**Supplementary Figure 1: NSM Maps for All Experiments of Llama2-7b and Zephyr**

Details non-silent mutation (NSM) distribution across all experiments for Llama2-7b(A) and Zephyr(B). The significantly and consistently sensitive column 39 in all Llama2-7b V matrices, except in layer 1 of matrix V, were highlighted with red arrows in Supplementary Figure 1B. Similarly, the significantly and consistently sensitive column 32 in all Zephyr V matrices were highlighted with red arrows in Supplementary Figure 1B.For ease of visualization, the images of all Up and Gates matrices are transposed in this and following figures.

**Supplementary Figure 2: COPA of Different Experiments of Llama2-7b and Zephyr**

Displays Correlative Complementary Patterns (COPA) for experiments including Java, Python0, Python10 for Llama2-7b(A), and Python0, Drosophila for Zephyr(B). The same phenomena were also observed in the other experiments.

**Supplementary Figure 3: Distribution of Destructive NSMs and Mmlu21 Scores for Llama2-7b and Zephyr**

The standard model provides correct answers to the first 14 questions and incorrect answers to the remaining 7 questions. Models with some NSMs can have more 14 correct answers.

**Supplementary Figure 4: Cosine Similarity Distribution Across Different Experiments of Llama2-7b and Zephyr**

Explores cosine similarity for Llama2-7b experiments in Drosophila, Newton and for Zephyr in Drosophila, Newton, P53. Excludes images without NSMs.

**Supplementary Figure 5: Malignant NSMs with Cosine Similarity ≤0.5 and Mmlu21 Scores ≤8**

Excludes images without NSMs. A significant part of NSMs are shared among different descriptive experiments of the same model.

**Supplementary Figure 6: Maps of NSM with RIHF for Llama2-7b and Zephyr**

Visualizes NSM distribution with Rare Initial Words of Highest Frequency (RIHF) across both models (See Method). Excludes images without NSMs.

**Supplementary Table 1: General Statistics of NSMs for Llama2-7b and Zephyr**

* Number of maximum, minimum and zero NSM of different experiments.
* Total number of NSMs of different matrices and layer across different experiments of Llama2-7b.
* Total number of NSMs of different matrices and layer across different experiments of Zephyr.
* Total number of distinct phenotypes of different matrices and layer across different experiments of Llama2-7b.
* Total number of distinct phenotypes of different matrices and layer across different experiments of Zephyr.
* Number of 64x64 squares in each different matrix type of Llama2-7b and Zephyr.

**Supplementary Table 2: NSMs of Mmlu21 experiment with Mmlu21 score <= 10 for Llama2-7b and Zephyr**

**Columns:**

* **Matrix:** Layer and matrix type
* **NSM Type:** Indicates whether the mutation is maximum, minimum, or zero.
* **NSM loc1:** The Y-coordinate of the mutation within the matrix, divided by 64 for alignment with Table 1.
* **NSM loc2:** The X-coordinate of the mutation within the matrix, divided by 64 for alignment with Table 1.
* **Mmlu21 Score:** The score obtained in the Mmlu21 experiment, with values ≤ 10.
* **Destructive NSM:** Indicates whether the mutation was considered destructive in the Mmlu21 experiment.
* **Cosine Similarity in Drosophila Experiment:** The cosine similarity measure between the mutated model's output and the standard model's output during the Drosophila experiment. Specifies "standard" if the MMLU21 NSM acts as a silent mutation in the Drosophila experiment and matches the output of the standard model.

**Supplementary Table 3: Mutations in different matrices switch Zephyr output to poetry and dialogue**

The second column contains the following information of mutations: mutation type; layer and matrix where the mutation occurs; X and Y coordinates within the matrix; RIHF, Cosine Similarity between mutation output and output of standard model of Drosophila experiment; Mmlu21 score of the 14/7 questions that standard model provides right/wrong answers. The score is noted as standard when the mutation output matches the standard model.

**Mutations from up to down in the table:** RIHF=”starring”: one mutation in layer 1; one mutation in layer 14, two mutations in layer 30. RIHF=”otta”: one mutation in layer 1. RIHF=”tugging”: one mutation in layer 14. The last mutation is a dysfunctional mutation in layer 1 producing chaotic output with RIHF ”starring” and has Mmlu21 score 4/4. All listed mutations are from MLP Down matrices.

**First Row Information:** Contains the inputs used for the experiment.

**Color Coding for Outputs:** Green: Poem; Yellow: Dialogue; Blue: First-person narrative blue; White: Other types of outputs.

**Supplementary Document 1: List of 151 Inputs Used in RIHF Calculation**

This document catalogs all 151 inputs used for calculating Rare Initial Words of Highest Frequency (RIHF). These inputs, generated by ChatGPT, are arranged linearly, with each line representing one input. The first 101 inputs feature more literary expressions, while the final 50 are factual statements.

**Supplementary Document 2A:** **151 Outputs for Zephyr**

This section consists of 151 blocks separated by "%%%%%%". Each block begins with one input from Supplementary Document 1 followed by its corresponding output from the standard Zephyr. This format provides a clear before-and-after snapshot of how the standard model processes each input, serving as a baseline for comparison.

**Supplementary Document 2B: 151 Outputs for Zephyr with “writer” mutations**

Includes data from 12 of "writer" mutations identified in our study:

* All 10 NSMs with the RIHF "starring" as highlighted in Figure 6D with red arrows, with or without blue outline.
* Two specific NSMs from Table 1 with RIHF "otta:" and "tugging" These mutations are highlighted in Figure 6D by light blue arrow with blue outline and yellow arrow with blue outline. The document covers all NSMs listed in Table 1.

For each NSM, the document presents:

* Mutation information (e.g., "1.down (0,38) minimum" indicates a minimum mutation in the Down matrix of layer 1 at coordinates (0,38)).
* 151 blocks separated by "%%%%%%", where each block contains one input from Supplementary Document 1 followed by the output from Zephyr carrying this specific NSM.

Additional Formatting Note: Non-ASCII characters within the document are denoted using the format "ord(xxx)", where "xxx" represents the Unicode code point of the character, ensuring clarity and precision in representing textual data.