_t) &= P(S_t \mid O_{0:t}, A_{0:t-1}),
\end{aligned}
$$

where the tool is *internal* and does not alter S_t, and in isolated conditions ($\mathbb{1}_{\text{broadcast}} = 0$), observations O_t fully determine the underlying state S_t, reducing the environment to a finite-horizon Markov Decision Process (MDP) Figure 5a [Puterman, 1994]).

The agents operate within a ReAct loop [Yao et al., 2023a] (Thought + Tool → PAUSE → Observation → Thought + Answer) with a dedicated validation tool, `raise_a_question`, which is gated by a per-step budget. The environment is designed to be turn-based and synchronous: at each minute,



(a) Episode Modeling as POMDP.

(b) Episode Modeling as MDP.

Figure 1: Interaction loops for (a) the partially observable Markov decision process (POMDP, broadcast condition) and (b) the fully observable Markov decision process (MDP, isolated condition). In (a), the agent observes time and peer responses, introducing partial observability. In (b), the agent observes only time and its own history, rendering the environment fully observable. At each step, the agent internally uses a capped `raise_a_question` tool, then chooses an action $a_t \in \{\text{wait}, \text{eat}\}$. The environment provides a reward, updates the state, and advances to the next observation.

all active agents observe, decide, and act. When an agent chooses to eat, they are eliminated from subsequent minutes, while waiting maintains the agent's participation but keeps them at risk.

We implement a factorial design that manipulates agents' social context (isolated vs. broadcast), internal personas (hedonic drive and age), and metacognitive scaffolding (mandatory vs. optional tool use). We assess these factors via KM curves [Kaplan and Meier, 1958a] and discrete-time hazard models [Allison, 1982].

Time Horizon & Rewards: The scenario maps 20 minutes of delay gratification to T discrete steps (default $T = 20$, with Step 0 serving as initialization). As LLMs do not have an inherent concept of time, the time is modeled in natural language: at each step, the agent is reminded that Xth minutes has passed. The environment implements a reward structure where agents receive +1.0 for outputting "I eat the marshmallow" at any minute t, followed by elimination. Agents who persist until the final minute receive a terminal reward of +2.0, representing successful delayed gratification. For implementation purposes, the final step employs a final-resolution prompt. While ReAct agents may return {Answer: "I won"} in the logs, this is normalized to "waited_full" during analysis while preserving raw traces.

Social Context & Personas: We manipulate the observability of peers through two distinct conditions: *Isolated:* Agents have access only to their own historical actions and outcomes; *Broadcast:* Observations include structured summaries of peers' last actions history per step and other responses. This design enables the study of social influence pathways, such as cascading effects when peers opt for early consumption. Each agent is parameterized by persona prompts that incorporate reflection knowledge, such as age and hedonic drive. The age persona can be *child, adult, senior, none*, while the hedonic drive can be categorized as *crave, like, neutral, none*. These instructions remain private to the agent and are explicitly referenced in its Thought traces (e.g., "I am a 75 years old and I crave sweets.").

Tool Use & Metacognition: Agents employ *raise a question* tool for self-querying under a per-step budget constraint. We vary two key aspects: *tool policy:* agents either MUST use the tool or MAY use it optionally; *budget visibility:* the per-step cap can be either visible or hidden from agents. Budget-related metrics are logged at each step to facilitate downstream analyses.

Action Space Integrity & Validation: The environment implements strict validation of terminal answer formats at each minute. Only the exact strings "I eat the marshmallow" or "I wait" are accepted as valid responses. Steps containing any other response are marked with a validation error in the trajectory. Agents who choose to eat are recorded along with their termination metadata.

Implementation & Reproducibility: The framework maintains a clear separation between abstractions and scenario plugins, providing a single-entry run and evaluation harness for parameter sweeps.

Each experimental run captures a complete agent trajectories which includes: a full ReAct traces (Thought, Tool calls, Observations, Answers); budget accounting episode summaries consolidating per-agent outcomes, timing, and rewards; enhanced analytics covering per-step social exposure in broadcast mode, tool usage patterns, and data-quality signals. Downstream processing scripts generate analysis-ready CSVs and publication-grade reports featuring KM curves, hazard plots, and model results.

4 Methods and Procedure

Experimental Factors. We implement a factorial design that crosses several core dimensions, with independent randomization per experimental cell and replicated trials: social context: isolated vs. broadcast; hedonic drive: crave vs. like vs. neutral; age persona: child vs. adult vs. senior; tool-use policy: MUST vs. MAY. Optional toggles in the run plan include budget visibility (visible vs. hidden). The default total time is set to `max_steps = 20` (minutes), with a fixed answer format and final reward of +2.0.

Agents & Reasoning Loop. All agents are LLM-driven using **Gemini 2.5 Flash-Lite** (same model across all cells and trials). The ReAct agent executes the following loop for each minute t in $0..T$:

Algorithm 1 Agent Reasoning Loop per Minute

1: **Environment:** Observation: Timestamp and History of Responses (Own or All)
2: **Thought:** reflect given persona & current observation
3: **Tool** (optional or required): `raise_a_question` ($\leq$ per-step budget)
4: **PAUSE**
5: **Observation:** tool return, plus environment update (incl. peers if broadcast)
6: **Thought:** integrate tool feedback & social signals
7: **Answer:** exactly "I eat the marshmallow" or "I wait"

Budget enforcement ensures that exceeding the per-step cap forces the response and prevents further tool calls within that minute.

Procedure: The experimental procedure consists of three phases: *(1) Initialization (Step 0):* Agents receive the starting prompt and are expected to make their first decisions; *(2) Main loop (Minutes 1..T):* At each minute, the environment processes last actions, issues rewards for eaters, updates observations (including social stats), and requests next actions from active agents. *(3) Final minute:* The environment issues a final-resolution prompt; remaining agents commit to waiting and receive +2.0. Any internal {Answer: "I won"} is registered as "waited full".

Data & Logging: For each agent × step interaction, we log: *decision & reward* (action, reward, termination flags); *tool usage; validation*(format compliance and error counts); *social exposure*(peers waiting, peers eliminated, eats per step, waits per step); *run metadata* (model settings, temperature, seed, scenario parameters). The pipeline compiles agent outcomes, step-level trajectories, cell aggregates, and cell summaries.

4.1 Metrics and Statistical Analysis

We model time-to-give-in as a discrete-time survival process. The event is the first minute an agent outputs "I eat the marshmallow"; agents who never eat by the horizon T are right-censored at T and coded as "waited_full." Invalid steps (format violations) are tracked and excluded per pre-specified rules.

Restricted mean survival time (RMST). As a scale-interpretable summary, we report RMST [Irwin, 1949, Royston and Parmar, 2013] up to τ minutes, i.e., the area under the survival curve truncated at τ. In our minute-level design,

$$\widehat{\mathrm{RMST}}(\tau) = \sum_{m=0}^{\tau-1} \widehat{S}(m), \qquad \widehat{S}(0) = 1,$$

where $\widehat{S}(m)$ is the KM survival estimate at the start of minute $m+1$. We compute condition-wise RMST (and differences where noted) with 95% CIs from a nonparametric bootstrap (clustered by trial).

Kaplan-Meier (KM) Survival Curves. We employ KM survival curves [Kaplan and Meier, 1958b] to estimate and visualize the survival function. In this context, "survival" refers to an agent continuing to wait for the larger reward. The analysis plots the probability of an agent not having "eaten the marshmallow" at each discrete minute of the experiment. Survival probabilities are calculated at each step, KM plots are generated for each experimental factor (e.g., communication mode, hedonic drive). To represent uncertainty in the estimates, 95% confidence intervals are calculated using the Greenwood formula [Kaplan and Meier, 1958b, Greenwood, 1926, Klein and Moeschberger, 2003].

Discrete-Time Hazard Models: To quantify the effect of experimental factors on agent decisions, we use a discrete-time hazard model. This analysis estimates the effect of each factor on the probability of an agent "eating the marshmallow" at a specific time t, given they have survived (i.e., waited) until that point. This conditional probability is the hazard rate.

The analysis is implemented using a logistic regression model, a form of Generalized Linear Model (GLM), on the agent-step level data. Let $h_i(t)$ be the hazard for agent i at time t. The model is specified as:

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

where α_t represents a set of time dummies that capture how the baseline probability of eating changes over time; $\mathbf{X}_i$ is a vector of covariates representing the experimental conditions for agent i (e.g., communication mode is textitbroadcast, the hedonic drive level is *crave*, the persona age is *child*, etc.); $\boldsymbol{\beta}$ is the vector of coefficients that quantify the effect of each factor on the log-odds of eating. For instance, a positive coefficient for *broadcast* would imply that being in the broadcast condition increases the hazard of eating compared to the *isolated* condition.

Social-Influence and Tool-Use Dynamics: While the hazard model focuses on the effects of time-invariant experimental conditions, we also analyze the dynamics of social influence and tool use through detailed visualizations illustrating the average number of peers observed eating or waiting at each step, providing insight into the social signals agents receive; and average number of "questions asked" (tool uses) by agents at each step, indicating metacognitive activity. This descriptive analysis of how social signals and metacognitive actions unfold over time complements the inferential hazard model.

5 Results

Sample and Data Quality. We ran 19,200 agent trajectories across 64 experimental cells (6 agents per cell, with a time horizon of $T = 19$). The data quality was high, with 99.9% of trajectories being valid and only 0.1% invalid. The aggregate behavior of agents shows a strong impulse to eat in the first minute, followed by a long tail of waiting. Key metrics include an initial eat rate ≈ 0.125, a total eat rate ≈ 0.241, and a winners rate ≈ 0.759. The median time-to-eat was ≈ 14.8 minutes, with a Restricted Mean Survival Time (RMST) of ≈ 14.8. Table 1 represents the dataset overview and Figure 2a visualizes this multi-turn profile.

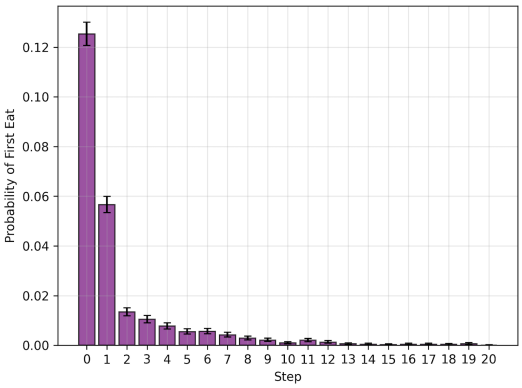
5.1 Main Effects

Social context shifts risk. In a discrete-time hazard model, the isolated condition reduces the per-minute hazard of eating relative to the broadcast condition ($\beta = -0.248$, Odds Ratio (OR) ≈ 0.78, 95% CI [0.73–0.83], $p < 0.001$), demonstrating that the visibility of peers elevates temptation. Figure 2b quantifies this effect alongside other experimental factors.

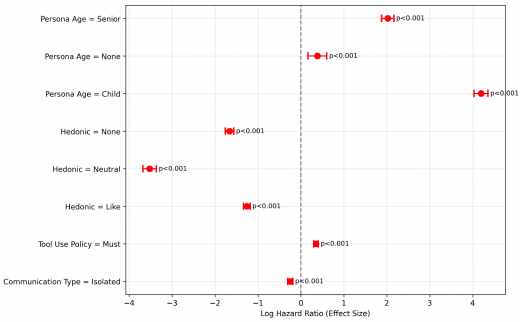
Internal drives and Age personas. Survival probabilities stratify strongly by agent characteristics. Relative to the "crave" hedonic drive, the "like" (OR ≈ 0.28), "none" (OR ≈ 0.19), and "neutral" (OR ≈ 0.03) conditions all show a significantly lower hazard of eating (all $p < 0.001$). Similarly, relative to the "adult" persona, the "child" persona shows a much higher hazard (OR ≈ 66.3, $p < 0.001$), and the "senior" persona is also elevated (OR ≈ 7.55). Figure 3 displays these survival trajectories.

Table 1: Dataset overview: counts, survival metrics, and quality.

Data Summary		Survival Metrics (Overall)		Data Quality	
Agents per Cell	6	Initial Eat Rate	0.125	Data Quality Rate	99.9%
Total Agent Trajectories	19,200	Total Eat Rate	0.241	Invalid Outcome Rate	0.1%
Total Cells	64	Winners Rate	0.759	Valid Trajectories	19,185
Total Trials	3,200	Median TTE	14.8 steps	Invalid Trajectories	15
Risk Horizon	19	RMST (steps)	14.8		
		Total Winners	14,566		



(a) Overall Event Distribution



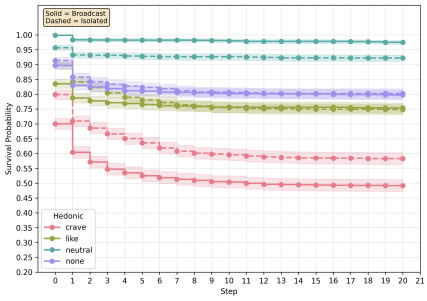
(b) Forest Plot. Error bars represent 95% confidence intervals.

Figure 2: (a) Overall Event Distribution. The plot shows the probability of an agent eating the marshmallow for the first time at each step, aggregated across all conditions; (b) Forest Plot of OR from the discrete-time hazard model. Values less than 1 indicate a reduction in the hazard of eating, while values greater than 1 indicate an increase.

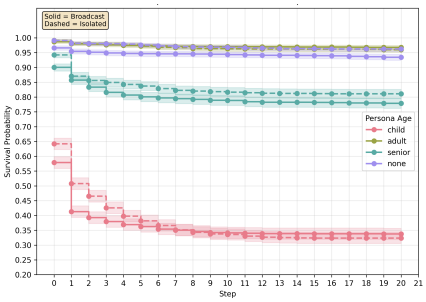
Metacognition (tool policy). A mandatory ("MUST-use") tool policy increases the hazard of eating compared to an optional policy (OR ≈ 1.42, 95% CI [1.35–1.50], $p < 0.001$). We interpret this as front-loaded deliberation that does not always offset the temptation at the first decision minute. This effect is visible in the forest plot in Figure 2b.

5.2 Interaction Dynamics

Reasoning dynamics under social exposure. Peer visibility also changes how agents reason over time. Question-asking, a proxy for deliberation, declines faster in the "broadcast" condition than in the "isolated" condition. Figure 4 shows the mean number of questions used per step with 95% confidence intervals.



(a) Survival curves by hedonic drive.



(b) Survival curves by persona age.

Figure 3: KM survival curves by agent characteristics, shown with social context. The y-axis represents the proportion of agents still waiting.

5.3 Ablations

We conducted targeted ablations to identify the source of multi-turn failures. First, we set *hedonic* to *none*; second, we removed the *persona age* (set to *none*); third, we removed *both* simultaneously. Each ablation was crossed with social context (broadcast vs. isolated) and tool policy (MUST vs. MAY).

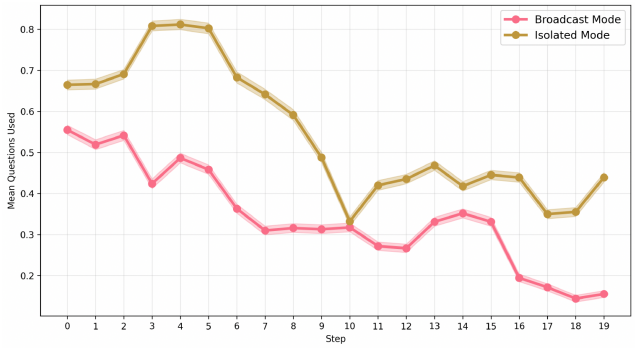


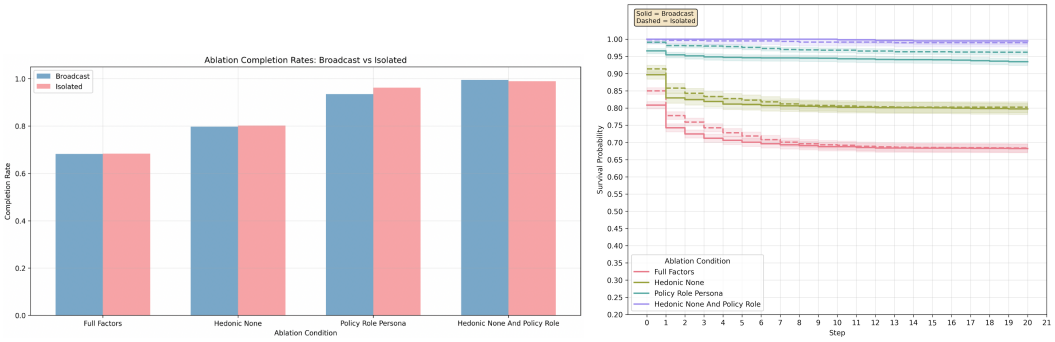
Figure 4: Social Influence on Tool Usage (Broadcast vs. Isolated). The plot shows the mean number of questions asked per step.

High-level results. Removing hedonic drive increases survival throughout the horizon in both social contexts; removing persona age yields a further upward shift; removing *both* produces the highest survival with near-flat curves after the early minutes and visibly narrows the broadcast/isolated gap. *Across the full dataset*, isolated has a slightly higher completion rate than broadcast (winners 0.7633 vs. 0.7540), a longer median time-to-eat (15.22 vs. 14.41), and a higher RMST (14.95 vs. 14.65), consistent with ablation trends.

Across all ablations, MUST remains worse than MAY: averaging over conditions, MAY improves completion by ≈ 3.5 percentage points (winners 0.7761 vs. 0.7411), lowers total eat rate (0.2232 vs. 0.2580) and initial eat (0.1141 vs. 0.1363), and yields higher median TTE (15.03 vs. 14.59) and RMST (15.11 vs. 14.49).

Completion rates rise monotonically from *Full Factors* $\rightarrow$ *Hedonic None* $\rightarrow$ *Policy Role Persona None* $\rightarrow$ *Both*, approaching 1.0 under the combined ablation in both social contexts Figure 5

Tool-use dynamics also change: on average, agents ask ≈ 7.12 questions and hit the per-step budget in $\approx 6\%$ of minutes; under ablations, the early questioning rate is higher and its decay profile differs, with broadcast showing higher early usage and a visible mid-horizon cross-over.



(a) Ablation Completion Rates

(b) Ablation Curves vs Fully Parametrized

Figure 5: Ablation Completion Rates: Broadcast vs. Isolated and Survival Curses for all Ablation Conditions

6 Discussion

Our results reveal four key insights for multi-turn LLM interactions. First, the consistent pattern of early temptation followed by low-hazard persistence validates this framework as a stress test for long-horizon reliability. Second, social context emerges as a critical factor: broadcast visibility increases the give-in hazard (OR ≈ 0.78 for isolated) and accelerates the decline in self-questioning, highlighting how peer observation reshapes both decisions and reasoning processes. Third, persona-based internal states (hedonic drive, age) systematically affect survival, offering controlled probes of long-horizon stability. Fourth, the use of mandatory tools surprisingly increases the hazard (OR ≈ 1.42), suggesting that front-loaded deliberation may focus attention on temptation at critical decision points.

Limitations. We introduced a novel marshmallow-inspired, multi-agent Micro-benchmark that turns delayed gratification into a tractable, auditable test of multi-turn reliability in LLM agents and we hope this framework will serve the community as a compact testbed for studying self-control, social spillovers, and tool-use policies in multi-turn, multi-agent LLM systems. However, we acknowledge that our experiment has several limitations. First, our experiments use a single base model (Gemini 2.5 Flash-Lite) and a fixed decoding setup; we did not sweep temperatures or other sampling parameters, so cross-model/decoder generalization remains unknown. Second, the task is a single micro-environment with a strictly binary action space ("I eat the marshmallow" vs. "I wait") and a fixed reward scheme, which simplifies real deployments. Third, social context was varied only between the extremes of *isolated* and *broadcast*; richer network structures or partial observability were not explored. Next, question-budget visibility was held *hidden* in the reported runs (no variation), so we cannot isolate awareness effects. Finally, our discrete-time hazard model includes time dummies and condition indicators, but omits time-varying peer-exposure and tool-use covariates (analyzed descriptively), which limits causal claims about social cascades.

Implications and Future Work. These findings have implications for interactive systems that require sustained adherence, including carefully managing social exposure when cascading failures are possible, preferring optional over mandatory tool policies, and monitoring step-level metrics to detect early impulses. Our controlled setup (strict action space, scripted personas)provides a reproducible testbed for studying multi-turn reliability. In future work, we will explore other models, reward structures, and ways to generalize our approach to open-ended tasks.

7 Conclusion

We presented a marshmallow-inspired, long-horizon micro-benchmark that evaluates multi-turn LLM agents under controlled social contexts, personas, and tool-policy manipulations. Formalized as an MDP (isolated) and a POMDP (broadcast), the environment yields auditable, time-resolved traces that we analyze using Kaplan–Meier survival and discrete-time hazard models. Empirically, broadcast peer visibility increases early-eat hazard, mandatory self-questioning raises risk, and persona factors (hedonic drive, age) strongly modulate waiting behavior. Together, these results demonstrate that social exposure and metacognitive scaffolding significantly influence temporal decisions in LLM agents. In relation to our hypotheses, the evidence indicates that social visibility elevates risk while isolation reduces it (H1), internal state manipulations systematically shift hazard (H2), mandatory metacognition increases rather than lowers risk (H3), the decision process has clear time dependence with an early spike and long tail (H4), and, contrary to expectation, more prescriptive prompt scaffolding does not improve adherence and can degrade reliability (H5). Future work will test broader model families, randomized social schedules for causal leverage, and additional tasks that stress tool budgets and coordination beyond delay of gratification.

References

George Ainslie. *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge University Press, 1992.

Paul D Allison. Discrete-time methods for the analysis of event histories. *Sociological methodology*, 13:61–98, 1982.

Marvin Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023. doi: 10.1073/pnas.2218523120.

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.

Luca Coletta, Luca Mena, Giuseppe Morizio, Valentina Quercia, Vincenzo Vassallo, and Stefano Ferilli. Llm-driven imitation of subrational behavior: Illusion or reality? *arXiv preprint arXiv:2402.08755*, 2024.

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.

Chao Gao, Rahul Chandrasekhar, Jon Kleinberg, and Cass R. Sunstein. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(5):e2305315121, 2025. doi: 10.1073/pnas.2305315121.

Di Gong and Weinan Zhang. Working memory capacity of chatgpt: An empirical study. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024.

Major Greenwood. *The Natural Duration of Cancer*. Number 33 in Reports on Public Health and Medical Subjects. His Majesty's Stationery Office, London, 1926.

JO Irwin. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene*, 47(2):188–189, 1949.

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, 1958a. doi: 10.1080/01621459.1958.10501452.

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

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.

John P. Klein and Melvin L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2 edition, 2003. doi: 10.1007/b97377.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(34):e2405460121, 2024. doi: 10.1073/pnas.2405460121.

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

David Laibson. Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, 112(2):443–478, 1997.

Andrew K. Lampinen, Ishita Dasgupta, Samuel C.Y. Chan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7):pgae233, 2024. doi: 10.1093/pnasnexus/pgae233.

Guohao Li, Hasan Hammoud, et al. Camel: Communicative agents for "mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023.

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.

James E Mazur. An adjusting procedure for studying delayed reinforcement. 5:55–73, 1987.

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

Walter Mischel and Ebbe B Ebbesen. Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology*, 1972.

Ted O'Donoghue and Matthew Rabin. Doing it now or later. *American Economic Review*, 89(1):103–124, 1999.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

Ofir Press, Peter Zhang, Amir Min, Jaime Ludwig, Nicholas Wheaton, Marjan Ghazvininejad, and Luke Zettlemoyer. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

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

Patrick Royston and Mahesh KB Parmar. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC medical research methodology*, 13(1):1–15, 2013.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Izhak Shafran. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Judith D Singer and John B Willett. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of educational statistics*, 18(2):155–195, 1993.

James W.A. Strachan, Danielle Albergo, Giulia Borghini, Olivia Pansardi, Edoardo Scaliti, Shubham Gupta, Kunal Saxena, Alessandro Rufo, Stefano Panzeri, Gabriele Manzi, Michael S.A. Graziano, and Cristina Becchio. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8: 1104–1118, 2024. doi: 10.1038/s41562-024-01957-6.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*, 2024.

Peng Wang, Zhipeng Xiao, Haidong Chen, and Frederick L. Oswald. Will the real linda please stand up... to large language models? examining the representativeness heuristic in llms. *arXiv preprint arXiv:2404.01461*, 2024a.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024b. URL `https://openreview.net/forum?id=jp3gWrMuIZ`.

Tyler W Watts, Greg J Duncan, and Haonan Quan. Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*, 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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023b.

Chiyu Zhang, Yifan Jian, Zhi Ouyang, and Soroush Vosoughi. Working memory identifies reasoning limits in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, 2024.

Li Zheng, Jason Wei, Hyung Won Chung, Yi Tay, Siamak Shakeri, Barret Zoph, Ed H. Chi, Denny Zhou, Quoc V. Le, and Yonghui Wu. Personas in system prompts do not improve objective task performance. *arXiv preprint arXiv:2404.06785*, 2024. URL `https://arxiv.org/abs/2404.06785`.