When Designs Explain Themselves: Report Cards for Evolutionary LLMs

Alex Siek [1] **Caishun Chen** [2] **Jian Cheng Wong** [3] **Yew-Soon Ong** [1]

[1]*Nanyang Technological University, Singapore* [2]*Centre for Frontier AI Research, A*STAR, Singapore* [3]*Institute of High Performance Computing, A*STAR, Singapore*. Correspondence to: Alex Siek ASIEK001@e.ntu.edu.sg.

1. Introduction

Large Language Models (LLMs) are increasingly used in evolutionary 3D design optimization. Systems like LLM2TEA [1] generate text prompts for vehicle designs, convert them to 3D meshes via Shap-E [2], and evaluate aerodynamic performance using surrogate models [3]. However, these frameworks guide the optimization using scalar fitness scores. The LLM knows *which* designs performed well, but not *why*, as it lacks the geometric feedback to inform specific successful or flawed features. While LLMs can diagnose mistakes in text and code generation tasks [4], reasoning-enabled evolution remains underexplored for 3D engineering design.

To address this, we introduce a diagnostic mechanism that augments evolutionary LLMs with structured reasoning through *Report Card*. For each generated design, the LLM analyzes depth maps alongside performance metrics to produce a region-level natural-language assessment of strengths and weaknesses. These diagnoses are fed back into the evolutionary loop, allowing the LLM to reason about which geometric properties to preserve or modify. This produces a traceable explanation of change, enabling human designers to interpret and validate the intentions behind each design decision.

2. Method

2.1 Baseline

LLM2TEA [1] operates in four stages: (1) an LLM generates text prompts describing car designs, (2) Shap-E [2] converts each prompt into a 3D mesh, (3) a surrogate model [3] predicts aerodynamic drag (C_d) and lift (C_l), and (4) a tournament-selection evolutionary algorithm selects parents for the next generation. Prompts are evolved via mutation and crossover modes. The LLM receives fitness scores ranked by performance, but no visual or geometric feedback about the designs themselves.

2.2 Report Card Generation

We augment the evolutionary loop with a diagnostic step between evaluation and prompt generation. After each generation is evaluated, every design receives a Report Card: an LLM analysis that takes as input a depth map rendered from the 3D mesh and the design's performance metrics (C_d, C_l). The LLM produces a per-region natural-language assessment (e.g., front, body, rear) identifying geometric strengths and weaknesses. For example: *"Front: well-tapered nose reduces frontal drag. Rear: abrupt cutoff likely causes flow separation and increased pressure drag."*

2.3 Reasoning-Informed Generation

Report cards are incorporated into prompt generation. Rather than generating new prompts based solely on fitness rankings, the LLM reads the diagnostic assessments of parent designs and uses them to inform its mutations. This closes the loop between evaluation and generation with readable feedback: the LLM knows not just *which* designs scored well, but *why*, and can target specific geometric features for improvement.

2.4 Interpretable Evolution Strategies

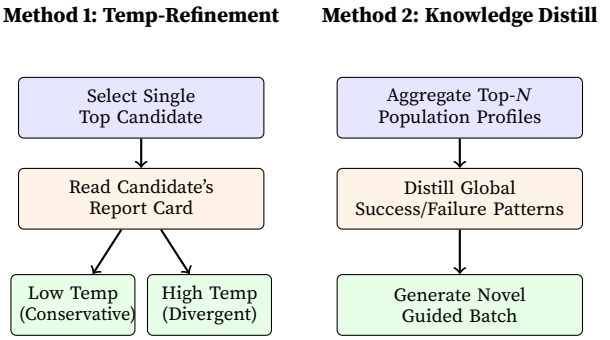

Fig. 1: Interpretable evolutionary strategies. Left: Temperature-Modulated Refinement generates a spectrum of fixes for a single candidate. Right: Historical Knowledge Distillation aggregates population-wide report cards to guide the generation of novel, diverse designs.

Traditional evolutionary algorithms guide mutation using scalar fitness scores. While effective for optimization, this "black box" approach offers little insight to human designers as to *why* a design evolved the way it did. By integrating natural-language Report Cards, we introduce two interpretable evolution strategies tailored to different design workflows.

1. Temperature-Modulated Refinement. This method allows for controlled, interpretable refinement around promising candidates. For a selected top-performing design, the LLM uses its specific Report Card to target diagnosed weaknesses. By generating refined prompts across a temperature gradient (e.g., $T \in \{0.7, 0.85, 1.0\}$), the system produces a spectrum of reasoned modifications. Lower temperatures yield conservative, direct fixes to identified flaws, while higher temperatures introduce divergent geometric interpretations. This provides human designers with a transparent set of alternatives grounded in the same initial diagnosis.

2. Historical Knowledge Distillation. Standard mutation can reduce population diversity by over-

Fig. 2: Traced design lineage across three generations. Left to right: (a) Gen 0 design ($C_d = 0.172$), (b) Gen 0 depth map used for first report card, (c) Gen 1 child addressing rear and wheel weaknesses ($C_d = 0.166$), (d) Gen 1 depth map used for second report card, (e) Gen 2 offspring after further targeted refinement ($C_d = 0.118$, 29% improvement over Gen 1).

exploiting top-performing parents. To preserve diversity while retaining functional improvements, this strategy distills knowledge across the entire population. The algorithm aggregates the Report Cards of the top N designs to identify global success patterns (features that consistently reduce drag) and failure patterns. This aggregated knowledge is then used to prompt the generation of an entirely novel batch of designs. Because the LLM is guided by distilled aerodynamic principles rather than the text of a single parent, it sustains geometric diversity while avoiding known aerodynamic flaws.

3. Qualitative Analysis

Figure 2 traces a lineage from the Top-K Noise strategy. Each generation shows the depth map, report card diagnosis, and resulting design change.

Gen 0. Starting prompt: *"A car in the shape of a flat, low-profile sedan with sloped rooflines"* ($C_d = 0.172$). The report card (Figure 2b) diagnoses: **Strengths** — Roof: *"sloped roofline reduces flow separation"*; Profile: *"low-profile silhouette minimizes frontal area"*. **Weaknesses** — Wheels: *"wheel wells slightly exposed, increasing tyre turbulence"*; Rear: *"abrupt tail angle creates a larger wake region"*.

Gen 1. The LLM directly targets the weaknesses: *"... tapered fastback rear for reduced wake, integrated smooth wheel covers to minimise tyre turbulence, flat undercarriage with a rear diffuser"* ($C_d = 0.166$). Identified strengths are preserved. The Gen 1 report card (Figure 2d) confirms partial improvement but notes the same issues persist at finer scale: Wheels: *"still slightly larger than optimal"*; Rear: *"angle still pronounced"*.

Gen 2. A second targeted refinement: *"... wheel well openings minimised for a tighter fit, integrated spoiler, flat bottom and rear diffuser for optimised undercarriage flow"* ($C_d = 0.118$, **29% improvement** over Gen 1).

The key contribution is **traceability**: a designer can follow the chain of report cards and understand why each design changed. This transforms the evolutionary process from a black box into an auditable sequence of design decisions.

4. Discussion

Depending on the evolution strategy, reasoning-informed methods can achieve comparable or better drag performance than the baseline. Figures A1 and A2 in Appendix A show the minimum drag and DPAR convergence curves across strategies. Both Temperature-Modulated Refinement and Knowledge Distillation maintain competitive performance while adding interpretability at each step.

One observed limitation stems from the text-to-3D generator. Shap-E [2] is sensitive to prompt specificity: as report cards accumulate detail across generations, prompts grow increasingly descriptive, and highly specific prompts can confuse the generator, producing geometries that diverge from the intended design. This is visible in Figure A1, where reasoning-informed strategies require more iterations to converge than the baseline. Addressing this text-to-geometry bottleneck — for instance through more controllable generative models or prompt abstraction — is an important direction for future work.

5. Conclusion

We present a *Report Card* mechanism that introduces structured reasoning into LLM-driven evolutionary design. By analyzing depth maps and performance metrics, the LLM produces per-region diagnostic assessments that explain geometric strengths and weaknesses of each generated design. These assessments are fed back into prompt generation, enabling reasoning-informed evolution where each design change is motivated by an explicit diagnosis. For engineers, this provides a readable trace of design reasoning, allowing them to audit, redirect, or build upon the AI's decisions. As text-to-3D generative models continue to develop, grounding evolutionary operators in visual reasoning offers a path toward more interpretable and controllable AI-driven design.

Acknowledgments

This project was supported by Nanyang Technological University under the URECA Undergraduate Research Programme.

Declaration on the Use of LLM

LLMs were used as general-purpose tools to improve clarity and language during the writing process. All research design, experiments, analysis, and conclusions were conducted and verified by the authors, who take full responsibility for the content.

References

[1] Melvin Wong, Jiao Liu, Thiago Rios, Stefan Menzel, and Yew-Soon Ong. LLM2TEA: An agentic AI designer for discovery with generative evolutionary multitasking. *IEEE Computational Intelligence Magazine*, 20(4):42–55, 2025.

[2] Heewoo Jun and Alex Nichol. Shap-E: Generating conditional 3D implicit functions. *arXiv preprint arXiv:2305.02463*, 2023.

[3] Binyang Song, Chenyang Yuan, Frank Permenter, Nikos Arechiga, and Faez Ahmed. Data-driven car drag prediction with depth and normal renderings. *Journal of Mechanical Design*, 146(5):051714, 2024.

[4] Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

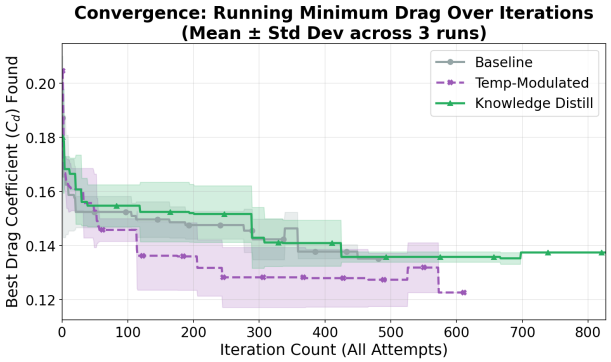

Fig. A1: Minimum drag (C_d) per generation across strategies. Lower is better.

Appendix A. Convergence Plots

Figure A1 shows minimum drag (C_d) convergence across generations for the baseline and reasoning-informed strategies. Figure A2 shows the corresponding Domain and Physical Alignment Rating (DPAR) [1], which penalises designs that achieve low drag through domain-invalid geometries (e.g., flat discs). Reasoning-informed strategies maintain competitive drag performance while producing geometrically valid designs, though they may require more iterations due to the Shap-E sensitivity to highly detailed prompts described in Section 4.

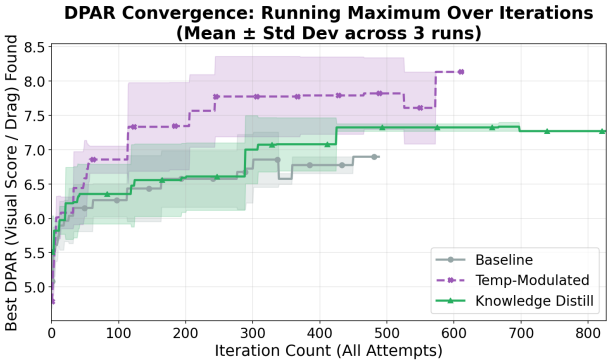

Fig. A2: DPAR convergence across strategies. Higher is better. DPAR balances domain validity with aerodynamic performance.