What Makes an LLM a Good Optimizer?

A Trajectory Analysis of LLM-Guided Evolutionary Search

Xinhao Zhang[1], Xi Chen[1,2], François Portet[1], Maxime Peyrard[1]

Grenoble Computer Science Lab, CNRS[1] Orange Innovation Grenoble[2]

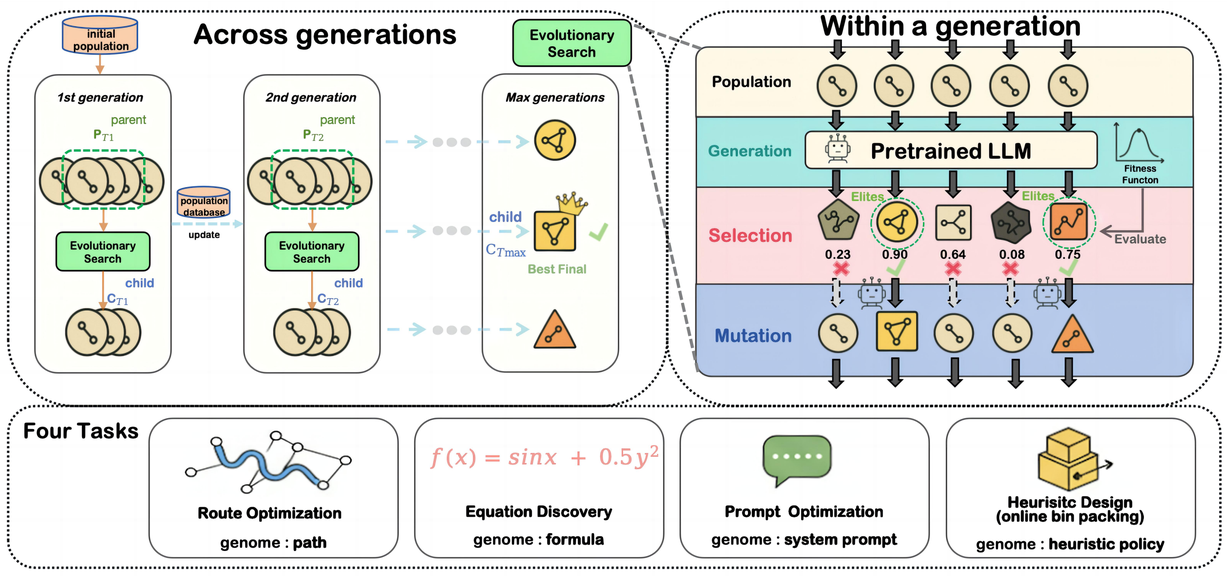
Advanced Language Processing School - ALPS

Motivation

- LLMs increasingly used as search operators within iterative optimization loops in a range of tasks, such as Automatic Prompt Engineering (APE) and Automatic Heuristic Design (AHD)
- Yet under the identical and controlled evolutionary loop, different LLMs still produce vastly different trajectories and outcomes
- Main Research Question: ***what characterize a good LLM optimizer?***

Experimental Setup

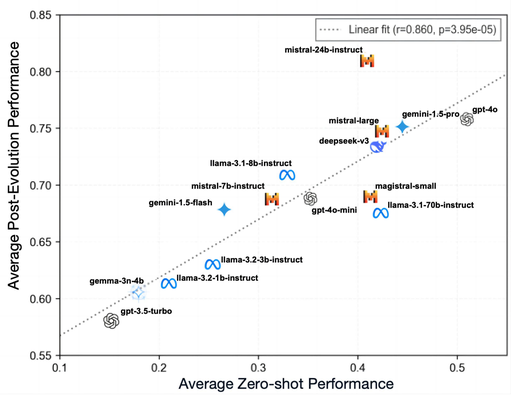
- Population initialization → Fitness-propotional selection → LLM-instructed mutation → Pool update
- Analysis scale up to 15 LLMs on 8 tasks across 4 domains, logging 72K+ candidates over 30 generations



Correlational Findings

✓ *Finding #1 Base capability matters, but is not enough*

- *SubRQ1 : Are model differences only a reflection of base model capability?*
- Zero-shot performance (temperature-swept best sample) psitively correlates with final performance
- Models with similar starting with similar staring ability diverge after optimization.



✓ *Finding #2 more novelty ≠ better optimization*

- We compute task-specific semantic distance between each child and thier parents
- Avg novelty of trajectories shows no correlation
- High novelty exploration fails to obtain fitness gains

✓ *Finding #3 breakthrough frequency is a key predictor*

- Breakthrough as a best-so-far imrpovement event
- Average breakthrough rate provided 3 times more explanatory power than with zero-shot performance and shows significant correlation
- Strong optimizers do not rely on rare big jumps, they consistently produce small and incremental improvements

Semantic Geometric Mechanism

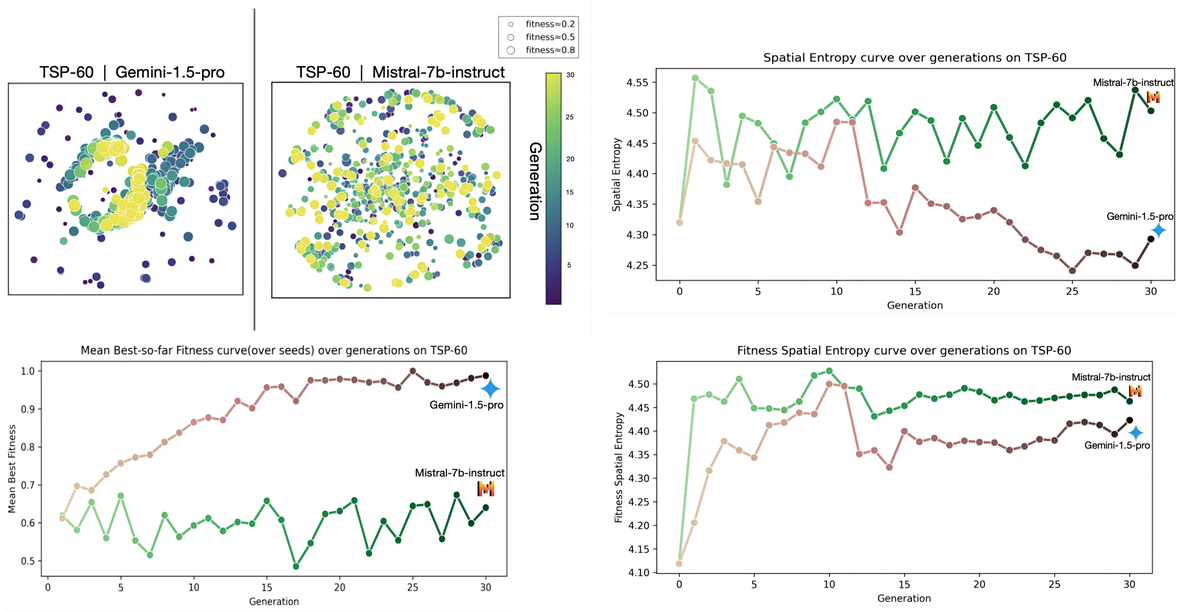
➢ ***SubRQ2 :*** *Why do some models' trajectories have more breakthroughs?*

➢ **Methods**

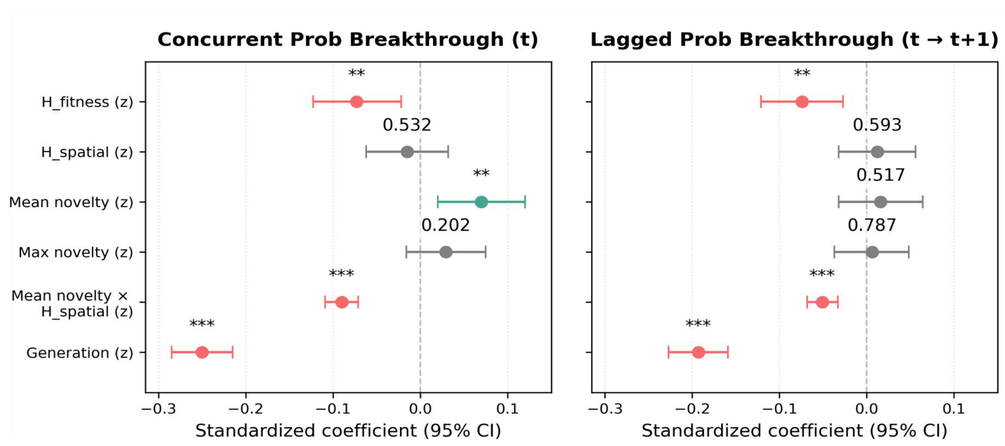
- We employ Multi-dimensional Scaling to project all the high-dimensinal solution embeddings from the trajectory into 2D semantic space
- We compute the spatial orgainization of search using kernel-based entropy on two different ways (H_{spatial} spatial entropy and H_{fitness} fitness spatial entropy)

➢ **Case Study on TSP-60**

- Strong optimizers*:* progressively localize search, exploit promosing regions, convert variation into improvements
- Weak optimizers: keep difting across distant regions, generate novelty without progress



Generation-Level Statistical Evidences



- Breakthroughs mostly occur in early generations
- Even maintainning multiple dispersed high-performing regions would hinder breakthrough production
- **Interaction Deeper Insights:** novelty increases the probability of breakthroughs only when search remains sufficiently localized

Implications and Future Work

- We find that **"optimizable ability"** is a distinct ability from "problem-solving ability", enabling future research on more effective optimizer
- Optimization quality depends more on **search dynamics**, not just model size; smaller models can still be competitive if they refine well
- On the role of novelty in LLM-Guided search, novelty acts as an immediate driver of exploratory breakthroughs, but its long-term utility depends on the search regime
- Good LLM-based optimizers as **local refiners**

Project Webpage 👇

😊Feel happy to discuss the future possibility followed by the contributions of this work, including but not limited to :

- Controllable search dynamics
- Training LLMs as search operators
- Explainablity of agentic AI systems

GETALP

Groupe d'Étude en Traduction Automatique
Traitement Automatisé des Langues et de la Parole

Laboratoire d'Informatique de Grenoble

2026