

Mercury Model Birthmarks: Structural Signatures of Token Cost, Latency, and Hidden Defects Across LLM Scales

Chen, Ho Yiing (norika)¹

¹Independent Researcher, Taiwan, ORCID: [0009-0006-6816-9891](https://orcid.org/0009-0006-6816-9891), norika at charenix.com

Draft: June 18, 2026

Abstract

Parameter count is a poor health report. A 3B, 7B, 14B, 32B, 70B, or larger language model may be fast, slow, concise, verbose, stable, fragile, or strangely wasteful for reasons that are not explained by size alone. This paper proposes the Mercury model birthmark: a stable structural signature that appears across prompts or tasks and reveals how a model was tokenized, trained, aligned, compressed, or architecturally constrained.

The birthmark is not a metaphor for mystery. It is a measurement object. I define token birthmarks, latency birthmarks, routing scars, dead-tissue regions, alignment reflexes, and compression scars using observable quantities: tokenizer burden, internal activation mass, route dispersion, repeated abnormal sites, and quality-adjusted cost. The central claim is severe: every model has marks. The question is where they are, whether they are harmless, and whether a prompt can expose them.

Author Note

The model-size story is too clean. People say 7B, 14B, 70B, 200B as if those numbers explain the creature. They do not. Parameter count is height and weight. Mercury is closer to an internal exam.

If humans built the model, there will be marks somewhere. The clean question is whether we can measure them without turning the idea into poetry.

No perfect body exists; the serious question is where the mark sits and whether it changes the answer.

1 Parameter Count Is Not Health

Let m be a model and x a prompt. Size gives us P_m , the parameter count. It does not directly give token cost, latency, internal route stability, or defect exposure. A larger model can answer faster in a suitable environment. A smaller model can waste time through poor tokenization, scattered routing, or fragile decoding.

The question is not whether large models are better. The question is what kind of internal cost each model pays to answer the same problem.

Size is only the first line on the chart; the birthmark is what the chart usually hides.

2 Birthmark Definition

Let $Z_m(x)$ be a vector of observable Mercury measurements:

$$Z_m(x) = [T_m(x), R_m(x), I_m(x), D_m(x), Q_m(x), U_m(x)],$$

where T is output token count, R is latency, I is internal activation mass, D is route dispersion, Q is quality, and U is unused or low-contribution region exposure.

A model birthmark is a stable abnormal signature:

$$B_m = \left\{ s : \Pr_{x \sim X} [\text{abn}_m(s, x) > \lambda] > \pi \right\}.$$

Here s may be a tokenizer pattern, layer, module, dimension, head group, or route family. Stability matters. One strange response is noise; repeated abnormality is a mark.

This is why the term should be earned. A birthmark needs recurrence across a prompt family, an identifiable internal address, and a measurable effect on cost, latency, or quality.

3 Token Birthmark

Tokenizer burden is

$$\text{TB}_m(x) = \frac{b_m(x)}{\text{chars}(x) + \epsilon}.$$

A token birthmark exists when a language, script, symbol class, or format consistently creates excess segmentation:

$$B_m^{\text{tok}} = \{x : \text{TB}_m(x) > \mu_{\text{TB}} + \sigma_{\text{TB}}\lambda\}.$$

Chinese, Japanese, mixed scripts, addresses, phone numbers, and legal or receipt formats are natural stress tests. A model can look slower because it was forced to chew the input into too many pieces.

4 Latency Birthmark

Let $R_m(x)$ be latency in a suitable environment for m . Define quality-adjusted latency:

$$\text{QL}_m(x) = \frac{R_m(x)}{Q_m(x) + \epsilon}.$$

A latency birthmark is not simply being slow. It is being slow in a repeated, task-specific way after environment and quality are accounted for:

$$B_m^{\text{lat}} = \{x : \text{QL}_m(x) > \lambda_R\}.$$

This separates a hardware mismatch from a model-side habit.

5 Routing Scar

Let $D_m(x)$ be route dispersion and $I_m(x)$ internal activation mass. A routing scar is a stable pattern where the model lights too much or scatters too widely for the quality it returns:

$$\text{Scar}_m(x) = \frac{I_m(x)D_m(x)}{Q_m(x) + \epsilon}.$$

The scar set is

$$B_m^{\text{route}} = \{s : \mathbb{E}_{x \in X_s} [\text{Scar}_m(x)] > \lambda_S\}.$$

This is the place where Mercury can be more than a benchmark. The benchmark says the model was slow or wrong. Mercury asks which internal streets were overused.

6 Dead Tissue

Some regions may remain consistently low contribution across prompts:

$$U_m(s) = \mathbb{E}_{x \sim X} [\mathbf{1}[a_s(x) < \tau]].$$

Dead tissue is not automatically safe to remove, but it is suspicious:

$$B_m^{\text{dead}} = \{s : U_m(s) > \lambda_U \wedge \text{risk}(s) < \rho\}.$$

This is where model surgery begins. A low-activity region must still be tested with ablation, PPL, task probes, and prompt stress.

The mark is a candidate, not a verdict.

7 Alignment Reflex

Instruction tuning and safety alignment can leave reflexes. Let $G_m(x)$ be the fraction of output spent on guard text, disclaimers, hedging, or irrelevant preface under a task where direct answer is permitted:

$$G_m(x) = \frac{\text{guard_tokens}_m(x)}{T_m(x) + \epsilon}.$$

An alignment reflex birthmark is

$$B_m^{\text{align}} = \{x : G_m(x) > \lambda_G \wedge Q_m(x) \text{ does not improve}\}.$$

This is one reason a model may use more tokens without doing more useful work.

In product terms, this is politeness turned into latency and battery drain.

8 Compression Scar

Quantization, pruning, distillation, or surgery can expose scars. Compare an original model m and a compressed model m' :

$$\Delta Z(x) = Z_{m'}(x) - Z_m(x).$$

The compression scar is

$$B_{m'}^{\text{comp}} = \{s : |\Delta I_s(x)| > \lambda_I \vee |\Delta Q(x)| > \lambda_Q\}.$$

This equation is useful because compression failures are often local before they become obvious in global metrics.

The same logic applies to model surgery. If a layer or dimension looks unused, Mercury does not declare victory; it asks whether removing or perturbing that site changes the route, the token count, or the task score.

9 Birthmark Vector

The model-level signature is

$$\mathcal{B}(m) = (B_m^{\text{tok}}, B_m^{\text{lat}}, B_m^{\text{route}}, B_m^{\text{dead}}, B_m^{\text{align}}, B_m^{\text{comp}}).$$

This vector is the health report. It does not replace benchmark scores. It explains why two models with similar scores can feel completely different in local use.

10 Relation To Existing Work

Scaling laws describe predictable relationships between model size, data, compute, and loss [1, 2]. Model cards ask authors to report intended use, evaluation, and limits [3]. Quantization and compression studies show that lower precision can preserve much while introducing new error modes [4]. Mechanistic interpretability gives language for components, circuits, and causal hypotheses [5].

Mercury birthmarks sit between those traditions. They are not only evaluation scores, and they are not yet full causal explanations. They are stable internal signatures that tell us where a model carries cost, fragility, or residue.

11 Limits

The word birthmark must not be abused. A mark is stable evidence, not a name we attach to anything ugly. Hardware, prompt sets, quality scoring, and compression baselines must be reported together, because a hidden environment mismatch can look exactly like a model-side scar.

The discipline matters because the idea is powerful. If used loosely, it becomes gossip. If measured carefully, it becomes a map of model health.

12 Conclusion

Mercury Model Birthmarks argues that every model carries structural marks. Some marks come from tokenization. Some come from latency habits, scattered routes, dead tissue, alignment reflexes, or compression wounds. Parameter count tells us how large the body is; Mercury asks how that body moves, where it wastes effort, where it hides defects, and which prompt will finally expose the mark.

References

- [1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. arXiv:2001.08361, 2020. <https://arxiv.org/abs/2001.08361>
- [2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre: Training Compute-Optimal Large Language Models, compute-optimal scaling analysis for large language model training. arXiv:2203.15556, 2022. <https://arxiv.org/abs/2203.15556>
- [3] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019. <https://arxiv.org/abs/1810.03993>
- [4] Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. arXiv:2210.17323, 2022. <https://arxiv.org/abs/2210.17323>
- [5] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits. *Transformer Circuits*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>