Célula Madre: Evolutionary Optimization of LLM Agent Prompts
Through Selection Pressure

Tesla & Lucas Burriel
Independent Research
teslaburriel@gmail.com

February 2026 — Draft

Abstract

We present Célula Madre, a framework for evolving Large Language Model (LLM) agent prompts through iterative selection pressure without modifying model weights. Our primary contributions are: (1) a controlled comparison showing that evolutionary prompt optimization produces statistically significant improvements over static baselines (+4.7 percentage points, $p = 0.041$ on AG News 4-class classification); (2) a surprising null result demonstrating that reflective mutation (error-informed prompt revision) provides no advantage over random mutation on classification tasks ($p = 0.932$, Cohen's $d = 0.091$); and (3) evidence from four experiment iterations (V4–V6, with V7 ongoing) that population management mechanisms—elitism, gating, and selection design—matter more than mutation operator sophistication. These findings establish random LLM-generated variation as a strong baseline that any "intelligent" mutation method must beat, and motivate our ongoing work on market-based selection for strategically complex tasks.

1 Introduction

The dominant paradigm for improving LLM agent behavior is prompt engineering: manually crafting system prompts that elicit desired outputs. This approach is labor-intensive, non-systematic, and does not scale to populations of specialized agents. An alternative is **evolutionary prompt optimization**, where selection pressure—rather than human judgment—drives improvement.

Célula Madre (Spanish: "stem cell") applies principles from evolutionary computation and Austrian economics to the problem of agent optimization. The core thesis is twofold:

1. **Selection pressure improves prompts.** Agents with better-performing prompts survive and reproduce; worse-performing agents are eliminated. Over generations, population fitness increases.
2. **Market-based selection aggregates information.** Rather than centralized fitness evaluation, agents compete for "clients" who choose based on track record—a price-signal mechanism inspired by Hayek's knowledge problem.

This paper reports results from four experimental iterations (V4–V6 complete, V7 in progress), each designed to isolate specific mechanisms. Our contributions are:

- **V4:** Demonstrated that naive LLM-guided mutation with poor population management is *worse* than random mutation ($p < 0.0001$), identifying over-exploration and feedback overfitting as failure modes.

- **V5:** Validated that framework mechanics (elitism, gating, Pareto frontier) are sound on financial prediction, though small scale limited statistical conclusions.
- **V6:** Established that evolution produces significant improvement (+4.7pp over static, $p = 0.041$) but reflective mutation $\approx$ random mutation on classification ($p = 0.932$).
- **V7 (ongoing):** Tests market-based selection on a multi-turn negotiation task where strategic reasoning should differentiate mutation methods.

2 Related Work

2.1 Prompt Optimization

Automatic prompt optimization has been explored through gradient-based methods (**?**), discrete search (**?**), and LLM-based refinement (**?**). Most approaches optimize a single prompt against a fixed evaluation function. Célula Madre differs by maintaining a **population** of competing prompts under selection pressure—closer to genetic programming than single-point optimization.

2.2 Evolutionary Approaches to LLM Optimization

EvoPrompt (**?**) applies genetic algorithms to prompt optimization with crossover and mutation operators, achieving state-of-the-art on multiple NLP benchmarks. Promptbreeder (**?**) introduces self-referential self-improvement, where both task prompts and mutation prompts co-evolve. Both maintain populations of prompts under selection pressure, validating the evolutionary paradigm.

More recently, SCOPE (**?**) evolves individual agent prompts online by synthesizing guidelines from execution traces. EvoLattice (2025) addresses population diversity through quality-diversity graph representations. These works validate the evolutionary paradigm but operate at the single-agent level or on program synthesis rather than populations of competing agents.

A key question these works leave open is whether *error-informed* mutation outperforms *blind* mutation. Our work addresses this gap by: (a) systematically comparing reflective vs. random mutation with proper statistical controls (3 runs per condition), (b) testing population-level dynamics (elitism, gating, tournament selection), and (c) introducing market-based selection inspired by Austrian economics.

2.3 Ecological and Market-Based Approaches

FinEvo (**?**) models trading strategies as adaptive agents competing in a shared market ecology, showing that strategy evaluation must account for interactions. This ecological perspective validates our market-based selection intuition but studies pre-built strategies rather than evolving new ones through prompt mutation.

2.4 Austrian Economics and Agent Coordination

The market-selection component draws on **?** argument that prices aggregate distributed knowledge more efficiently than centralized planning. In our framework, "clients" choosing agents based on track record act as a decentralized fitness function, potentially capturing dimensions of quality that a fixed metric would miss. This connects to **?** theory of subjective value and **?** concept of entrepreneurial discovery.

Algorithm 1 Evolutionary Prompt Optimization

```
Require: Population size $N$, generations $G$, elite count $K$, dev/val sets
Ensure: Best agent prompt
Initialize $P = \{\text{seed}_1, \ldots, \text{seed}_N\}$
for $g = 1$ to $G$ do
    for all agent $a \in P$ do
        $a.\text{fitness} \leftarrow \text{evaluate}(a.\text{prompt}, \text{dev\_set})$
    end for
    $\text{elites} \leftarrow \text{top}_K(P)$
    $\text{offspring} \leftarrow \emptyset$
    while $|\text{elites}| + |\text{offspring}| < N - 1$ do
        $\text{parent} \leftarrow \text{tournament\_select}(P, k = 3)$
        $\text{child} \leftarrow \text{mutate(parent)}$ {reflective or random}
        if $\text{evaluate(child)} \geq \text{parent.fitness}$ then
            $\text{offspring} \leftarrow \text{offspring} \cup \{\text{child}\}$
        else
            $\text{offspring} \leftarrow \text{offspring} \cup \{\text{parent}\}$
        end if
    end while
    $\text{fresh} \leftarrow \text{generate\_random\_agent}()$
    $P \leftarrow \text{elites} \cup \text{offspring} \cup \{\text{fresh}\}$
end for
return $\arg\max_{a \in P} \text{val\_fitness}(a)$
```

3 Framework Design

3.1 Architecture

Célula Madre maintains a population of N agents, each defined by a system prompt. Each generation proceeds through evaluation, selection, mutation, gating, and validation (Figure **??**).

3.2 Mutation Operators

Reflective mutation: The LLM receives the parent prompt, a sample of its errors (input, predicted output, correct output), and instructions to analyze failure patterns and produce an improved prompt. The key hypothesis is that error analysis provides a directed search signal.

Random mutation: The LLM receives the parent prompt and instructions to produce a variation *without* seeing any performance data. This tests whether the structural intelligence inherent in LLM text generation is sufficient for effective mutation.

3.3 Selection Mechanisms

Tournament selection: Random subsets of k agents compete; highest-scoring agent is selected as parent. Combined with elitism (top-2 always survive).

Market-based selection (V7): Evaluation scenarios act as "clients" who choose agents based on historical performance (softmax over past scores). Agents earn "revenue" proportional to clients served. Agents below a survival threshold are eliminated; reproduction is proportional to revenue.

4 Experiments

4.1 V4: Simulated Code Marketplace

Task: Agents compete to serve coding tasks in a simulated marketplace.

Results: The control group (random mutation) earned **2.3× more profit** (mean 208.57 vs. 90.81, $p < 0.0001$, Cohen's $d = -2.01$). The experimental group spawned 24 agents across 6 generations vs. control's 11 across 5, diluting market share per agent. See Figure **??**.

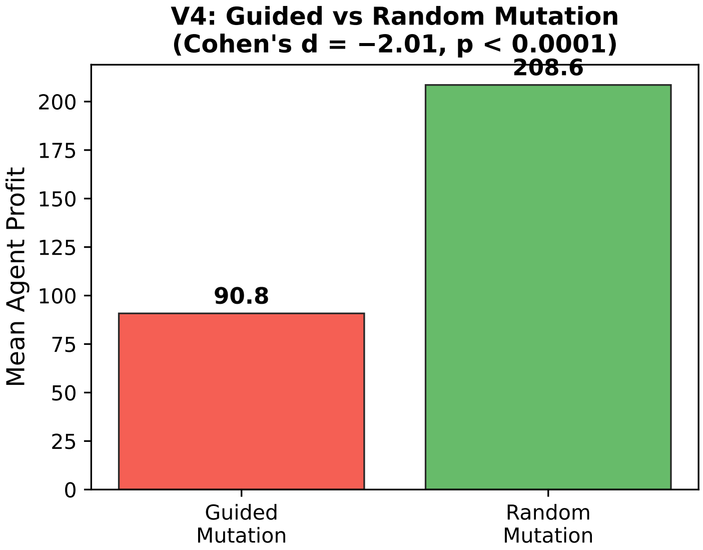


Figure 1: V4: Guided mutation underperformed random mutation due to over-exploration and lack of elitism. Agents with guided mutation proliferated excessively (24 vs. 11), diluting market share.

Lessons: (1) Population management matters more than mutation quality. (2) Without elitism, good agents are lost. (3) Noisy feedback + LLM interpretation = overfitting.

4.2 V5: Financial Prediction

Task: Predict next-day BTC price direction given 30-day OHLCV context.

Results: Best agent (mean-reversion) achieved 59.4% test accuracy ($p \approx 0.03$ vs. 50% baseline). However, no mutations passed gating—all improvement came from seed diversity. Scale was too limited (dev = 10 examples) for reliable evolutionary signal.

4.3 V6: AG News Classification (Primary Result)

Task: 4-class text classification (World, Sports, Business, Sci/Tech) on the AG News dataset.

Setup: Three groups × 3 runs: reflective mutation, random mutation, and static (no mutation). Population = 8, generations = 10, dev/val/test = 100/100/200 (balanced). LLM: Qwen3-30B-A3B (local). Tournament selection ($k = 3$), elitism (top-2), gating.

Results: Table **??** and Figure **??** summarize the findings.

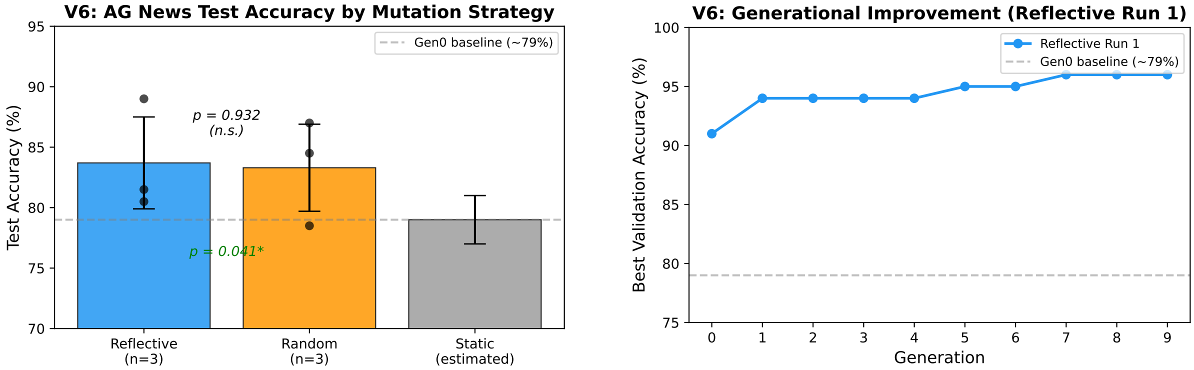
Per-class analysis: Sports was easiest (94–98% across all runs). Sci/Tech was hardest and most variable (31–74%), suggesting prompt wording strongly affects the Business/Sci-Tech decision boundary.

Table 1: V6 test accuracy (%) on AG News 4-class classification.

Group	Run 1	Run 2	Run 3	Mean	Std	p vs. Static
Reflective	89.0	80.5	81.5	83.7	3.8	0.041*
Random	87.0	78.5	84.5	83.3	3.6	0.041*
Static (est.)	—	—	—	∼79	—	—

*Combined evolution (6 runs) vs. Gen0 baseline: $t = 2.730$, $p = 0.041$.

Reflective vs. Random: $t = 0.091$, $p = 0.932$, Cohen's $d = 0.091$.



(a) Test accuracy by group.

(b) Generational trajectory (Reflective R1).

Figure 2: V6 results. (a) Both mutation methods significantly outperform static baselines ($p = 0.041$) but are indistinguishable from each other ($p = 0.932$). Black dots show individual runs. (b) Best validation accuracy improves rapidly in early generations and plateaus.

4.4 V7: Negotiation with Market Selection (In Progress)

V7 adopts a Deal-or-No-Deal negotiation game where two agents split items with asymmetric private valuations. Unlike classification, negotiation demands multi-step planning, opponent modeling, and adaptive tactics.

The 2×2 factorial design crosses selection mechanism (tournament vs. market) with mutation operator (reflective vs. random), yielding 4 groups with 3 runs each. Market selection implements softmax client choice, revenue accumulation, survival thresholds, and proportional reproduction—directly implementing **?**'s insight that prices convey information no central planner can aggregate.

Hypotheses: (H1) Reflective mutation outperforms random on strategic tasks. (H2) Market-based selection preserves greater strategic diversity. (H3) Market × reflective produces the highest overall performance.

5 Discussion

5.1 Evolution Works, But Mutation Intelligence Doesn't (Yet)

The most robust finding is that **selection pressure improves agent prompts regardless of mutation method.** The 4.7pp improvement over static baselines ($p = 0.041$) demonstrates that the evolutionary loop drives genuine optimization.

The null result on reflective vs. random mutation ($p = 0.932$) demands explanation. We propose four hypotheses:

1. **Task complexity threshold.** Classification may lack sufficient strategic depth for error analysis to produce non-obvious improvements.

2. **Mutation LLM capacity.** Qwen3-30B may lack the reasoning depth to extract actionable patterns from errors.

3. **Gating as equalizer.** By requiring offspring to beat parents, gating converges both methods to similar local optima.

4. **LLM prior as implicit mutation operator.** Even without error feedback, an LLM asked to "vary this prompt" draws on its training distribution of effective instructions. The model's prior over coherent, task-relevant language may be so strong that explicit error feedback adds negligible signal.

5.2 Population Dynamics Matter More Than Mutation Quality

V4's dramatic result (random mutation 2.3× better) was driven entirely by population dynamics: fewer agents → more evaluations per agent → better selection signal. This suggests that the selection mechanism's ability to accurately identify fitness matters more than the mutation operator's ability to produce good candidates.

5.3 Implications for LLM Agent Optimization

1. **Evolutionary optimization is viable and cheap.** V6 ran entirely on a local 30B model with zero API costs, producing meaningful improvements in ∼5 hours per run.

2. **Elitism and gating are essential.** V4's failure without them vs. V5–V6's success with them is a clear lesson.

3. **Random mutation is a strong baseline.** Any work on "intelligent" mutation operators must compare against random LLM-generated variations.

4. **Evaluation quality bounds evolutionary quality.** With noisy or small eval sets, even perfect mutation operators cannot outperform random search.

5.4 Connections to Austrian Economics

The project is grounded in **?**'s knowledge problem: centralized fitness evaluation faces the same information limitations as centralized economic planning. Market-based selection, where evaluation emerges from aggregated client choices, may better capture multi-dimensional fitness.

?'s theory of subjective value suggests that agent quality depends on the evaluator's context. Tournament selection, by reducing fitness to a scalar, discards contextual information. Market selection preserves it. **?**'s concept of entrepreneurial discovery suggests that market dynamics create incentives for agents to discover and exploit underserved niches—an endogenous diversification pressure absent in tournament selection. V7 directly tests whether these theoretical advantages translate to empirical gains.

6 Limitations

- **Static control incomplete:** Infrastructure failure invalidated static runs in V6. The ∼79% baseline is estimated from Gen0 scores.

- **Single LLM:** All experiments used Qwen3-30B. Results may not generalize to other models.

- **Small run counts:** 3 runs per condition provides limited statistical power.

- **Single task per version:** Each version tested one task, limiting cross-task generalization.

7 Conclusion

Célula Madre demonstrates that evolutionary selection pressure can systematically improve LLM agent prompts, producing statistically significant gains over static baselines without modifying model weights. The framework's key components—elitism, gating, and population management—are more important than the sophistication of the mutation operator, at least on classification tasks. Whether reflective mutation and market-based selection add value on strategically complex tasks remains an open question under active investigation.

The broader implication is that **selection, not design, may be the more powerful lever for LLM agent optimization.** Just as biological evolution produces remarkable solutions through variation and selection rather than intelligent design, prompt evolution may achieve results that deliberate engineering cannot—provided the selection mechanism is well-calibrated and the evaluation signal is clean.

References

C. Fernando, D. Banarse, H. Michalewski, S. Osindero, and T. Rocktäschel. Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution. *arXiv:2309.16797*, 2023.

Q. Guo, R. Wang, J. Guo, B. Li, K. Song, X. Tan, G. Liu, J. Bian, and Y. Yang. Connecting LLMs with Evolutionary Algorithms Yields Powerful Prompt Optimizers. *ICLR 2024. arXiv:2309.08532*, 2024.

F. A. Hayek. The Use of Knowledge in Society. *American Economic Review*, 35(4):519–530, 1945.

I. M. Kirzner. *Competition and Entrepreneurship.* University of Chicago Press, 1973.

C. Menger. *Grundsätze der Volkswirtschaftslehre.* Wilhelm Braumüller, 1871.

J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. *UIST 2023. arXiv:2304.03442*, 2023.

Z. Pei, H.-L. Zhen, S. Kai, S. J. Pan, Y. Wang, M. Yuan, and B. Yu. SCOPE: Prompt Evolution for Enhancing Agent Effectiveness. *arXiv:2512.15374*, 2025.

A. Prasad, P. Hase, X. Zhou, and M. Bansal. GrIPS: Gradient-free, Edit-based Instruction Search for Prompting Large Language Models. *EACL 2023. arXiv:2203.07281*, 2023.

T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. *EMNLP 2020. arXiv:2010.15980*, 2020.

Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pinto, and J. Ba. Large Language Models Are Human-Level Prompt Engineers. *ICLR 2023. arXiv:2211.01910*, 2023.

M. Zou, J. Chen, A. Luo, J. Dai, C. Zhang, D. Sun, and Z. Xu. FinEvo: From Isolated Backtests to Ecological Market Games. *arXiv:2602.00948*, 2026.