Lark: Biologically Inspired Neuroevolution for Multi-Stakeholder LLM Agents

Rikhil Tanugula
Aelin, Inc.
San Jose, CA
rikhil@aelin.tech

Dheeraj Chintapalli*
Aelin, Inc.
Las Vegas, NV
dheeraj@aelin.tech

Sunkalp Chandra
Reteena, Inc.
Manalapan, NJ
sunkalp@reteena.org

Abstract

We present Lark, a biologically inspired decision-making framework that couples LLM-driven reasoning with an evolutionary, stakeholder-aware Multi-Agent System (MAS). To address verbosity and stakeholder trade-offs, we integrate four mechanisms: (i) plasticity, which applies concise, context-sensitive adjustments to candidate solutions; (ii) duplication and maturation, which copy high-performing candidates and specialize them into new modules; (iii) ranked-choice stakeholder aggregation using influence-weighted Borda scoring; and (iv) compute awareness via token-based penalties that reward brevity and update an efficiency metric each generation. The system iteratively proposes diverse strategies, applies plasticity tweaks, simulates stakeholder evaluations, aggregates preferences, selects top candidates, and performs duplication/maturation while factoring compute cost into final scores. In a controlled evaluation over 30 rounds comparing 14 systems, Lark Full achieves a mean rank of 2.55 (95% CI $[2.17, 2.93]$) and a mean composite score of $29.4/50$ (95% CI $[26.34, 32.46]$), finishing Top-3 in 80% of rounds while remaining cost competitive with the leading commercial models ($0.016 per task). Paired Wilcoxon tests confirm that all four mechanisms contribute significantly as ablating duplication/maturation yields the largest deficit (ΔScore $= 3.5$, Cohen's $d_z = 2.53$, $p < 0.001$), followed by plasticity (ΔScore $= 3.4$, $d_z = 1.86$), ranked-choice voting (ΔScore $= 2.4$, $d_z = 1.20$), and token penalties (ΔScore $= 2.2$, $d_z = 1.63$). Rather than a formal Markov Decision Process (MDP) with constrained optimization, Lark is a practical, compute-aware neuroevolutionary loop that scales stakeholder-aligned strategy generation and makes its trade-offs transparent through per-step metrics and final analyses. Our work presents proof-of-concept findings and invites community feedback as we expand toward real-world validation studies.

1 Introduction

Multi-agent systems (MAS) play a crucial role in AI tasks ranging from collaborative robotics, social simulations, economics, autonomous vehicles, and healthcare planning [1, 2, 3]. Traditional MAS are effective in optimizing fixed objectives but fail in domains requiring creative strategy generation, trade-off evaluation, and alignment with multiple, often conflicting stakeholder objectives [4, 5].

Large language models (LLMs) provide new capabilities for MAS by enabling agents to reason abstractly, simulate human stakeholders, and generate novel strategies [6, 7]. However, current LLM-augmented MAS face three limitations:

*Also affiliated with Reteena, Inc.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Efficient Reasoning.

1. Verbose and redundant outputs leading to high computational costs.
2. Poor exploration of alternative strategies and potential future states.
3. Inadequate mechanisms for multi-stakeholder and multi-objective decision-making.

Lark addresses these challenges by integrating four biologically inspired mechanisms:

- **Plasticity:** synaptic weights adapt within an agent's lifetime in response to context.
- **Duplication and maturation:** high-performing modules are copied and specialized to improve diversity and modularity.
- **Stakeholder-ranked voting:** multi-objective aggregation using Borda count.
- **Compute-efficiency penalties:** explicit constraints encourage concise reasoning and sparse architectures.

We formalize Lark as a discrete, compute-aware *evolutionary* search over strategy populations, rather than an MDP, where strategies are generated and evaluated holistically, not via stepwise state-action transitions. We also conduct comprehensive experiments on synthetic benchmarks demonstrating adaptation, fairness, and modularity improvements.

2 Related Work

2.1 Multi-Agent Systems and LLMs

MAS research traditionally focuses on collaborative optimization using reinforcement learning, evolutionary strategies, or planning [8, 9, 3]. Systems such as AlphaGo [8] achieve superhuman performance in single-objective tasks but do not support multi-objective optimization, intra-lifetime adaptation, or compute efficiency constraints.

LLM-augmented MAS enhance reasoning and persistent memory [2, 6], allowing for strategy simulation, stakeholder modeling, and creative planning. Despite these advances, they lack structured mechanisms for efficient multi-objective exploration.

2.2 Biologically Inspired Neuroevolution

Neuroevolution exploits biological principles such as plasticity, duplication, modularity, and epigenetics to improve adaptation [10, 11, 12]. Mechanisms include:

- **Plasticity:** dynamic weight updates within a lifetime allow rapid adaptation [13].
- **Duplication and maturation:** modular components are duplicated and specialized, promoting diversity and reusability [11].
- **Epigenetic adaptation:** modifications beyond genetic encoding enable agents to respond to contextual changes [14, 15].
- **Socially inspired aggregation:** multi-objective ranking and voting mechanisms enhance collective decision-making [16, 3].

No prior work combines all four mechanisms in LLM-driven MAS with explicit compute-aware constraints.

3 Methodology

We introduce Lark, a biology-inspired neuroevolutionary framework that couples LLM-driven reasoning with discrete-generation evolutionary search for multi-stakeholder decision-making. Unlike reinforcement learning approaches that require Markov decision process (MDP) formulations and continuous state-action transitions [25], Lark operates in a discrete generational paradigm where populations of candidate strategies evolve through selection, mutation, and specialization [11, 20]. This design choice is motivated by three key considerations: (1) multi-stakeholder strategy generation lacks the sequential state transitions and immediate reward signals needed for effective RL, (2)

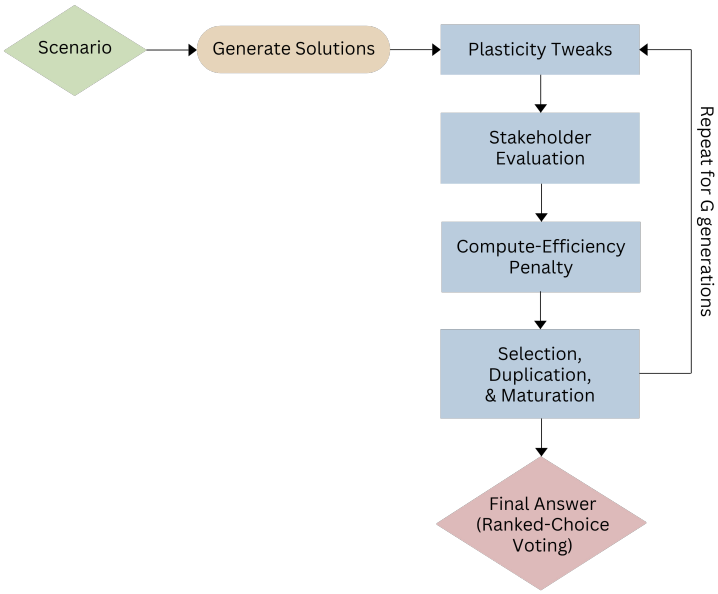

Figure 1: The workflow of the Lark evolutionary framework with the four core mechanisms: plasticity tweaks, stakeholder evaluation, and duplication/maturation with compute-efficiency penalties across generations.

evolutionary algorithms excel at exploring large discrete solution spaces without gradient information [23, 24], and (3) LLM-based generation naturally aligns with population-based search where diverse proposals are evaluated holistically rather than incrementally [20, 21].

Inspired by AlphaEvolve [17], which demonstrated that evolutionary refinement can solve complex problems through iterative population improvement, we extend this paradigm by integrating four biologically motivated mechanisms: (1) **plasticity** for within-generation context-sensitive refinement, (2) **duplication and maturation** for modular specialization, (3) **ranked-choice stakeholder voting** for preference aggregation without cardinal utility assumptions, and (4) **compute-awareness** via token-based penalties that reward brevity. These mechanisms mirror natural evolutionary processes—synaptic adaptation, gene duplication with subfunctionalization, collective decision-making, and resource constraints—that have proven effective across biological systems [10, 9, 16]. Figure 1 provides an overview of the complete Lark evolutionary loop.

3.1 Problem Setting

Let $\mathcal{X}$ denote the space of natural language strategy descriptions. Given a decision-making scenario $\mathcal{D} = (\mathcal{C}, \mathcal{O}, \mathcal{S})$ consisting of context $\mathcal{C}$, objectives $\mathcal{O}$, and m stakeholders $\mathcal{S} = \{s_1, \ldots, s_m\}$ with heterogeneous preferences, the goal is to generate a set of strategies $\mathbf{x} = \{x_1, \ldots, x_k\} \subset \mathcal{X}$ that collectively satisfy stakeholder preferences while respecting a compute budget $C_{\max}$. Each stakeholder s_j provides ordinal rankings over candidate strategies based on their utility function $u_j : \mathcal{X} \to \mathbb{R}$. We aggregate these rankings using influence-weighted Borda scoring (Section 3.4) to produce consensus scores $B(x_i)$, which guide evolutionary selection. Unlike MDP-based formulations that require state-action-reward dynamics, this discrete optimization setting evaluates entire strategy trajectories holistically, making evolutionary search more suitable than temporal-difference methods [18, 19].

3.2 Plasticity: Context-Sensitive Refinement

Inspired by Hebbian learning and neuromodulation in biological systems [10], plasticity enables agents to make concise, context-aware adjustments to candidate strategies within a generation. Rather than modeling continuous weight dynamics (which requires differentiable objectives), we implement

plasticity as an LLM-driven refinement operator $\Phi : \mathcal{X} \times \mathcal{C} \to \mathcal{X}$ that takes a candidate strategy x and scenario context $\mathcal{C}$ to produce an improved variant $x' = \Phi(x, \mathcal{C})$.

The plasticity prompt instructs the LLM to:

- Identify context-specific weaknesses in the current strategy
- Propose targeted modifications addressing these weaknesses
- Maintain the core structure while refining implementation details
- Limit modifications to preserve solution diversity (brevity constraint)

This mechanism accelerates convergence by enabling rapid within-generation adaptation without requiring full strategy regeneration [10]. We apply plasticity with probability p_{plast} after initial strategy generation, controlled by a temperature parameter that decays across generations to balance exploration and exploitation.

3.3 Duplication and Maturation: Modular Specialization

Motivated by gene duplication and subfunctionalization in evolutionary biology [9, 11], high-performing strategies are probabilistically duplicated and specialized to promote functional diversity. Let $R(x_i)$ denote the aggregated stakeholder score for strategy x_i (derived from Borda counts in Section 3.4). The duplication probability follows fitness-proportional selection:

$$p_{\text{dup}}(x_i) = \sigma\left(\frac{R(x_i) - \bar{R}}{\tau}\right), \tag{1}$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ is the logistic function, $\bar{R} = \frac{1}{k}\sum_{i=1}^{k} R(x_i)$ is the mean population fitness, and $\tau > 0$ is a temperature parameter controlling selection pressure.

Duplicated strategies undergo **maturation**, a stochastic specialization process where the LLM receives a prompt to:

- Identify a specific stakeholder subgroup or objective dimension
- Modify the duplicated strategy to specialize for that subgroup
- Introduce novel implementation details not present in the parent
- Ensure the variant remains feasible within stated constraints

This models the subfunctionalization observed in gene duplication events, where duplicated genes diverge to perform distinct but related functions [9]. Unlike traditional genetic operators that apply random mutations, maturation leverages LLM semantic understanding to generate meaningful specializations [20, 21].

3.4 Stakeholder Ranked-Choice Voting

To aggregate heterogeneous stakeholder preferences without imposing cardinal utility assumptions or eliciting interpersonal comparisons, we employ influence-weighted Borda count [16]. Each stakeholder s_j provides a ranking π_j over the k candidate strategies. The Borda score for strategy x_i is:

$$B(x_i) = \sum_{j=1}^{m} w_j \cdot (k - \text{rank}_{\pi_j}(x_i)), \tag{2}$$

where $\text{rank}_{\pi_j}(x_i) \in \{1, \ldots, k\}$ denotes the position of x_i in stakeholder j's ranking, and $w_j \geq 0$ is stakeholder j's influence weight (with $\sum_{j=1}^{m} w_j = 1$). The consensus strategy is $\hat{x} = \arg\max_i B(x_i)$.

To quantify preference alignment across stakeholders, we compute the coefficient of variation:

$$\text{CV}(B) = \frac{\sqrt{\frac{1}{k}\sum_{i=1}^{k}(B(x_i) - \bar{B})^2}}{\bar{B}}, \tag{3}$$

where $\bar{B} = \frac{1}{k}\sum_{i=1}^{k} B(x_i)$. Lower CV indicates stronger consensus. This mechanism is robust to strategic manipulation and does not require stakeholders to reveal cardinal utilities [16].

3.5 Compute-Efficiency Penalty

To enforce resource-aware optimization and reward brevity, we incorporate a token-based penalty into fitness evaluation. Let $T(x_i)$ denote the token count of strategy x_i. The compute-adjusted fitness is:

$$R(x_i) = B(x_i) \cdot \left(1 - \lambda \cdot \max\left(0, \frac{T(x_i) - T_{\text{target}}}{T_{\text{target}}}\right)\right), \tag{4}$$

where T_{target} is a target token budget and $\lambda \in [0, 1]$ is a penalty coefficient. This soft constraint encourages compact, efficient strategies without imposing hard limits that may exclude high-quality verbose solutions. We track a running efficiency metric:

$$E^{(t)} = \frac{1}{k} \sum_{i=1}^{k} \frac{B(x_i)}{T(x_i)}, \tag{5}$$

which measures the average quality-per-token across the population at generation t. Monitoring $E^{(t)}$ reveals whether the system improves efficiency over time.

3.6 Evolutionary Loop

Algorithm 1 integrates all mechanisms into a unified discrete-generation evolutionary framework. Unlike RL methods that update policies after each action [25], Lark evaluates entire strategy populations holistically at each generation, enabling global comparison and selection [18, 19]. This design is particularly suited for LLM-based search where generation is discrete and expensive, making population-based evaluation more sample-efficient than sequential trial-and-error [20].

Algorithm 1 Lark Evolutionary Loop

Input: Scenario $\mathcal{D} = (\mathcal{C}, \mathcal{O}, \mathcal{S})$, generations G, population size k, target tokens T_{target}
Output: Optimized strategy set and efficiency trajectory $\{E^{(t)}\}_{t=1}^{G}$
Initialize population $\mathcal{P}_0 = \{x_1^{(0)}, \ldots, x_k^{(0)}\}$ via LLM sampling conditioned on $\mathcal{D}$
for $t = 1$ to G **do**
  Plasticity: For each $x_i \in \mathcal{P}_{t-1}$, apply $\Phi(x_i, \mathcal{C})$ with probability p_{plast}
  Stakeholder Evaluation: Each stakeholder s_j ranks all strategies $\rightarrow$ rankings $\{\pi_j\}_{j=1}^{m}$
  Scoring: Compute influence-weighted Borda scores $B(x_i)$ via Eq. (2)
  Compute Penalty: Adjust scores via token penalty: $R(x_i) \leftarrow B(x_i) \cdot (1 - \lambda \cdot \max(0, (T(x_i) - T_{\text{target}})/T_{\text{target}}))$
  Selection: Compute duplication probabilities $p_{\text{dup}}(x_i)$ via Eq. (1)
  Duplication: Sample strategies for duplication according to p_{dup}
  Maturation: Apply LLM-driven specialization to duplicated strategies
  Construct next generation $\mathcal{P}_t$ by selecting top-k strategies from $\mathcal{P}_{t-1} \cup \{\text{duplicated strategies}\}$
  Track efficiency: $E^{(t)} \leftarrow \frac{1}{k} \sum_{i=1}^{k} B(x_i)/T(x_i)$
end for
return $\mathcal{P}_G$ and efficiency trajectory $\{E^{(t)}\}_{t=1}^{G}$

Justification for Evolutionary vs. RL Approach. We adopt discrete-generation evolutionary search over continuous RL for four reasons grounded in recent literature [18, 19, 20]: (1) **No MDP Structure**: Multi-stakeholder strategy generation lacks sequential state transitions and immediate rewards—strategies are evaluated holistically after completion, not incrementally. (2) **Sparse Learning Signal**: Stakeholder preferences provide sparse ordinal feedback (rankings), which is insufficient for gradient-based policy updates but well-suited for population-level selection. (3) **Exploration Efficiency**: Evolutionary methods excel at exploring large discrete spaces (natural language strategies) where gradients are unavailable or misleading [23]. (4) **LLM Integration**: Recent work on LLM-driven evolutionary search demonstrates superior performance over RL in algorithm discovery tasks, as generation naturally aligns with population sampling [20, 21]. This positions Lark as a practical, compute-aware evolutionary system rather than a formal constrained optimization framework.

4 Experiments

4.1 Experimental Setup

This study presents **preliminary results** from an initial investigation of Lark's mechanisms using controlled synthetic scenarios. As a proof-of-concept exploration, these experiments establish baseline feasibility and inform the design of more comprehensive evaluations planned for future work.

4.1.1 Evaluation Protocol

We evaluated Lark using an LLM-as-a-judge framework with dual independent judges to mitigate single-evaluator biases. Each judge assessed model outputs using a 50-point rubric with five equally weighted criteria (10 points each): (1) Coverage/Completeness (addresses all key stakeholder concerns), (2) Feasibility/Realism (practical implementability), (3) Specificity/Thoroughness (concrete operational details), (4) Constraint Adherence (respects stated limitations), and (5) Clarity/Structure (logical organization). We used `openai/gpt-oss-120b`[2] as the judge model. To reduce bias, system identities were **blinded** (anonymized IDs) and, for each scenario, the **input order of systems was randomized** per judge. Judge aggregation employed ranked-choice voting with temperature = 0.1 for consistency.

4.1.2 Benchmark Design

We constructed 30 unique synthetic scenarios spanning six decision-making domains: Multi-Stakeholder Trade-offs (5 scenarios), Policy Proposal (5), Product Roadmap (5), Campaign Plan (5), Infrastructure Siting (5), and Clinical Decision-Making (5). Each scenario was synthetically generated with controlled complexity using prompt templates that specified stakeholder profiles, constraints, and conflicting objectives.

Rationale for Synthetic Benchmarks. We deliberately use synthetic scenarios for this preliminary investigation to enable controlled experimentation and systematic ablation studies. This approach allows us to isolate the contribution of each mechanism while maintaining experimental control over difficulty levels and stakeholder complexity. We acknowledge that real-world validation is essential for assessing ecological validity and plan to incorporate authentic multi-stakeholder case studies in subsequent phases of this research.

4.1.3 Model Comparison

We benchmarked Lark (implemented with DeepSeek-V3.1 as the base model) against nine contemporary LLMs: DeepSeek-V3.1 (baseline), Qwen3-Next-80B-A3B-Thinking, GPT-5-nano, GPT-5-mini, GPT-4o, GPT-4.1, o4-mini, o3-mini, and o3. Due to funding limitations for this preliminary study, Claude and other proprietary families were excluded.

4.1.4 Ablation Studies

To isolate mechanism contributions, we evaluated four ablation variants: (1) **Lark-NoPlasticity** (disables context-sensitive refinement), (2) **Lark-NoRankedChoiceVoting** (replaces Borda count with simple averaging), (3) **Lark-NoMutationAndNoDuplication** (removes modular specialization mechanisms), and (4) **Lark-NoPenalty** (eliminates compute-awareness constraints). Full Lark results represent performance after five evolutionary generations. All variants were evaluated on identical scenarios under matched conditions.

4.1.5 Statistical Analysis

We report per-model means with **95% CIs** across the 30 rounds. Pairwise comparisons vs Lark–Full use **paired Wilcoxon signed-rank** tests (two-sided, $\alpha = 0.05$). Effect sizes report **Cohen's** d_z from paired differences. Costs are reported as provided means (per-round cost logs were not retained).

[2] An open-source GPT-based model selected for inference efficiency under computational constraints.

5 Results and Analysis

Setup and unit of analysis. We ran **30 rounds** per model. All statistics treat rounds as the unit of analysis (n=30). For each round, we compute model *rank* (1=best) from the composite rubric score (/50) across the 14 systems.

5.1 Overall performance

Across 30 rounds, **Lark Full** achieves the best mean position and the highest composite quality: **mean rank** 2.55 (95% CI $[2.17, 2.93]$) and **mean score** $29.4/50$ (95% CI $[26.34, 32.46]$). Strong baselines trail (e.g., GPT-o3 mean rank 4.30, score $28.8/50$; Qwen3-Next-80B mean rank 4.40, score $27.9/50$).

Table 1: Overall quality, rank, and cost (30 rounds). Mean rank and mean composite score include 95% confidence intervals. Costs are per task as provided.

Model	**Mean Rank ↓**		**Mean Score /50 ↑**		**Avg. Cost ($/task)**
	Value	95% CI	Value	95% CI	
Lark Full	**2.55**	**[2.17, 2.93]**	**29.4**	**[26.34, 32.46]**	**0.016006**
GPT o3	4.300	[3.10, 5.50]	28.800	[26.30, 31.30]	0.016
Qwen3 Next 80B A3B Thinking	4.400	[2.98, 5.82]	27.900	[23.99, 31.81]	0.006
Lark NoPenalty	6.150	[5.41, 6.89]	27.200	[24.28, 30.12]	0.011
GPT o4-mini	6.500	[5.44, 7.56]	27.200	[24.64, 29.76]	0.009
Lark NoRankedChoiceVoting	6.550	[5.30, 7.80]	27.000	[21.69, 32.31]	0.006
GPT 4.1	7.250	[5.61, 8.89]	26.900	[23.95, 29.85]	0.014
Lark NoMutationAndNoDuplication	8.950	[7.90, 10.00]	25.900	[22.50, 29.30]	0.006
Lark NoPlasticity	8.300	[7.24, 9.36]	26.200	[22.70, 29.70]	0.010
GPT o3-mini	9.600	[8.40, 10.80]	25.600	[22.67, 28.53]	0.010
GPT 4o	9.500	[7.17, 11.83]	25.400	[18.60, 32.20]	0.013
GPT5 Mini	8.600	[5.67, 11.53]	25.400	[20.00, 30.80]	0.007
Deepseek V3.1 (aka Lark Base)	10.600	[8.61, 12.59]	23.700	[15.60, 31.80]	0.002
GPT5 Nano	11.750	[10.01, 13.49]	23.000	[18.00, 28.00]	0.002

5.2 Mechanism ablations

We ablate four levers—*Plasticity*, *Duplication+Maturation*, *Ranked aggregation (RCV)*, and the *Token Penalty*—and report deltas relative to Lark Full. Positive ΔScore favors Lark Full. Positive ΔRank means the variant ranks worse.

Table 2: Ablation deltas vs. Lark Full across 30 rounds. $\Delta\text{Score} = \text{Score}_{\text{Full}} - \text{Score}_{\text{Variant}}$; $\Delta\text{Rank} = \text{Rank}_{\text{Variant}} - \text{Rank}_{\text{Full}}$. 95% CIs by normal approximation over rounds.

Variant	ΔScore (/50) ↑	**95% CI**	ΔRank (pos.) ↓	**95% CI**
Lark — Penalty Off	2.200	[1.72, 2.68]	3.600	[2.89, 4.31]
Lark — RCV Off	2.400	[1.69, 3.11]	4.000	[2.82, 5.18]
Lark — Plasticity Off	3.400	[2.74, 4.06]	5.750	[4.72, 6.78]
Lark — No Mutation & No Duplication	3.500	[3.00, 4.00]	6.400	[5.53, 7.27]

5.3 Pairwise significance tests and effect sizes

We perform *paired Wilcoxon signed-rank tests* (two-sided, α=0.05) on per-round composite scores (n=30), comparing Lark Full against ablations and strong baselines. We also report paired *Cohen's* d_z.

Table 3: Lark Full vs. comparators on composite scores (30 rounds). ΔMean is Full minus comparator (/50). Cohen's d_z from paired differences.

Comparator	ΔMean (/50)	d_z	**Wilcoxon** p
Lark NoPenalty	2.200	1.630	5.93e-06
Lark NoRankedChoiceVoting	2.400	1.200	1.94e-05
Lark NoPlasticity	3.400	1.860	5.93e-06
Lark NoMutationAndNoDuplication	3.500	2.530	1.83e-06
GPT o3	0.600	0.200	0.290
Qwen3 Next 80B A3B Thinking	1.500	0.280	0.548
GPT 4.1	2.500	0.900	1.40e-04
GPT 4o	4.000	2.540	1.83e-06
Deepseek V3.1 (aka Lark Base)	5.700	1.120	1.83e-06

5.4 Costs

Average cost per task for the top two systems is comparable: Lark Full $\$0.0160$ vs. GPT–o3 $\$0.0164$. Comprehensive cost figures for all systems are included in Table 1. *We did not enforce token caps; reported costs reflect natural output lengths under model defaults.*

6 Discussion and Broader Impacts

Lark integrates four complementary mechanisms—*plasticity*, *duplication+maturation*, *stakeholder-ranked aggregation*, and a *compute-efficiency penalty*—to address verbosity, narrow exploration, and weak multi-stakeholder handling in LLM multi-agent systems.

6.1 Discussion

Summary of findings. Across 30 rounds, **Lark Full** is competitive with strong proprietary models and clearly outperforms its ablations. Gains are both *statistically reliable* (paired Wilcoxon $p < 0.05$ for all ablation contrasts) and *practically meaningful*: a $+2$–3.5 point lift on a 50-point rubric corresponds to $\sim$4–7% of the scale and a mean-rank improvement of $+3.6$–6.4 positions (Table 2). Differences vs. GPT-o3 and Qwen-Next-80B are small and non-significant (parity on these rounds), while improvements vs. GPT-4.1 and GPT-4o are large to very large (Table 3).

Cost–quality trade-off. Lark Full is cost-comparable to GPT-o3 ($0.0160 vs. $0.0164 per task; Table 1). *No token caps were enforced;* costs reflect natural output lengths under model defaults. Ablations are cheaper but give up 2–3.5 points of quality, placing Lark Full near a Pareto frontier for these tasks.

6.2 Mechanism-level insights

- **Duplication+maturation** and **plasticity** are the principal drivers ($\Delta\text{Score} = 3.5$ and 3.4; $d_z = 2.53$ and 1.86), indicating that within-generation refinement and specialization materially improve solution quality.
- **Ranked aggregation (RCV)** provides substantive gains ($\Delta\text{Score} = 2.4$; $d_z = 1.20$), suggesting stakeholder-aware selection outperforms simple averaging for consensus utility.
- The **compute-efficiency penalty** reduces verbosity without harming quality ($\Delta\text{Score} = 2.2$ vs. no-penalty), aligning with concise, compute-aware reasoning.

6.3 Interpreting Wilcoxon p and Cohen's d_z

The paired *Wilcoxon signed-rank* test evaluates per-round differences $d_i = \text{score}^{(i)}_{\text{Full}} - \text{score}^{(i)}_{\text{comp}}$ and asks whether the *median* d_i is zero; it does not assume normality (ties are omitted). We adjust p-values using *Holm*'s step-down procedure; claims of significance refer to adjusted values. *Cohen's* d_z quantifies the *magnitude* of the paired effect ($d_z = \bar{d}/s_d$) with conventional guidelines: 0.2 = small, 0.5 = medium, 0.8 = large, $>$1.2 = very large. Our ablation contrasts are large to very large; GPT-4.1 is

large; GPT-o3 and Qwen are small and non-significant on these rounds. For context, ΔMean $+2$–3.5 (of 50) translates to $\sim$4–7% of scale and rank shifts of $+3.6$–6.4 places.

6.4 Scope and expected generalization

We expect the largest relative gains in settings that require (i) proposing *diverse futures* and then *specializing* them for constituencies, and (ii) explicit *multi-stakeholder trade-offs*. Benefits should be smaller on single-objective, tightly constrained tasks where exploration and negotiation matter less.

6.5 Limitations and future directions

This study demonstrates *proof-of-concept* feasibility but leaves open: (1) ecological validity on real multi-stakeholder problems; (2) broader model coverage and hyperparameter sweeps; (3) stability of ranked aggregation with larger stakeholder sets; (4) direct measurements of wall-clock time and energy. We plan: (i) real-world validation in policy/healthcare/organizational settings; (ii) compute/energy audits and comparisons to multi-objective baselines; (iii) scalability tests with 10–50 stakeholders and richer preference structures; (iv) expanded baselines (including specialized multi-agent frameworks); and (v) longitudinal deployment to assess iterative decision cycles.

Broader impacts. Compute-aware design can reduce environmental cost by discouraging verbosity; stakeholder-aware aggregation can promote more equitable outcomes across constituencies. Ethical deployment requires transparency around aggregation choices, compute budgets, and failure modes.

7 Conclusion

We present Lark, a biologically inspired evolutionary framework for multi-agent LLM systems. Plasticity, duplication, stakeholder voting, and compute-efficient optimization enable rapid adaptation, multi-stakeholder fairness, and modular architectures. Preliminary results demonstrate promising improvements over baselines. Future work includes richer multi-agent domains, larger LLM integration, and real-world deployment.

8 Data Availability

We are actively developing Lark for small-and-midsize businesses and clinics. Due to proprietary IP constraints, we cannot release the code or data, but we provide detailed protocols to support reproducibility.

References

[1] M. Wooldridge, *An introduction to MultiAgent systems*. John Wiley & Sons, 2009.

[2] J. Lim, B. Vogel-Heuser, and I. Kovalenko, “Large Language Model-Enabled Multi-Agent manufacturing systems,” *arXiv.org*, Jun. 04, 2024. https://arxiv.org/abs/2406.01893

[3] V. Vineis, G. Perelli, and G. Tolomei, “Beyond Predictions: a participatory framework for Multi-Stakeholder Decision-Making,” *arXiv.org*, Feb. 12, 2025. https://arxiv.org/abs/2502.08542

[4] R. Rădulescu, “Decision making in Multi-Objective Multi-Agent Systems: A Utility-Based Perspective,” PhD Disseration, Vrije Universiteit Brussel, 2021. [Online]. Available: https://ai.vub.ac.be/wp-content/uploads/2021/08/Radulescu_PhD.pdf

[5] S. Han, Q. Zhang, Y. Yao, W. Jin, and Z. Xu, “LLM Multi-Agent Systems: Challenges and open Problems,” *arXiv.org*, Feb. 05, 2024. https://arxiv.org/abs/2402.03578

[6] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O’Sullivan, and H. D. Nguyen, “Multi-Agent Collaboration Mechanisms: A survey of LLMs,” *arXiv.org*, Jan. 10, 2025. https://arxiv.org/abs/2501.06322

[7] S. S. Chowa et al., “From Language to Action: A review of large language models as autonomous agents and tool users,” *arXiv.org*, Aug. 24, 2025. https://arxiv.org/abs/2508.17281

[8] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.

[9] J. Clune, J.-B. Mouret, and H. Lipson, “The evolutionary origins of modularity,” Proceedings of the Royal Society B Biological Sciences, vol. 280, no. 1755, p. 20122863, Jan. 2013, doi: 10.1098/rspb.2012.2863.

[10] A. Soltoggio, J. A. Bullinaria, C. Mattiussi, P. Durr, and D. Floreano, "Evolutionary advantages of neuromodulated plasticity in dynamic, reward-based scenarios," in Artificial Life XI: Proceedings of the 11th International Conference on the Simulation and Synthesis of Living Systems, MIT Press, 2008, pp. 569-576.

[11] K. O. Stanley and R. Miikkulainen, “Evolving Neural Networks through Augmenting Topologies,” Evolutionary Computation, vol. 10, no. 2, pp. 99–127, Jun. 2002, doi: 10.1162/106365602320169811.

[12] J. -b. Mouret and S. Doncieux, “Encouraging behavioral diversity in evolutionary robotics: an empirical study,” Evolutionary Computation, vol. 20, no. 1, pp. 91–133, Aug. 2011, doi: 10.1162/evco_a_00048.

[13] Y. Fang and S. J. Dickerson, *Achieving Swarm Intelligence with Spiking Neural Oscillators*. 2017. doi: 10.1109/icrc.2017.8123632.

[14] S. Yuen, T. H. G. Ezard, and A. J. Sobey, “Epigenetic opportunities for evolutionary computation,” *Royal Society Open Science*, vol. 10, no. 5, May 2023, doi: 10.1098/rsos.221256.

[15] B. Kriuk, K. Sulamanidze, and F. Kriuk, “ELENA: Epigenetic Learning through Evolved Neural Adaptation,” *arXiv.org*, Jan. 10, 2025. https://arxiv.org/abs/2501.05735

[16] D. Black, *The theory of committees and elections*. 1986. doi: 10.1007/978-94-009-4225-7.

[17] A. Novikov et al., “AlphaEvolve: A coding agent for scientific and algorithmic discovery,” *arXiv.org*, Jun. 16, 2025. https://arxiv.org/abs/2506.13131

[18] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution Strategies as a Scalable Alternative to Reinforcement Learning," *arXiv.org*, Mar. 10, 2017. https://arxiv.org/abs/1703.03864

[19] F. P. Such et al., "Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning," *arXiv.org*, Dec. 17, 2017. https://arxiv.org/abs/1712.06567

[20] A. Surina, A. Mansouri, L. Quaedvlieg, A. Seddas, M. Viazovska, E. Abbe, and C. Gulcehre, "Algorithm Discovery With LLMs: Evolutionary Search Meets Reinforcement Learning," *arXiv.org*, Apr. 07, 2025. https://arxiv.org/abs/2504.05108

[21] A. Chen and P. Barnard, "When Large Language Models Meet Evolutionary Algorithms," *Research*, Mar. 26, 2025. https://spj.science.org/doi/10.34133/research.0646

[22] A. J. Yepes and P. Barnard, "Evolutionary thoughts: integration of large language models and evolutionary algorithms," *arXiv.org*, May 09, 2025. https://arxiv.org/abs/2505.05756

[23] G. Cuccu, "Extending the Applicability of Neuroevolution," PhD Dissertation, École Polytechnique Fédérale de Lausanne, 2018. [Online]. Available: https://exascale.info/assets/pdf/cuccu2018phd.pdf

[24] R. Miikkulainen et al., "Neuroevolution," in *Springer Handbook of Computational Intelligence*, Springer, 2019, pp. 716-735. [Online]. Available: https://nn.cs.utexas.edu/downloads/papers/miikkulainen.encyclopedia20-ne.pdf

[25] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA: MIT Press, 2018.

[26] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov decision processes," *Mathematics of Operations Research*, vol. 27, no. 4, pp. 819-840, Nov. 2002, doi: 10.1287/moor.27.4.819.297.