Reasoning-Guided Evolutionary Prompt Optimization for Improved Financial Problem Solving

Leandro A. Loss[1,2] **Pratikkumar Dhuvad**[2]
[1]ESSCA School of Management [2]AML RightSource
[leandro.loss,pratikkumar.dhuvad]@amlrightsource.com

Abstract

Prompt quality remains a primary bottleneck for deploying Large Language Models (LLMs) in high-stakes domains such as finance. Prior automated prompt optimization work has relied on ad-hoc heuristics or on LLM evaluators that lack explicit, stepwise reasoning, limiting the quality of discovered prompts. We introduce a Genetic Algorithm (GA) framework that uses *thinking models* (OpenAI's GPT-omni and GPT-5 variants) both to generate candidate prompts (initialization, crossover, mutation) and to evaluate their outputs, so evolution is guided by models that perform structured, multi-step inference. We evaluate this approach on a challenging Financial Math Reasoning benchmark, comparing GPT-5, GPT-5-mini, GPT5-nano, and GPT-o4-mini against non-thinking baselines, GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, and GPT-4o. Fitness computation was standardized using GPT-5-nano as output evaluator, creating a consistent test bed for comparisons. Our results show that reasoning-enabled optimization consistently produces stronger prompts than non-thinking optimization and manually engineered prompts. More specifically, we show in this study that GA-evolved prompts exceeded manual prompts in 7 out of 8 model versions and yielded an average around 11% higher fitness over non-thinking baseline. These findings demonstrate that combining evolutionary search with reasoning-capable LLMs substantially improves automated prompt engineering for financial reasoning tasks.

1 Introduction

LLMs have rapidly become foundational in natural language processing, powering applications across domains such as healthcare, law, and finance [1, 17]. Despite their capabilities, performance remains strongly dependent on high-quality input prompts. Manual prompt engineering is labor-intensive, vendor-specific, and often infeasible for teams without domain expertise [10].

To address this challenge, recent studies have explored automated prompt optimization. Gradient-guided methods such as AutoPrompt [15], reinforcement learning approaches like RLPrompt [4], and evolutionary strategies such as EvoPrompting [2] have all shown measurable success. Previously reported work has demonstrated that GAs can autonomously evolve prompts by replacing traditional operators with LLM-driven meta-prompts [11]. Later extensions introduced self-evaluating pipelines where LLMs also judge outputs for fitness, reducing reliance on human labels [12].

However, a critical limitation in prior GA frameworks lies in the *type* of models used as evaluators and evolutionary operators. Non-thinking models such as GPT-4.1 can generate prompts and evaluate answers but often lack the reasoning depth to judge correctness on nuanced tasks like financial mathematics. This introduces noise into optimization and constrains achievable fitness.

In this work, we introduce thinking models, reasoning-capable LLMs such as GPT-5, as both the *engine* and the *judge* of GA-based prompt optimization. Thinking models generate more structured and

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

semantically consistent offspring during initialization, crossover and mutation, while simultaneously serving as higher-quality evaluators of candidate prompt outputs. Our experiments on a Financial Math Reasoning dataset show that this dual use of reasoning-capable LLMs improves average fitness by 11% compared to non-thinking GPT-4.x baselines, establishing a new benchmark for automated prompt engineering in finance.

2 Related Literature

Prompt optimization: Manual prompt design remains a limiting factor for scaling LLMs across industries [10]. Automated methods span gradient-based optimization [15], reinforcement learning [4, 19], and meta-prompting frameworks [8, 18]. While effective, these approaches often require human supervision or handcrafted reward functions.

Evolutionary computation: GAs are well-suited for non-linear, large search spaces [7, 5]. Integration with LLMs is increasingly studied: Meyerson et al. [13] demonstrated crossover guided by language models, while Chen et al. [2] applied evolutionary search to neural architectures. Guo et al. [6] and Hsieh et al. [9] confirmed the effectiveness of GA-driven prompt optimization. Prior works established GA frameworks with either string-matching fitness [11] or non-thinking LLM judges [12].

Algorithm I: LLM-based GA for Prompt Optimization

0: **input**→ population size, generations, mutation prob., *LLM name*, *data*
1: Initialize population with *system*, *init. meta-prompts*, and data
2: Evaluate fitness using *fitness* and *system fitness prompts*
3: While generations not reached:
4: Selection via trio tournament
5: Pair crossover using *crossover meta-prompt*
6: Conditional mutation using *mutation meta-prompt*
7: Re-evaluate fitness
8: Increment generation count
9: End While loop
10: **output**← prompt with highest fitness

(a)

Role	Prompt
system	You are an AI that helps people solve problems. Avoid comments outside the proposed prompt as the user will use your answer to integrate with another downstream system.
initial.	Create a [LLM_MODEL] prompt that solves the problem exemplified by the following examples: [SAMPLE_QUESTION_1][SAMPLE_ANSWER_1] [SAMPLE_QUESTION_1][SAMPLE_ANSWER_2] ...
crossover	Given the following drafts for two prompts that aim to solve the same particular problem, create a better prompt using only ideas from them. [PROMPT_A] [PROMPT_B]
mutation	Given the following draft for a prompt that aims to solve a particular problem, create a better prompt using ideas from it. [PROMPT_A]
system fitness	You are an AI that validates automated answers against ground truth. You only answer 'yes' or 'no' with no extra comments, notes, or explanations as the user will use your answer to integrate with another downstream system.
fitness	My ground truth is [TRUTH]. Does the automated output [ANSWER] linguistically, symbolically, conceptually, or fundamentally match my ground truth?

(b)

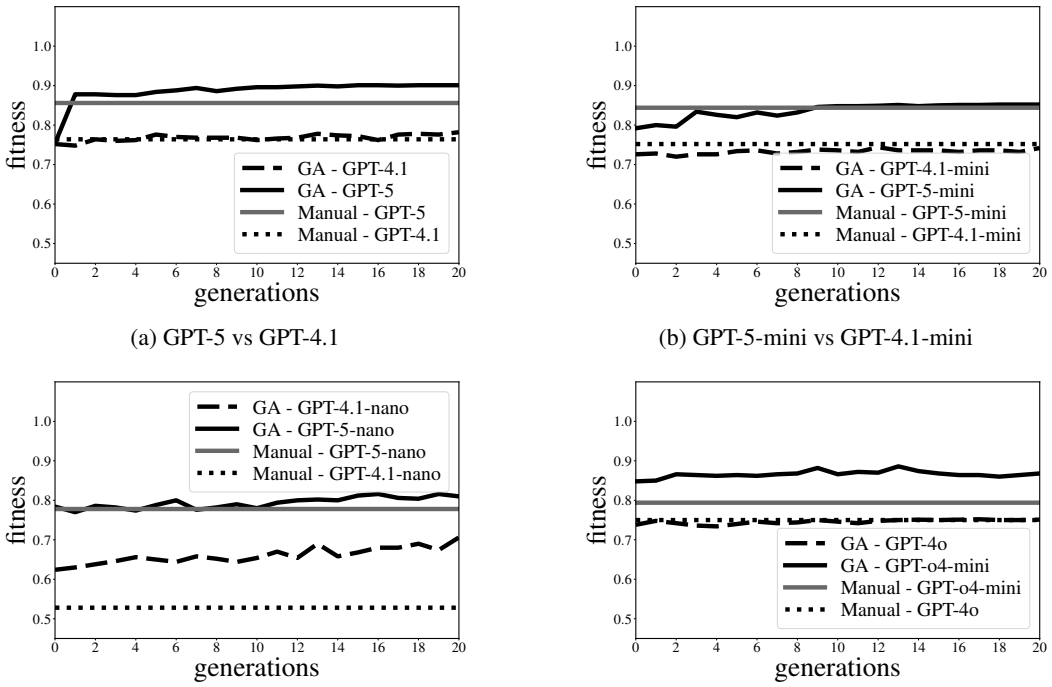
Table 1: The GA and meta-prompt designed used in [12]. (a) Algorithmic steps. (b) Meta-prompts guiding initialization, crossover, mutation, and fitness evaluation.

Thinking models: Recent reasoning-capable LLMs (e.g., GPT-5) provide explicit stepwise inference and structured decision-making [14]. Prior research has shown that chain-of-thought prompting can improve reasoning in non-thinking models [16], but our work is the first to systematically compare thinking vs. non-thinking models within GA-driven prompt optimization, using them simultaneously as evolutionary operators and fitness evaluators.

3 Proposed Method

We extend prior GA-based prompt optimization [11, 12] by using reasoning-enabled ("thinking") models to drive *both* candidate generation and fitness evaluation, so evolution can create and reward prompts that elicit multi-step reasoning.

GA with Thinking Models: GA implementation is outlined in Table. 1a where each individual is a candidate prompt. In short: (i) initialization uses thinking models with in-context examples; (ii) crossover/mutation operate via meta-prompts (Table. 1b) with reasoning models guiding semantic coherence; (iii) selection uses reasoning-grounded fitness so the GA rewards task-relevant, multi-step



(a) GPT-5 vs GPT-4.1

(b) GPT-5-mini vs GPT-4.1-mini

(c) GPT-5-nano vs GPT-4.1-nano

(d) GPT-o4-mini vs GPT-4o

Figure 1: Comparison of GA best fitness across generations. All models were optimized and judged with GPT-5-nano as the fitness function.

behaviors. This dual role makes generation and evaluation mutually informed while keeping the description compact.

Fitness Function: Fitness is measured as the average accuracy over a 500-sample evaluation set. All outputs are judged by GPT-5-nano using concise binary (yes/no) meta-prompts, providing a consistent baseline across models. While this standardization ensures fair comparison, it also introduces the limitation that fitness scores may reflect biases or variability in the judge model itself.

Dataset: We use the Financial Math Reasoning dataset [3] ($\sim$1,500+ pairs), the same benchmark used in our prior works to preserve comparability. Notably, in our earlier studies this has been the only dataset where LLM-based optimization underperformed manual prompting, largely because tasks require preprocessing intermediate results before the final computation. In this work, we address that gap by supplying the GA with reasoning-enabled models better suited for step-by-step problem solving.

Models Compared: We evaluate OpenAI's thinking models GPT-5, GPT-5-mini, GPT-5-nano, and GPT-o4-mini, against some of their non-thinking models GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, and GPT-4o.

Experimental Setup: GA parameters mirror prior work with population size of 20, 20 evolving generations, and trio-tournament selection. While we also track average population fitness across runs, for simplicity we report only best-of-generation results in the tables and figures. Textual comments on averages are added when relevant. Results are contrasted between pairs of thinking and non-thinking models, as well as against manual prompting. While GA runs are initialized with problem-agnostic meta-prompts, the manual prompts serve as a human-engineered benchmark, allowing us to test whether autonomous evolution can match or surpass expert-designed prompts.

4 Results and Discussion

Among eight LLM variants evaluated on the Financial Math Reasoning dataset, GA evolution yielded best-performing prompts that surpassed manual prompts in 7 cases, with an average best improvement of $+4.3\%$. Average population gains, however, were modest ($+0.9\%$). This indicates that the GA

Table 2: Financial Math results: manual prompts vs. GA-evolved prompts. Here, "Manual" represents the results achieved by manually engineered prompts, "GA Best" represents the fitness by the best prompt evolved by the evolutionary algorithm, and "Diff-Best" represents the difference between GA Best and Manual. Fitness of all individuals was judged by GPT-5-nano.

Model	Manual	GA Best	Diff-Best
GPT-4.1	0.764	**0.782**	0.018
GPT-4.1-mini	**0.752**	0.742	-0.010
GPT-4.1-nano	0.528	**0.706**	0.178
GPT-4o	0.750	**0.751**	0.001
GPT-5	0.856	**0.901**	0.045
GPT-5-mini	0.844	**0.852**	0.008
GPT-5-nano	0.778	**0.810**	0.032
GPT-o4-mini	0.794	**0.868**	0.074
Overall Average	0.7583	**0.8015**	0.0433
Non-thinking Average	0.6985	**0.7453**	0.0468
Thinking Average	0.8180	**0.8578**	0.0398

reliably finds high-quality individuals while population-level quality remains model-dependent. When comparing thinking and non-thinking models, the former began at a substantially higher manual baseline (average manual fitness ≈ 0.818 vs. 0.699 for non-thinking models), so absolute improvements for thinking models are smaller even when evolution finds better prompts. We also observe that the smaller variants derived the largest relative gains (e.g., GPT-4.1-nano: $+0.178$), suggesting the GA is particularly useful when base-model prompting is weak.

While prior work demonstrated that GA-based prompting can surpass manual baselines in general NLP settings [11, 12], performance on the Financial Math Reasoning benchmark remained inconsistent. The present study closes this gap by showing consistent gains in 7 out of 8 model variants when reasoning-capable LLMs are used as both operators and evaluators. The GA increases the fitness of the best individuals more consistently than it does for the population average. This implies the method is effective at exploring promising prompt heuristics but that additional mechanisms (e.g., diversity maintenance, elitism tuning, or adaptive mutation) are likely required to shift the entire population distribution.

We employed GPT-5-nano as a judge in fitness evaluation to keep measurement consistent. However, LLM judges can misjudge equality and introduce bias and variance in fitness estimates [12]. Also, for Financial Math Reasoning, manual prompts that explicitly elicit Chain-of-Thought (CoT) can outperform automated candidates in some strong models [11, 16]. Therefore, future automated searches should consider including CoT-style meta-operators or explicitly expose CoT as a target behavior.

Table 2 reports fitness on the Financial Math Reasoning dataset for non-thinking models (GPT-4.1 family, GPT-4o) and thinking models (GPT-5 family, GPT-o4-mini). For each model, we compare manual prompting with the best GA-evolved individual. Diff-Best measures the gain (or loss) relative to manual prompting.

5 Conclusion

This study showed that integrating reasoning-enabled LLMs (GPT-5s and omni) into GA-based prompt optimization leads to more reliable evolution and evaluation of prompts for financial reasoning. Unlike earlier work that used string-matching fitness [11] or non-thinking evaluators [12], this study demonstrates that "thinking models" reduce noise in fitness and generate semantically coherent offspring, yielding consistent gains: GA prompts outperform manual prompts in 7 out of 8 cases, and results in fitness averages around 11% higher than GPT-4.1 and GPT-4o baselines.

These results highlight the practical role of evolutionary prompting: it can substantially boost weaker models while still improving stronger ones. Future work will focus on diversity-preserving operators, richer fitness objectives, and explicit chain-of-thought meta-prompts [16] to further automate and generalize prompt optimization in finance and beyond.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[2] Angelica Chen, David Dohan, and David So. Evoprompting: Language models for code-level neural architecture search. In *Advances in Neural Information Processing Systems*, volume 36, pages 7787–7817, 2023.

[3] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *Proceedings of EMNLP 2022*, 2022.

[4] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proc. of the 2022 Conf. on Empirical Methods in NLP*, pages 3369–3391, 2023.

[5] David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[6] Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In *International Conference on Learning Representations (ICLR)*, 2024.

[7] John H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[8] Yutai Hou, Hongyuan Dong, Xinghao Wang, Bohan Li, and Wanxiang Che. Metaprompting: Learning to learn better prompts. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3251–3262, 2022.

[9] Cho-Jui Hsieh, Si Si, Felix X Yu, and Inderjit S Dhillon. Automatic engineering of long prompts. In *Findings of the Association for Computational Linguistics*, pages 10672–10685, 2024.

[10] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 100–114, 2022.

[11] Leandro A. Loss and Pratikkumar Dhuvad. From manual to automated prompt engineering: Evolving llm prompts with genetic algorithms. In *Proceedings of the IEEE Congress on Evolutionary Computation (CEC)*, 2025.

[12] Leandro A. Loss and Pratikkumar Dhuvad. An llm-based genetic algorithm for prompt engineering. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO Companion)*, 2025.

[13] Elliot Meyerson, Mark J Nelson, Herbie Bradley, Adam Gaier, Arash Moradi, Amy K Hoover, and Joel Lehman. Language model crossover: Variation through few-shot prompting. In *ACM Transactions on Evolutionary Learning and Optimization*, volume 4, pages 1–40, 2023.

[14] OpenAI. Hello gpt-5. `https://openai.com`, 2024. Accessed 2025-08-26.

[15] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proc. of the 2020 Conference on Empirical Methods in NLP*, pages 4222–4235, 2020.

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[17] Stephen Wu, Ozan Irsoy, Shengjie Lu, Volodymyr Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, Daniel Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[18] Qinyuan Ye, Maxamed Axmed, Reid Pryzant, and Fereshte Khani. Prompt engineering a prompt engineer. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 355–385, 2024.

[19] Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. Tempera: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract/introduction claim a GA framework that uses thinking models both for evolution and judging, evaluation on a Financial Math Reasoning benchmark with standardized GPT-5-nano judgment, and empirical gains (e.g., GA best exceeds manual in 7/8 models, average $\sim$+4.3%). These claims are supported by the method description and the reported table/figures in Results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper acknowledges judge bias/variance from LLM evaluators, limited dataset scope (Financial Math Reasoning only), uneven population-average improvements, and suggests mitigations (diversity/elitism tuning, CoT operators).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper is empirical/methodological and presents no theorems or formal proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: While GA parameters (population, generations, selection, runs), models, dataset, and judge are specified, key reproduction details are missing (exact meta-prompts/operators, random seeds, API versions, full data split list beyond "500-sample split," and evaluation scripts).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not state that code, prompts, or scripts are released, nor does it provide links/instructions for accessing them.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides enough setup detail to interpret results (dataset domain/size, standardized GPT-5-nano judge, GA hyperparameters, number of runs, compared models), even if not sufficient for full replication.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Although results are averaged over 20 runs, the paper does not report confidence intervals, standard errors/standard deviations, or statistical tests.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The paper does not disclose hardware, API throughput/costs, execution time per run, or total compute.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses non-sensitive benchmark data, involves no human subjects or personally identifiable information, and discusses limitations and potential biases of LLM judges.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
item[] Answer: [No]

Justification: The paper does not include a dedicated broader-impacts discussion (e.g., risks in financial decision-support, fairness, error propagation, or misuse).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No new high-risk models or scraped datasets are released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: While related work is cited, licenses/terms for any datasets/models used are not specified in the paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects or crowdsourcing are involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human-subjects research is conducted.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The core method uses LLMs (thinking models) both as GA operators (initialization, crossover, mutation) and as fitness judges (GPT-5-nano), and this usage is explicitly described throughout the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.