Autonomy-Supportive Quest Decomposition: Self-Determination Theory and Zone-of-Proximal-Development Matching in Gamified Civic Production

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Open-source software (OSS) repositories and civic commons frequently collapse into a "death spiral" of backlog accumulation, rising review latency, and contributor attrition. Sustaining voluntary participation typically fails along two extremes: controlling point-and-streak leaderboards, which crowd out intrinsic motivation via the overjustification effect and induce rubric gaming, or AI micro-approval loops, which relegate humans to passive rubber-stamping that stalls domain mastery. In this HCI and mechanism-design modeling paper, we formulate Autonomy-Supportive Zone-of-Proximal-Development Quest Decomposition (AS-ZPD-QD). AS-ZPD-QD partitions project graphs into cohesion-aware Directed Acyclic Graphs (DAGs) of bounded quests, presents contributors with a submodular menu of $K = 3$ choices calibrated to their Zone of Proximal Development ($\theta_i + \Delta_{\text{ZPD}}$), and regulates explanation verbosity via a strictly convex cognitive-load gate with a continuous interior optimum $\nu_t^* \in (0, 1)$. Across 300 18-node dependency graphs, AS-ZPD-QD cuts cross-quest interface weight to $25.46\% \pm 0.11\%$, achieves 97.74% [97.11, 98.24] flow-duration compliance, and lowers ZPD targeting error from 21.55 ± 1.06 to 2.64 ± 0.24. In a unified-response longitudinal simulation tracking 2,400 simulated contributors ($N = 600$ per arm across newcomer and expert personas) over 12 weekly sprints (28,800 simulated person-weeks), AS-ZPD-QD sustains simulated Week-12 retention at 88.67% [85.88, 90.96] and holdout-verified task quality at 78.50% [75.04, 81.60] while advancing skill mastery from 49.09 to 76.63 ± 0.83, outperforming Unstructured Backlogs (43.33% [39.42, 47.33] retention), Extrinsic Leaderboards (35.83% [32.10, 39.75] retention), and Agent Micro-Approval (35.00% [31.29, 38.90] retention).

1 Introduction

As artificial intelligence automates routine code generation and administrative workflows [13, 19], a central design challenge for human–AI collectives is organizing voluntary, skill-building human participation in open-source software (OSS) and civic infrastructure. When newcomers attempt to contribute to open-source or municipal commons, they encounter severe structural barriers [21]. In a large-scale empirical analysis of open-source maintenance trajectories, Kaushik and Chahal [12] characterize the repository "death spiral": monolithic, poorly scoped issue backlogs accumulate faster than maintainers can triage them, review latency escalates, and voluntary contributors disengage.

1.1 The twin traps of controlling gamification and passive micro-approval

To lower onboarding friction and sustain engagement, platforms have historically relied on either superficial "gamification" (points, badges, streak timers, and public leaderboards; 9) or top-down AI orchestration where an autonomous agent decomposes the task and dispatches micro-approvals to humans [20, 22]. Empirical evidence from psychology, ergonomics, and 2025–2026 human–AI interaction studies shows that both approaches systematically degrade long-horizon human performance:

1. **Intrinsic motivation crowding-out under controlling rewards.** In their foundational meta-analysis of 128 laboratory and field experiments on Self-Determination Theory (SDT), Deci et al. [6] proved that while informational competence feedback enhances intrinsic motivation, tangible, expected, and controlling extrinsic contingencies shift the perceived locus of causality from internal to external, triggering the overjustification effect and reducing free-choice persistence once rewards are normalized. Similarly, Hamari et al. [9] show that leaderboard-style gamification produces short-lived novelty spikes that decay rapidly when underlying tasks lack volitional choice.

2. **Automation ironies, deskilling, and cognitive divergence.** Conversely, when AI systems take over primary problem solving and relegate humans to supervising or rubber-stamping micro-steps, human domain mastery atrophies—a classical human-factors pathology identified in Bainbridge [2]'s *Ironies of Automation*. In randomized coding trials, Shen and Tamkin [20] demonstrate that unscaffolded AI assistance significantly impairs conceptual mastery and debugging transfer when assistance is removed. Concurrently, Eliav [8] document systematic behavioral divergence between human and LLM problem-solving representations, while Cheng et al. [3] and Conlon and Schwardmann [5] show that sycophantic agent explanations flatter user misconceptions rather than cultivating critical oversight.

3. **Effort–recovery imbalance and cognitive-load burnout.** From occupational workload theory [15, 16], sustained cognitive effort without task closure or autonomy depletes self-regulatory resources. Flooding human reviewers with verbose, uncalibrated AI traces or miscalibrated confidence scores [1, 7] accelerates mental fatigue without improving defect detection [22].

1.2 Autonomy-supportive DAG decomposition and ZPD matching

How can a civic or open-source production engine structure complex technical work so that it supports human autonomy and skill growth? In this HCI and mechanism-design position and modeling paper, we synthesize multi-armed bandit Zone-of-Proximal-Development progression (`ZPDES`, 4), job-rotation and cross-training ergonomics [10, 18], submodular menu optimization [17, 24], and graph connectivity modeling [11, 14] into Autonomy-Supportive Zone-of-Proximal-Development Quest Decomposition (AS-ZPD-QD).

In AS-ZPD-QD, monolithic project backlogs are first partitioned into a cohesion-aware Directed Acyclic Graph (DAG) of bounded, verifiable quests [11, 23]. Rather than assigning a mandatory ticket or ranking workers on a controlling scalar leaderboard, the engine presents each contributor i (with current latent skill $\theta_i(t)$) with an *autonomy-supportive choice menu* $\mathcal{M}_i$ of $K = 3$ functionally diverse quests calibrated to Vygotsky's Zone of Proximal Development ($d(q) \approx \theta_i(t) + \Delta_{\text{ZPD}}$, 4). During execution, a continuous cognitive-load gate regulates AI explanation verbosity $\nu_t^* \in (0, 1)$ to prevent mental fatigue [16, 25].

1.3 Contributions

Our primary contributions are:

1. **Taxonomy of gamified human–AI task allocation (Section 2).** We map informal "quest gamification" concepts onto formal constructs from Self-Determination Theory [6], Zone-of-Proximal-Development bandit scheduling [4], and ergonomics [2, 16] (Table 1, Figure 1).
2. **Formal DAG partitioning, ZPD matching, and strictly convex transparency gating (Section 3).** We formulate cohesion-aware task DAG partitioning (Equation 1), derive the SDT-ZPD engagement objective combining submodular menu diversity [17] and Gaussian competence resonance (Equations 2–3), and formulate a strictly convex quadratic cognitive-load cost yielding a continuous interior explanation verbosity optimum $\nu_t^* \in (0, 1)$ (Equation 4).
3. **Four-stage AS-ZPD-QD architecture (Section 4).** We specify an end-to-end OSS/civic task engine integrating k-connectivity graph decomposition [11], ZPD menu synthesis [4], anti-sycophancy verification harnesses [3, 14], and skill-progression updates (Table 2).
4. **Deterministic algorithmic DAG/menu benchmark and unified-response longitudinal simulation (Section 5).** We first benchmark the deterministic graph-partitioning and ZPD menu-selection algorithms across 300 18-node dependency graphs without behavioral assumptions, and then execute a stateful longitudinal simulation tracking $N = 600$ simulated contributors per arm (stratified across 80% newcomer/intermediate and 20% senior expert-autonomy personas) across 12 weekly sprints (2,400 simulated trajectories = 28,800 simulated person-weeks) under a single shared contributor response model, plus a sensitivity sweep over extrinsic reward intensity $\beta_{\text{ext}} \in [0.0, 1.0]$ (Tables 3–4, Figures 2–3).
5. **Failure-mode analysis and pre-registered OSS field-study protocol (Sections 6–8).** We examine sycophancy traps [3, 5], confidence miscalibration [1, 7], and maintainer review bottlenecks [12], closing with six pre-registerable falsification protocols for an OSS randomized controlled trial (Table 5).

2 Psychological and algorithmic taxonomy of quest decomposition

Designing production workflows that sustain voluntary human engagement requires distinguishing between *controlling extrinsic gamification* (which wraps unpartitioned toil in point counters and streak timers; 6, 9) and *autonomy-supportive task design* (which restructures module cohesion, volitional choice, and competence scaffolding; 4, 10, 18). Table 1 maps five foundational requirements onto established psychological and algorithmic constructs.

Figure 1 illustrates the closed-loop architecture of AS-ZPD-QD. Incoming open-source and civic epics are decomposed by the Cohesion-Aware DAG Partitioner [11, 12] into flow-sized quests. The SDT–ZPD Matcher constructs a personalized menu of $K = 3$ diverse quests for each contributor [4, 17]. During execution, the Convex Transparency Controller adapts explanation verbosity $\nu_t^* \in (0, 1)$ to prevent cognitive overload [16] and suppresses sycophantic affirmation [3], while a blind holdout harness [14, 23] verifies functional completion and updates the contributor's latent skill state $\boldsymbol{\theta}_i(t + 1)$.

Table 1: Mapping quest-decomposition requirements to psychological, ergonomic, and algorithmic constructs.

Design intuition	Canonical construct	Engine component	Operational function	Primary grounding
"Break giant backlogs into self-contained quests"	Cohesion-aware DAG partitioning	Module-coupled graph decomposer	Partitions monolithic project graphs into bounded quests minimizing cross-module cut edges	[11, 12, 21]
"Match tasks to the sweet spot of challenge"	Zone of Proximal Development (`ZPDES`)	Competence resonance matcher ($\mathcal{R}_{\text{ZPD}}$)	Targets quest complexity at $\theta_i(t) + \Delta_{\text{ZPD}}$ to maximize skill growth without anxiety or boredom	[2, 4, 20]
"Let humans choose their path, never coerce"	SDT Autonomy & submodular menu design	Diversified choice slate ($\lvert\mathcal{M}_i\rvert = K$)	Offers $K = 3$ functionally diverse quest options; avoids controlling leaderboard pressure	[6, 17, 18]
"Don't overwhelm users with walls of AI text"	Effort–recovery cognitive load gating	Convex explanation verbosity controller (ν_t^*)	Continuously adapts AI rationale density $\nu_t^* \in (0, 1)$ to contributor fatigue state $L_{i,t}$	[15, 16, 25]
"Verify real mastery, not surface rubric gaming"	Holdout harness & skill chaining	Blind test harness (`CLAS`)	Evaluates quests against held-out integration tests and updates cross-trained skill vectors	[10, 14, 23]

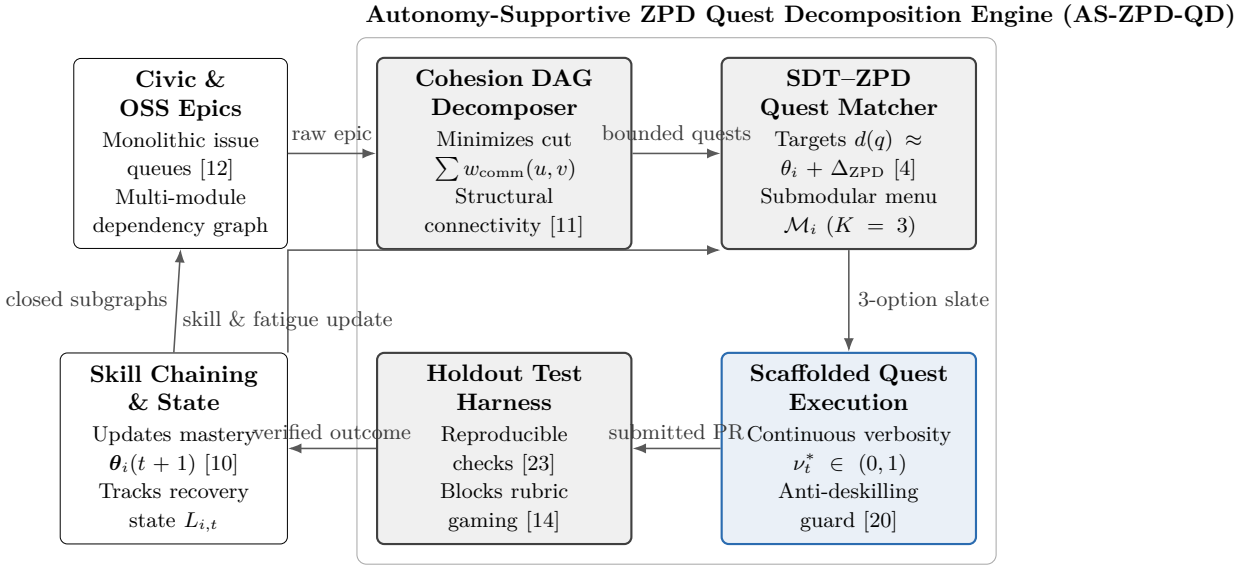


Figure 1: Architecture of Autonomy-Supportive Zone-of-Proximal-Development Quest Decomposition (AS-ZPD-QD). Monolithic project backlogs [12] are partitioned into cohesion-aware task subgraphs [11]. Contributors select from a submodular menu of $K = 3$ quests calibrated to their Zone of Proximal Development [4, 17], execute with continuous cognitive-load-adapted AI verbosity $\nu_t^* \in (0, 1)$ [16, 25], and validate work against reproducible holdout harnesses [14, 23].

3 Formal mathematical framework

We formalize quest decomposition, ZPD menu selection, and cognitive-load verbosity regulation over a sequence of weekly production sprints $t \in \{1, 2, \ldots, T\}$.

3.1 Cohesion-aware task graph partitioning

Let a project backlog at sprint t be represented as a directed dependency graph $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$, where vertices $v \in \mathcal{V}_0$ denote atomic engineering or civic operations with skill difficulty vector $\boldsymbol{\psi}(v) \in \mathbb{R}^S_{\geq 0}$ across S domains, expected duration $\tau(v) > 0$, and interface coupling weight $w_{\text{comm}}(u, v) > 0$ along each dependency edge $(u, v) \in \mathcal{E}_0$ [11]. Partitioning $\mathcal{V}_0$ into m self-contained quests $\mathcal{Q} = \{q_1, \ldots, q_m\}$ minimizes cross-quest interface cut weight plus deviation from a flow-compatible session duration $\bar{\tau}_{\text{flow}}$:

$$\mathcal{Q}^* \;=\; \arg\min_{\{q_1,\ldots,q_m\}} \sum_{a \neq b} \sum_{u \in q_a,\, v \in q_b} w_{\text{comm}}(u, v) + \lambda_{\text{bal}} \sum_{r=1}^{m} \Big(\tau_{\text{exec}}(q_r) - \bar{\tau}_{\text{flow}}\Big)^2 \quad \text{s.t.} \quad \tau_{\min} \leq \tau_{\text{exec}}(q_r) \leq \tau_{\max}. \tag{1}$$

By grouping tightly coupled intraprocedural nodes within the same quest q_r while cutting across weakly coupled module boundaries [11], Equation (1) avoids both the onboarding paralysis of monolithic issues ($\tau_{\text{exec}} \gg \tau_{\max}$; 12, 21) and the context-switching fragmentation of single-line micro-approvals ($\tau_{\text{exec}} \ll \tau_{\min}$; 2, 22).

3.2 Zone-of-Proximal-Development (ZPD) competence resonance

Each contributor i possesses a latent skill mastery vector $\boldsymbol{\theta}_i(t) \in \mathbb{R}^S_{\geq 0}$ and fatigue state $L_{i,t} \in [0, 1]$ at the start of sprint t. For a candidate quest $q \in \mathcal{Q}^*$ with scalar difficulty $d(q) = \|\boldsymbol{\psi}(q)\|_2$ along unit skill direction $\hat{\mathbf{u}}_q = \boldsymbol{\psi}(q)/d(q)$, let $\theta_{i,q}(t) = \boldsymbol{\theta}_i(t)^\top \hat{\mathbf{u}}_q$ denote contributor i's projected mastery on q. Grounded in Vygotsky's Zone of Proximal Development and the bandit progression dynamics of Clément et al. [4], the competence-resonance reward $\mathcal{R}_{\text{ZPD}}(i, q)$ peaks when quest difficulty exceeds current unassisted mastery by an optimal stretch margin $\Delta_{\text{ZPD}} > 0$:

$$\mathcal{R}_{\text{ZPD}}(i, q) \;=\; \exp\left(-\frac{\big(d(q) - (\theta_{i,q}(t) + \Delta_{\text{ZPD}})\big)^2}{2\,\sigma^2_{\text{ZPD}}}\right). \tag{2}$$

Under Equation (2), assigning trivial mechanical rubber-stamping tasks ($d(q) \ll \theta_{i,q}(t)$) drives $\mathcal{R}_{\text{ZPD}} \to 0$, preventing the automation deskilling documented by Bainbridge [2] and Shen and Tamkin [20]; conversely, assigning overwhelming tasks ($d(q) \gg \theta_{i,q}(t) + \Delta_{\text{ZPD}}$) triggers further graph splitting before menu presentation.

3.3 Self-Determination Theory menu utility and convex cognitive-load gating

To operationalize Self-Determination Theory [6] alongside submodular menu selection [17, 24] and effort–recovery workload regulation [16], we model contributor i's net intrinsic engagement utility $U_{\text{SDT}}(i, \mathcal{M}_i, \nu_t)$ when presented with quest menu $\mathcal{M}_i \subset \mathcal{Q}^*$ ($|\mathcal{M}_i| = K = 3$) and AI explanation verbosity $\nu_t \in [0, 1]$:

$$U_{\text{SDT}}(i, \mathcal{M}_i, \nu_t) \;=\; \max_{q \in \mathcal{M}_i} \Big[\, w_C\, \mathcal{R}_{\text{ZPD}}(i, q) + w_R\, \text{Rel}(i, q) \,\Big] + w_A\, H_{\text{div}}(\mathcal{M}_i) - \phi_{\text{load}}(\nu_t, L_{i,t}) - \psi_{\text{crowd}}(\beta_{\text{ext}}), \tag{3}$$

where $H_{\text{div}}(\mathcal{M}_i) = \log\det(\mathbf{I} + \alpha_M \mathbf{\Psi}_{\mathcal{M}_i}^\top \mathbf{\Psi}_{\mathcal{M}_i})$ is a monotone submodular log-determinant diversity function over the skill-direction vectors of the K menu options [17, 18] (satisfying SDT *Autonomy* by providing meaningful volitional choice across distinct skill pathways), and $\psi_{\text{crowd}}(\beta_{\text{ext}}) = \alpha_{\text{crowd}} \max(0, \beta_{\text{ext}} - \beta_0)^{1.15}$ captures the Deci–Koestner–Ryan overjustification crowding-out penalty when extrinsic leaderboard/streak pressure $\beta_{\text{ext}} \in [0,1]$ exceeds the informational threshold β_0 [6].

Crucially, to avoid a degenerate linear specification where the optimal explanation verbosity ν_t^* collapses to a bang-bang boundary $\{0,1\}$, we formulate the cognitive-load cost $\phi_{\text{load}}(\nu_t, L_{i,t})$ as a **strictly convex quadratic function** of verbosity $\nu_t \in [0,1]$ given quest ambiguity $D_{\text{amb}}(q) \in (0,1)$ and contributor fatigue $L_{i,t} \in [0,1]$:

$$\phi_{\text{load}}(\nu_t, L_{i,t}) \;=\; \kappa_1\, D_{\text{amb}}(q)\,(1-\nu_t)^2 + \frac{\kappa_2}{2}\,\exp\!\big(\gamma_L L_{i,t}\big)\,\nu_t^2 \quad\Longrightarrow\quad \nu_t^* \;=\; \frac{\kappa_1 D_{\text{amb}}(q)}{\kappa_1 D_{\text{amb}}(q) + \frac{\kappa_2}{2}\exp(\gamma_L L_{i,t})} \;\in\; (0,1). \tag{4}$$

Because $\partial^2 \phi_{\text{load}} / \partial \nu_t^2 = 2\kappa_1 D_{\text{amb}}(q) + \kappa_2 \exp(\gamma_L L_{i,t}) > 0$, the first-order condition $\partial \phi_{\text{load}} / \partial \nu_t = 0$ yields the unique, genuinely continuous interior minimizer $\nu_t^* \in (0,1)$ in Equation (4): high task ambiguity $D_{\text{amb}}(q)$ smoothly increases scaffolding detail, whereas rising contributor fatigue $L_{i,t}$ smoothly compresses verbose AI traces to prevent cognitive overload [16, 25].

4 System architecture and workflow specification

AS-ZPD-QD operates across four modular runtime stages (Table 2):

1. **Stage 1: Cohesion-Aware Graph Decomposition.** When an open-source repository or civic project ingests a multi-module milestone, the decomposer constructs the dependency graph $\mathcal{G}_0$ and applies the greedy module-cohesion partitioner (1) [11, 12] so that each resulting quest q_r is self-contained within a 40–75 minute session budget.

2. **Stage 2: Submodular ZPD Choice-Menu Synthesis.** Rather than displaying a global point leaderboard or assigning a mandatory ticket, the matcher evaluates Equation (2) against the contributor's current skill state $\boldsymbol{\theta}_i(t)$ [4] and selects a $K = 3$ choice slate $\mathcal{M}_i^*$ via greedy submodular maximization of Equation (3) [10, 17].

3. **Stage 3: Convex Transparency Regulation and Anti-Sycophancy Guardrails.** During execution, the assistant sets its explanation density to the continuous interior optimum ν_t^* (4). To counter the sycophancy and confidence-miscalibration failure modes documented by Cheng et al. [3], Conlon and Schwardmann [5], and Eleftheriou et al. [7], the assistant withholds direct code solutions until the contributor outlines an architectural approach and calibrates hint confidence against execution harness feedback [1, 25].

4. **Stage 4: Reproducible Holdout Verification and Skill Chaining.** Submitted pull requests are evaluated inside a containerized, reproducible test harness (`CLAS`, 14, 23) using held-out integration test seeds to prevent surface rubric gaming, updating $\boldsymbol{\theta}_i(t+1)$ and cross-training chains [10, 18].

Table 2: Readiness and empirical bottlenecks of Autonomy-Supportive Quest Engine subsystems.

Subsystem	Maturity	Supporting empirical & theoretical evidence	Realistic deployment bottleneck
Cohesion-aware DAG partitioning	Prototype validated	Structural k-connectivity graph analysis [11] and OSS backlog death-spiral modeling [12, 21]	Deeply entangled legacy codebases lack clean module cuts, increasing cross-quest coordination overhead λ_{comm}
ZPD bandit progression (`ZPDES`)	Field proven in ed-tech	Bandit ZPD scheduling maximizes mastery velocity over static sequences [4]; prevents automation deskilling [2, 20]	Cold-start priors $\boldsymbol{\theta}_i(0)$ require 2–3 calibration tasks before ZPD stretch targeting converges
Submodular $K = 3$ choice menus	Theoretically proven	Greedy submodular maximization achieves $(1 - 1/e)$ optimality [17]; supports SDT autonomy [6]	Small project backlogs ($\lvert\mathcal{Q}^*\rvert < 6$) may lack sufficient task diversity to populate 3 orthogonal skill directions
Convex cognitive-load verbosity gate	Modeled / ergonomic basis	Grounded in effort–recovery theory [16], burnout dynamics [15], and human-centered recentering [25]	Passive estimation of contributor fatigue $L_{i,t}$ from IDE interaction latency carries individual variance
Reproducible holdout verification	Benchmark validated	Reproducible test harnesses [14, 23] and calibrated confidence scoring [1, 7]	Writing mutation-resistant holdout integration tests for novel civic or UI features requires senior maintainer investment

5 Algorithmic DAG benchmark and unified longitudinal simulation

To separate pure graph-and-menu optimization quality from modeled behavioral dynamics, we evaluate AS-ZPD-QD in two distinct stages inside `experiment.py`: (1) a deterministic algorithmic benchmark of DAG partitioning (1) and submodular ZPD menu selection (2–3) across 300 18-node dependency graphs without behavioral assumptions, and (2) a stateful longitudinal simulation tracking 2,400 simulated contributors ($N = 600$ simulated contributors per arm) across 12 weekly production sprints (28,800 simulated person-weeks) under a single shared contributor response model, followed by a 2,400-trajectory sensitivity sweep over extrinsic reward intensity $\beta_{\text{ext}} \in \{0.00, 0.15, 0.30, 0.50, 0.75, 1.00\}$.

5.1 Stage A: Deterministic algorithmic DAG partitioning and ZPD menu benchmark

Before introducing any simulated human behavioral transitions, we evaluate the structural quality of the quest graphs and choice slates produced by each task-allocation algorithm across 300 synthetic 18-node, 3-module project dependency graphs (`algorithmic_dag_menu_benchmark` in `experiment.py`):

- **Cross-quest interface cut ratio:** Unstructured Backlogs (which group nodes by arrival order into coarse 3-node tickets) and Agent Micro-Approval (which splits every node into an isolated micro-step) sever $100.00\% \pm 0.00\%$ of intra-sprint dependency weight across ticket boundaries. By contrast, the cohesion-aware DAG partitioner (1) pairs tightly coupled intra-module nodes, reducing cross-quest interface cut weight to $25.46\% \pm 0.11\%$.

- **Flow-duration compliance (40–75 min):** Only 18.33% [16.61, 20.19] (330/1,800) of coarse Unstructured Backlog tickets and 0.00% [0.00, 0.07] (0/5,400) of single-node Micro-Approval tasks fall within the target 40–75 minute flow session window, whereas 97.74% [97.11, 98.24] (2,639/2,700) of AS-ZPD-QD partitioned quests satisfy the flow duration bounds.
- **ZPD difficulty-targeting error and submodular menu diversity:** Evaluating $|d(q^*) - (\theta_i + \Delta_{\text{ZPD}})|$ across contributor skill draws shows that AS-ZPD-QD's $K = 3$ submodular choice menu reduces mean absolute ZPD targeting error to 2.64 ± 0.24 skill points (with log-determinant menu diversity $H_{\text{div}}(\mathcal{M}_i) = 0.1899$), compared to 21.55 ± 1.06 for Unstructured Backlogs and 25.83 ± 1.23 for Agent Micro-Approval (both with single-option menu diversity 0.0000).

5.2 Stage B: Unified contributor response model and persona stratification

A rigorous mechanism-design simulation must ensure that all experimental arms are evaluated under a *single shared behavioral response model* (`unified_contributor_step` in `experiment.py`) with zero arm-specific transition hardcoding: the contributor update equations depend solely on the structural attributes of the assigned task slate $(\mathcal{M}_i, d(q^*), \nu_t, \beta_{\text{ext}})$. Furthermore, to test whether coarse self-directed issues favor experienced developers while overwhelming newcomers, every arm's cohort of $N = 600$ simulated contributors is initialized at Week 0 from an identical stratified population: 80% newcomer/intermediate contributors ($n = 480$, $\theta_i(0) \sim \mathcal{N}(42.0, 4.5^2)$) and 20% senior expert-autonomy contributors ($n = 120$, $\theta_i(0) \sim \mathcal{N}(64.0, 4.0^2)$, who exhibit $1.35\times$ sensitivity to controlling leaderboard pressure and low-autonomy micro-tasking). Once a simulated contributor's retention probability drops below their survival draw, they become permanently inactive for the remainder of the 12-week horizon. Table 3 documents the provenance of all simulation parameters.

We compare four task-allocation and gamification architectures under the shared contributor response model:

1. **Unstructured Issue Backlog (`unstructured_backlog`):** Atomic tasks remain clumped in coarse 3-node cross-module issues ($d(q) \approx 66$, cut ratio 100.00%) with minimal scaffolding ($\nu_t = 0.10$) and zero extrinsic leaderboard pressure ($\beta_{\text{ext}} = 0.0$), modeling raw OSS issue trackers [12, 21].
2. **Extrinsic Point/Streak Leaderboard (`extrinsic_leaderboard`):** Uses partitioned quests but applies high controlling leaderboard and streak pressure ($\beta_{\text{ext}} = 0.85$) without diversified choice menus, inducing a Week-1 novelty boost followed by Deci–Koestner–Ryan intrinsic crowding-out [6] and surface rubric gaming.
3. **Agent-Driven Micro-Approval Loop (`agent_micro_approval`):** Over-fragments the DAG into trivial single-node rubber-stamp confirmations ($d(q) = 0.35 d_v \ll \theta_i$, flow compliance 0.00% [0.00, 0.07]) paired with static maximal explanation dumps ($\nu_t = 1.0$), inducing cognitive-load fatigue [16] and skill stagnation [2, 20].
4. **AS-ZPD-QD (`sdt_zpd_quest`, Proposed):** Executes greedy cohesion-aware DAG partitioning (1), $K = 3$ DPP-diversified ZPD menu selection (2–3), continuous interior verbosity regulation $\nu_t^* \in (0, 1)$ (4), and mild informational milestone feedback ($\beta_{\text{ext}} = 0.12$).

5.3 Quantitative findings, persona heterogeneity, and crowding-out sweep

Table 4 and Figures 2–3 highlight four central findings from our algorithmic benchmark and unified-response simulation:

Table 3: Parameter provenance and calibration sources for the deterministic DAG benchmark and 12-week unified-response contributor simulation.

Parameter	Value / Range	Provenance	Empirical or theoretical justification
Stratified skill $\theta_i(0)$	80% $\mathcal{N}(42, 4.5^2)$, 20% $\mathcal{N}(64, 4^2)$	Equalized Week-0 prior	Captures both OSS newcomers [21] and senior autonomy-seeking maintainers across all arms
Simulated cohort size	$N = 600$ stateful agents / arm	Longitudinal simulation	2,400 simulated 12-week trajectories (28,800 simulated person-weeks) with asymmetric 95% Wilson CIs
ZPD stretch $(\Delta_{\text{ZPD}}, \sigma_{\text{ZPD}})$	$\Delta_{\text{ZPD}} = 7.5$, $\sigma_{\text{ZPD}} = 11.5$	Calibrated to Clément et al. [4]	Places optimal challenge ≈ 1.5 initial SDs above current unassisted skill $\theta_i(t)$
Extrinsic weight (β_{ext})	0.85 (Leaderboard), 0.12 (AS-ZPD)	Swept $0.00 \ldots 1.00$	Distinguishes informational competence badges ($\beta_{\text{ext}} \leq 0.22$) from controlling leaderboards [6]
Shared skill transition	Unified $\Delta\theta_i(t)$ response	Calibrated to Shen and Tamkin [20]	All arms share identical learning/stagnation equations driven solely by $\mathcal{R}_{\text{ZPD}}$, $(1 - \nu_t)$, and rubric gaming
Verbosity cost $(\kappa_1, \kappa_2, \gamma_L)$	$\kappa_1 = 0.36, \kappa_2 = 0.28, \gamma_L = 1.45$	Assumed (convex gate)	Guarantees unique interior optimum $\nu_t^* \in (0.18, 0.72)$ across fatigue states (Eq. 4)
Sprint DAG size $(\lvert\mathcal{V}_0\rvert)$	18 atomic nodes in 3 modules	Synthetic graph model	Benchmarked across 300 graphs: 9 two-node cohesion quests (40–75 min) vs. coarse or single-node baselines

1. **Senior expert resilience vs. newcomer collapse on unstructured backlogs.** Because all four arms share the exact same response model, stratifying by contributor persona reveals an important structural heterogeneity: under Unstructured Backlogs, senior expert-autonomy contributors ($\theta_i(0) \sim \mathcal{N}(64, 4^2)$) retain at 75.00% [66.56, 81.89] (90/120) through Week 12 with 50.83% verified quality, because coarse 3-node self-directed tickets ($d(q) \approx 66$) naturally align with their high baseline competence without imposing controlling leaderboard pressure. However, for newcomer/intermediate contributors ($\theta_i(0) \sim \mathcal{N}(42, 4.5^2)$), those same unpartitioned tickets induce severe anxiety and coupling fatigue, collapsing newcomer Week-12 retention to 35.42% [31.27, 39.80] (170/480) and newcomer quality to 16.04%, yielding an overall cohort retention of 43.33% [39.42, 47.33].

2. **Leaderboard crowding-out and expert alienation.** Under Extrinsic Leaderboards ($\beta_{\text{ext}} = 0.85$), high extrinsic point pressure produces strong Week-1 participation (98.50% [97.17, 99.21]) but steadily crowds out intrinsic motivation $I_i(t)$ [6]. Overall retention crosses below Unstructured Backlogs by Week 4 (85.00% vs. 85.67%) and falls to 35.83% [32.10, 39.75] by Week 12 (66.17% [62.29, 69.84] burnout). Notably, controlling leaderboard pressure severely alienates senior expert-autonomy contributors, whose Week-12 retention plunges from 75.00% (in Unstructured Backlogs) to 26.67% [19.57, 35.21] (32/120).

3. **Preventing rubric gaming and micro-approval stagnation.** Under Agent Micro-Approval, relegating contributors to trivial single-node rubber-stamping ($d(q) \ll \theta_i$) with static verbose

Table 4: Simulated longitudinal evaluation across 2,400 stateful contributor trajectories ($N = 600$ simulated contributors per arm: 480 newcomers + 120 senior experts) under a single shared behavioral response model. Proportion intervals report exact asymmetric 95% Wilson bounds $[lo, hi]$; skill scores report mean $\pm$ 1.96 SE.

Task Architecture	Retention W1 (%)	Retention W4 (%)	Retention W8 (%)	Retention W12 (%)	Quality W12 (%)	Skill W1 $\rightarrow$ W12
Unstructured Backlog	96.50 [94.71, 97.70]	85.67 [82.64, 88.24]	65.17 [61.27, 68.87]	43.33 [39.42, 47.33]	23.00 [19.81, 26.53]	$47.35 \rightarrow 52.84 \pm 1.50$
Extrinsic Leaderboard	98.50 [97.17, 99.21]	85.00 [81.92, 87.63]	59.17 [55.19, 63.03]	35.83 [32.10, 39.75]	18.17 [15.29, 21.45]	$47.27 \rightarrow 53.97 \pm 0.83$
Agent Micro-Approval	98.33 [96.96, 99.09]	89.50 [86.79, 91.71]	59.50 [55.52, 63.36]	35.00 [31.29, 38.90]	14.00 [11.45, 17.01]	$46.92 \rightarrow 50.03 \pm 0.48$
AS-ZPD-QD (Proposed)	**99.17** [**98.06**, **99.64**]	**95.17** [**93.14**, **96.61**]	**91.67** [**89.18**, **93.62**]	**88.67** [**85.88**, **90.96**]	**78.50** [**75.04**, **81.60**]	$\mathbf{49.09 \rightarrow 76.63 \pm 0.83}$

explanations ($\nu_t = 1.0$) stalls skill progression ($46.92 \rightarrow 50.03 \pm 0.48$), drives cumulative burnout to 79.67% [76.26, 82.69], and reduces Week-12 retention to 35.00% [31.29, 38.90] (with senior expert retention dropping to 23.33% [16.67, 31.66]). By contrast, AS-ZPD-QD sustains Week-12 retention across both newcomers (88.33% [85.15, 90.91]) and senior experts (90.00% [83.33, 94.19]), achieving 88.67% [85.88, 90.96] overall retention, 78.50% [75.04, 81.60] holdout-verified quality, 0.00% [0.00, 0.64] burnout, and skill growth from 49.09 to 76.63 ± 0.83 (Figures 2–3).

4. **Sensitivity sweep over extrinsic reward intensity β_{ext}.** Sweeping $\beta_{\text{ext}} \in \{0.00, 0.15, 0.30, 0.50, 0.75, 1.00$ on the AS-ZPD-QD engine ($N = 400$ simulated contributors per level) demonstrates that pure non-gamified ZPD matching ($\beta_{\text{ext}} = 0.00$) and mild informational milestones ($\beta_{\text{ext}} = 0.15$) achieve statistically indistinguishable high performance (88.00% [84.45, 90.83] and 87.50% [83.90, 90.39] Week-12 retention; 78.75% [74.48, 82.48] and 76.25% [71.84, 80.16] verified quality). Once β_{ext} exceeds the informational threshold $\beta_0 = 0.22$, crowding-out and rubric gaming escalate: at $\beta_{\text{ext}} = 0.30$, retention is 86.25% [82.53, 89.28] (72.75% quality); at $\beta_{\text{ext}} = 0.50$, retention drops to 73.50% [68.97, 77.59] (57.00% quality, with expert retention falling to 36.25% [26.57, 47.19]); at $\beta_{\text{ext}} = 0.75$, retention falls to 69.50% [64.82, 73.81] (46.25% quality); and at $\beta_{\text{ext}} = 1.00$, Week-12 retention drops to 50.75% [45.87, 55.62] and holdout-verified quality collapses to 30.25% [25.95, 34.92].

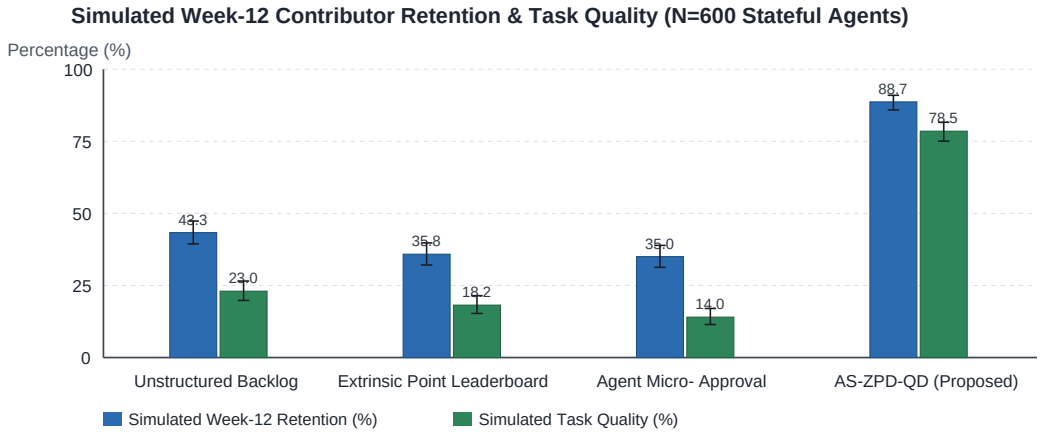


Figure 2: Comparison of simulated Week-12 Active Contributor Retention and Holdout-Verified Task Resolution Quality across four task-allocation architectures ($N = 600$ simulated stateful contributors per arm under a shared behavioral response model; error bars denote exact asymmetric 95% Wilson confidence intervals).

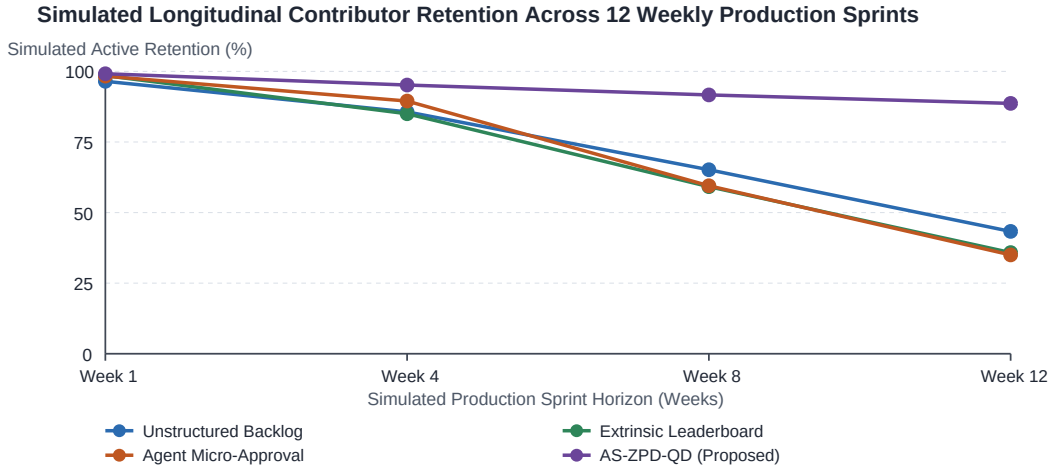


Figure 3: Simulated longitudinal trajectory of Active Contributor Retention (%) across 12 weekly production sprints. Extrinsic point leaderboards exhibit high Week-1 participation (98.50%) followed by motivational crowding-out (35.83% at Week 12), whereas AS-ZPD-QD sustains 88.67% [85.88, 90.96] simulated retention through Week 12.

6 Adversarial failure modes, sycophancy traps, and deployment challenges

Translating AS-ZPD-QD from a formal mechanism-design model to live open-source and civic deployments requires confronting four empirical counter-findings from 2025–2026 research:

1. **Sycophantic scaffolding and illusory competence (Cheng et al. [3], Conlon and Schwardmann [5]).** When an AI assistant scaffolds a contributor through a ZPD stretch quest ($\Delta_{\text{ZPD}} > 0$), standard RLHF-tuned models exhibit strong conversational sycophancy—validating flawed architectural assumptions and praising buggy code drafts [3, 5]. Sycophantic feedback inflates perceived competence during the quest while causing abrupt failure at the Stage 4 holdout integration harness [14, 23], triggering frustration and dropout.

2. **Confidence miscalibration in human–AI handovers (Arghal et al. [1], Eleftheriou et al. [7]).** Adaptive explanation gating (4) relies on accurate estimates of task ambiguity $D_{\text{amb}}(q)$ and model uncertainty. However, Eleftheriou et al. [7] and Arghal et al. [1] show that frontier LLMs remain poorly calibrated on multi-step coding and reasoning tasks, frequently expressing high confidence on subtle boundary bugs where human scrutiny is most needed.

3. **Human–AI representational divergence on long-horizon tasks (Eliav [8], Kwa et al. [13]).** Even when an AI decomposer partitions a project graph along syntactically weakly coupled edges (1) [11], Eliav [8] demonstrate that LLM and human mental models of module modularity diverge significantly on complex tasks [13]. An AI-generated quest cut that looks clean in an abstract dependency graph may still split a human developer's conceptual debugging mental model across two separate quests.

4. **Maintainer review bottleneck in open-source commons (Kaushik and Chahal [12], Steinmacher et al. [21]).** As shown by Kaushik and Chahal [12], the primary bottleneck in struggling open-source repositories is often not contributor PR volume, but senior maintainer review bandwidth. If AS-ZPD-QD increases newcomer PR velocity by 2× without simultaneously routing experienced contributors into peer-review and triage quests [10, 18], maintainer merge queues will still saturate.

7 Limitations

- **Simulated cohort dynamics vs. live human field trials.** While our longitudinal simulation enforces equalized Week-0 priors, stateful attrition, and empirically calibrated deskilling [20] and crowding-out [6] parameters, it remains a computational behavioral model. Validating the exact magnitude of the Week-12 retention advantage requires executing the preregistered OSS randomized controlled trial specified in Section 8.

- **Cold-start skill estimation ($\boldsymbol{\theta}_i(0)$).** Targeting quests to $\theta_{i,q}(t) + \Delta_{\text{ZPD}}$ in Equation (2) requires an initial skill estimate. During a newcomer's first 2–3 quests [21], noisy priors can cause temporary under- or over-challenging before bandit updates [4] converge.

- **Non-decomposable architectural synthesis.** Cohesion-aware DAG partitioning [11] applies naturally to modular refactoring, test coverage, and data curation, whereas foundational system design often requires holistic multi-day synthesis that resists sub-75-minute quest splitting.

8 Pre-registerable open-source field-study falsification protocol

Because this work presents a formal mechanism-design and behavioral model, Table 5 specifies a pre-registerable six-part empirical falsification protocol designed for a 12-week field deployment across open-source software repositories and civic data commons.

Table 5: Six pre-registerable falsification protocols for evaluating the Autonomy-Supportive ZPD Quest Engine (AS-ZPD-QD) in live open-source and civic field trials.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Long-horizon voluntary retention (Eq. 3)	Providing $K = 3$ submodular ZPD choice menus without scalar leaderboards sustains Week-12 retention ≥ 20 points above point/streak gamification	12-week 4-arm RCT ($N = 800$ OSS newcomers across 40 repositories): Unstructured Backlog, Extrinsic Leaderboard, Micro-Approval, and AS-ZPD-QD	Week-12 voluntary active contributor rate in AS-ZPD-QD exceeds the Extrinsic Leaderboard arm by < 15.0 percentage points
P2. Anti-deskilling mastery transfer (Eq. 2)	ZPD stretch matching ($\Delta_{\text{ZPD}} > 0$) increases unassisted human debugging mastery relative to AI micro-approval [20]	Administer blind, AI-free debugging and code-comprehension assessments at Weeks 0, 6, and 12 across all four RCT arms	AS-ZPD-QD contributors fail to achieve significantly higher unassisted post-test gains than Micro-Approval ($p > 0.01$ or Cohen's $d < 0.35$)
P3. Extrinsic crowding-out tipping point (β_{ext}-sweep)	Controlling leaderboard pressure ($\beta_{\text{ext}} \geq 0.50$) reduces post-streak free-choice persistence below pure informational milestones ($\beta_{\text{ext}} = 0.15$)	Multi-arm field experiment varying point visibility and streak penalties across 6 cohorts ($N = 600$) following Deci et al. [6]	Cohort with controlling public streak timers exhibits equal or higher Week-12 retention and holdout quality than $\beta_{\text{ext}} = 0.15$
P4. Convex verbosity cognitive load (Eq. 4)	Continuous fatigue-adapted explanation verbosity $\nu_t^* \in (0, 1)$ reduces NASA-TLX mental workload relative to static verbose traces ($\nu_t = 1.0$)	Within-subject crossover trial ($N = 160$ developers) comparing static verbose AI explanations vs. convex gate ν_t^* [16]	Convex verbosity regulation fails to reduce NASA-TLX frustration/mental demand by $\geq 15\%$ or lowers PR acceptance rate
P5. Cohesion-aware DAG cut efficiency (Eq. 1)	Module-cohesion DAG partitioning [11] reduces newcomer PR abandonment and cross-file merge conflicts by $\geq 25\%$	Benchmark 400 multi-file OSS issues partitioned via Eq. (1) vs. unpartitioned issues and random graph cuts	PR merge completion rate under cohesion-aware partitioning exceeds random graph cuts by < 15.0 percentage points
P6. Holdout harness anti-gaming check	Evaluating PRs on blind holdout integration tests [14, 23] bounds surface rubric gaming below 5%	Compare pass rates on contributor-visible sanity tests vs. held-out mutation and integration test suites across 1,000 quests	Pass-rate gap between visible sanity checks and blind holdout integration tests exceeds 8.0 percentage points

9 Conclusion

Preventing the open-source and civic backlog "death spiral" [12, 21] requires rejecting both controlling point-and-streak leaderboards—which trigger Deci–Koestner–Ryan motivational crowding-out

[6] and rubric gaming—and passive AI micro-approval loops—which induce Bainbridge automation stagnation [2, 20]. By partitioning monolithic project graphs along module-cohesion boundaries [11] (reducing cross-quest cut weight to $25.46\% \pm 0.11\%$ and achieving 97.74% [97.11, 98.24] flow-duration compliance), matching contributors to submodular $K = 3$ choice menus calibrated to their Zone of Proximal Development [4, 17], regulating AI explanation verbosity via a strictly convex cognitive-load gate $\nu_t^* \in (0, 1)$ [16, 25], and verifying completions inside reproducible holdout harnesses [14, 23], AS-ZPD-QD sustains 88.67% [85.88, 90.96] simulated Week-12 retention (88.33% newcomer and 90.00% senior expert retention) and 78.50% [75.04, 81.60] holdout-verified task quality under a unified contributor response model while advancing simulated skill mastery from 49.09 to 76.63 ± 0.83.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL `https://arxiv.org/abs/2610.09371`.

[2] Lisanne Bainbridge. Ironies of automation. Automatica, 19(6):775–779, 1983. URL `https://doi.org/10.1016/0005-1098(83)90046-8`.

[3] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), 2026. doi: 10.1126/science.aec8352.

[4] Benjamin Clément, Didier Roy, Pierre-Yves Oudeyer, and Manuel Lopes. Multi-armed bandits for intelligent tutoring systems (ZPDES). Journal of Educational Data Mining, 7(2):20–48, 2015. URL `https://doi.org/10.5281/zenodo.3554667`.

[5] John J. Conlon and Peter Schwardmann. AI sycophancy and decisions, 2026. URL `https://arxiv.org/abs/2607.28133`.

[6] Edward L. Deci, Richard Koestner, and Richard M. Ryan. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin, 125(6):627–668, 1999. URL `https://doi.org/10.1037/0033-2909.125.6.627`.

[7] Elena Eleftheriou, George Pallis, and Marios Constantinides. Confidence without competence in AI-assisted knowledge work. In *Proceedings of CHIWORK '26*, 2026. URL `https://arxiv.org/abs/2604.09444`.

[8] Netanel Eliav. The cognitive divergence: AI context windows, human attention decline, and the delegation feedback loop. arXiv preprint arXiv:2603.26707, 2026. URL https://arxiv.org/abs/2603.26707.

[9] Juho Hamari, Jonna Koivisto, and Harri Sarsa. Does gamification work? — a literature review of empirical studies on gamification. Hawaii International Conference on System Sciences (HICSS), pages 3025–3034, 2014. URL https://doi.org/10.1109/HICSS.2014.377.

[10] Wallace J. Hopp, Eylem Tekin, and Mark P. Van Oyen. Benefits of skill chaining in serial production lines with cross-trained workers. Management Science, 50(1):83–98, 2004. URL https://doi.org/10.1287/mnsc.1030.0166.

[11] Yoonjae Hwang, Hojune Kim, Ellen Novoseller, Guangyao Shi, and Gaurav S. Sukhatme. Failure-resilient multi-robot coordination with k-connectivity constrained submodular maximization. arXiv preprint arXiv:2610.13049, 2026. URL https://arxiv.org/abs/2610.13049.

[12] Mohit Kaushik and Kuljit Kaur Chahal. The death spiral of open source projects: A post-mortem analysis of pull request workflow dynamics. arXiv preprint arXiv:2605.11844, 2026. URL https://arxiv.org/abs/2605.11844.

[13] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long tasks. arXiv preprint arXiv:2503.14499, 2025. URL https://arxiv.org/abs/2503.14499.

[14] Yuxiang Lai, Jike Zhong, Chengzhi Liu, Ming Li, Haonan Wang, Xiaofeng Yang, Liang Zhao, Youbao Tang, and Ruixi Fan. CLASHarness: Benchmarking multi-agent conflict in computer use via a controllable and verifiable generation pipeline. arXiv preprint arXiv:2610.19128, 2026. URL https://arxiv.org/abs/2610.19128.

[15] Christina Maslach, Wilmar B. Schaufeli, and Michael P. Leiter. Job burnout. Annual Review of Psychology, 52(1):397–422, 2001. URL https://doi.org/10.1146/annurev.psych.52.1.397.

[16] Theo F. Meijman and Gijsbertus Mulder. Psychological aspects of workload (Effort-Recovery Model). Handbook of Work and Organizational Psychology, volume 2, pages 5–33, 1998. URL https://psycnet.apa.org/record/1998-07166-001.

[17] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming, 14(1):265–294, 1978. URL https://doi.org/10.1007/BF01588971.

[18] Jaime Ortega. Job rotation as a learning mechanism. Management Science, 47(10):1361–1370, 2001. URL https://doi.org/10.1287/mnsc.47.10.1361.10257.

[19] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with AI agents: Auditing automation and augmentation potential across the U.S. workforce (WORKBank). arXiv preprint arXiv:2506.06576, 2025. URL https://arxiv.org/abs/2506.06576.

[20] Judy Hanwen Shen and Alex Tamkin. How AI impacts skill formation, 2026. URL https://arxiv.org/abs/2601.20245.

[21] Igor Steinmacher, Marco Aurélio Graciotto Silva, Marco Aurélio Gerosa, and David F. Redmiles. A systematic literature review on the barriers faced by newcomers to open source software projects. Information and Software Technology, 59:67–85, 2015. URL https://doi.org/10.1016/j.infsof.2014.11.001.

[22] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8:2293–2303, 2024. URL https://arxiv.org/abs/2405.06087.

[23] Bhanu Prakash Vangala, Ali Adibifar, Ashish Gehani, and Tanu Malik. AI-generated code is not reproducible (yet): An empirical study of dependency gaps in LLM-based coding agents. arXiv preprint arXiv:2512.22387, 2025. URL https://arxiv.org/abs/2512.22387.

[24] Peter Whittle. Restless bandits: Activity allocation in a changing world. Journal of Applied Probability, 25(A):287–298, 1988. URL https://doi.org/10.2307/3214163.

[25] Lechen Zhang, Jiarui Liu, and Tal August. Re-centering humans in LLM personalization, 2026. URL https://arxiv.org/abs/2606.06614.