The Single-File Test: Frontier LLMs in the Wild

A Longitudinal Public-Interface Evaluation of First-Output Web Generation & Social Reach

DATASET: 12/10/2025 - 02/04/2026

SCOPE: 17 Experiments | 68 Runs

MODELS: GPT, Gemini, Grok, Claude

THE PREMISE & THE PROTOCOL

This research studies how frontier LLMs generate complete web applications when tested through public web UIs, mirroring natural, non-technical end-user interaction.

CRITICAL CONSTRAINTS

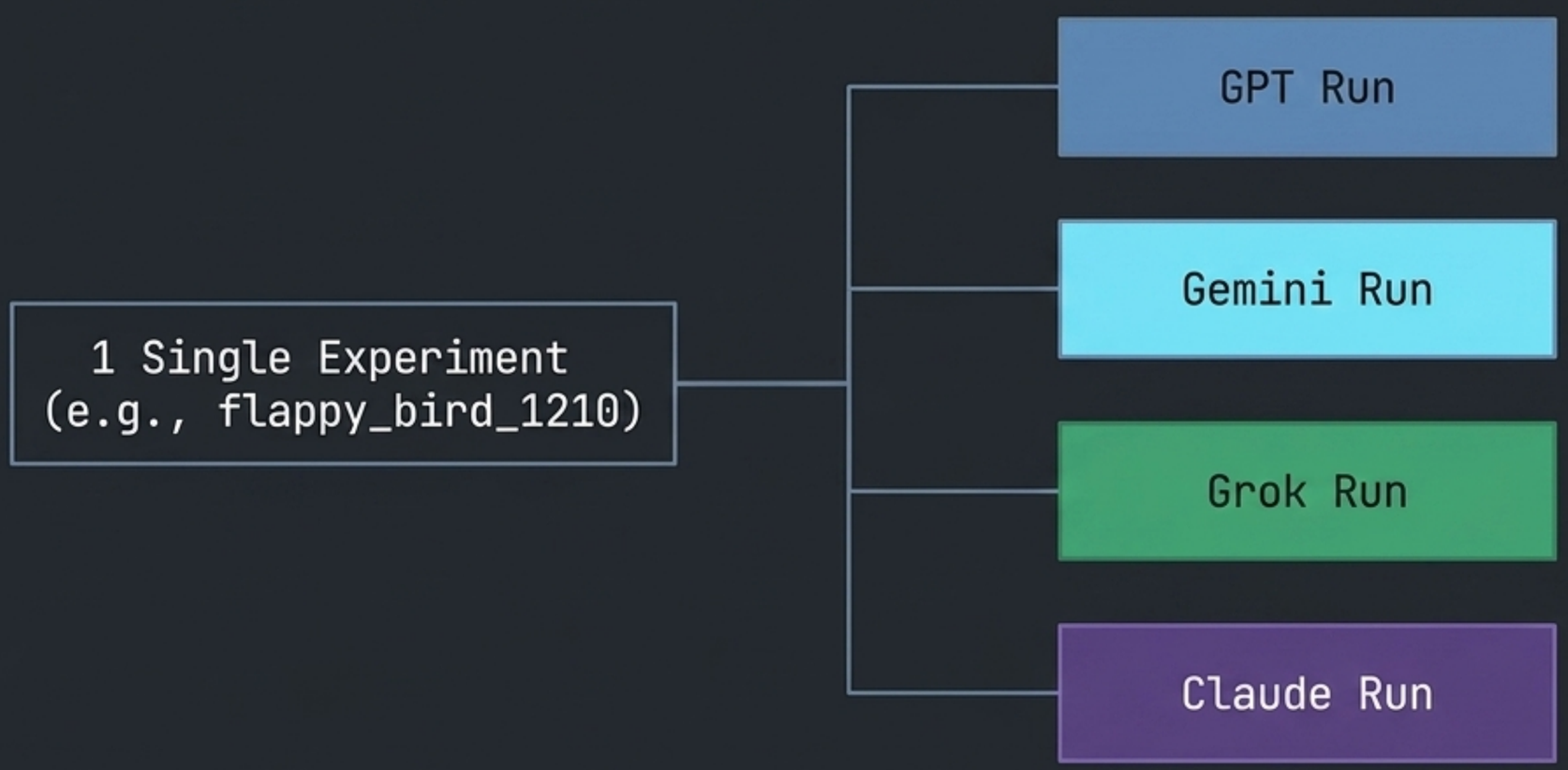
LAB SETTINGS ✓ THE WILD

- **OBSERVATIONAL, NOT CAUSAL:** A comparison of public interfaces under a fixed user protocol, not a clean benchmark of intrinsic capability.
- **DYNAMIC ENVIRONMENTS:** Interface versions changed over time, Claude was accessed via arena.ai, and timing conditions fluctuated.

The First-Output Protocol: Each model received the same prompt once. Zero retries. debugging. Zero repair prompts. Zero iterative steering.

Prompt → Public Chat UI → First HTML Output → Browser Execution → Dual Scoring (Human + Gemini) → Social Posting → 24-Hour Telemetry Collection

DATA ARCHITECTURE: N=17 TO N=68



17 Experiments × 4 Models = 68 Model-Level Rows

DIAGNOSTIC READOUT

```
MISSING_VALUES: 0
DUPLICATE_ROWS: 0
VARIABLES_TRACKED: 48
```

(Spanning Model Metadata, Timing, Reasoning Stats, Quality Scores, Audio Specs, and Social Engagement)

The Evaluator Matrix (Browser Video Scoring)

Human Evaluator

Scale:
0-10 (Decimals allowed)

Blind Status:
Partially blinded to model identity

Role:
Primary author judgment on Prompt Adherence, Functional Correctness, and UI quality based on rendered browser video.

Gemini Evaluator

Scale:
1-10 (Integers only)

Blind Status:
Blind-chat judgment

Role:
Secondary judge (capable of video input). Used strictly for directional agreement and leniency checks, not as a perfectly controlled reliability test.

TARGET ACQUISITION: PREDICTING CODE VERBOSITY

Input Word Count (Prompt length)

TYPE: NUMERIC | RANGE: 20-512
UNIT: TOKENS

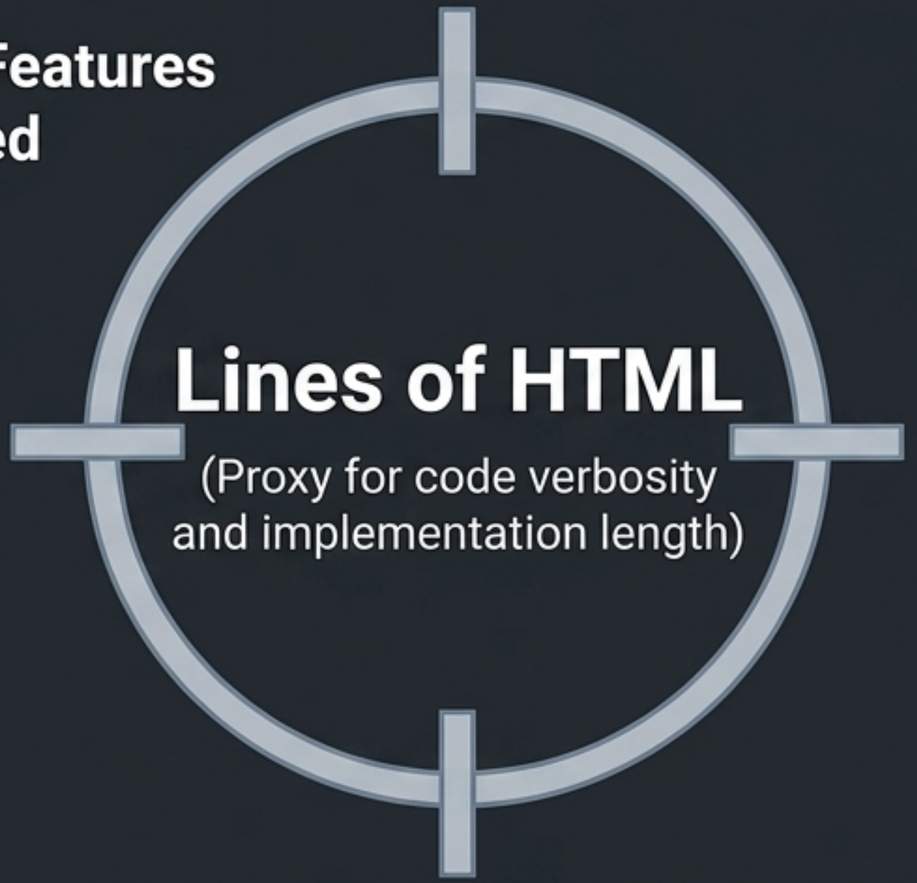
Model Family (Categorical identifier)

TYPE: CATEGORICAL | CLASSES: 4
VALUES: GPT, Gemini, Grok, Claude

Prompt Text (TF-IDF encoded)

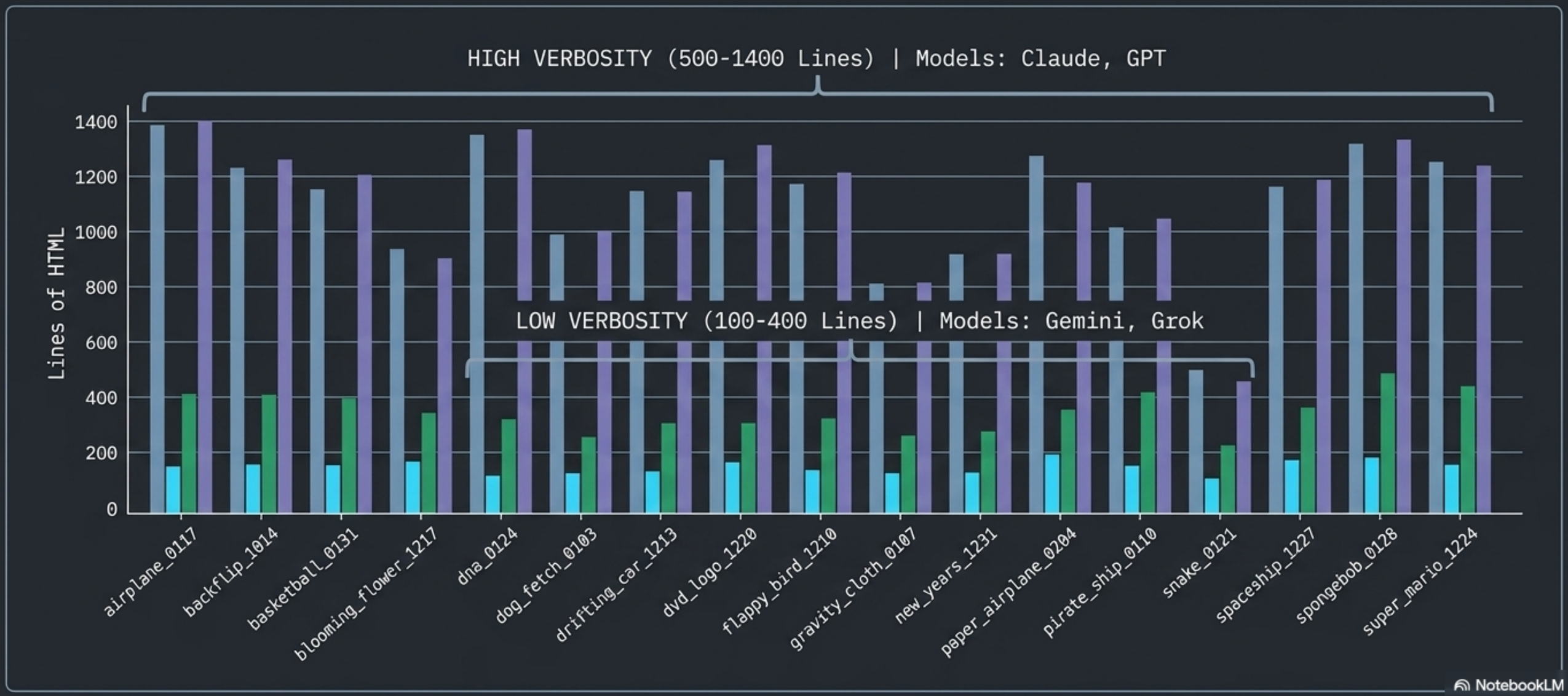
TYPE: SPARSE VECTOR | DIMS: HIGH
FEATURE: TERM FREQUENCY

Candidate Features Tested



DATA LEVEL: N=68 (Model-Level Generations) | METHODOLOGY: Supervised Ridge Regression (L2 regularization) via Leave-One-Experiment-Out grouped validation.

The Verbosity Spectrum



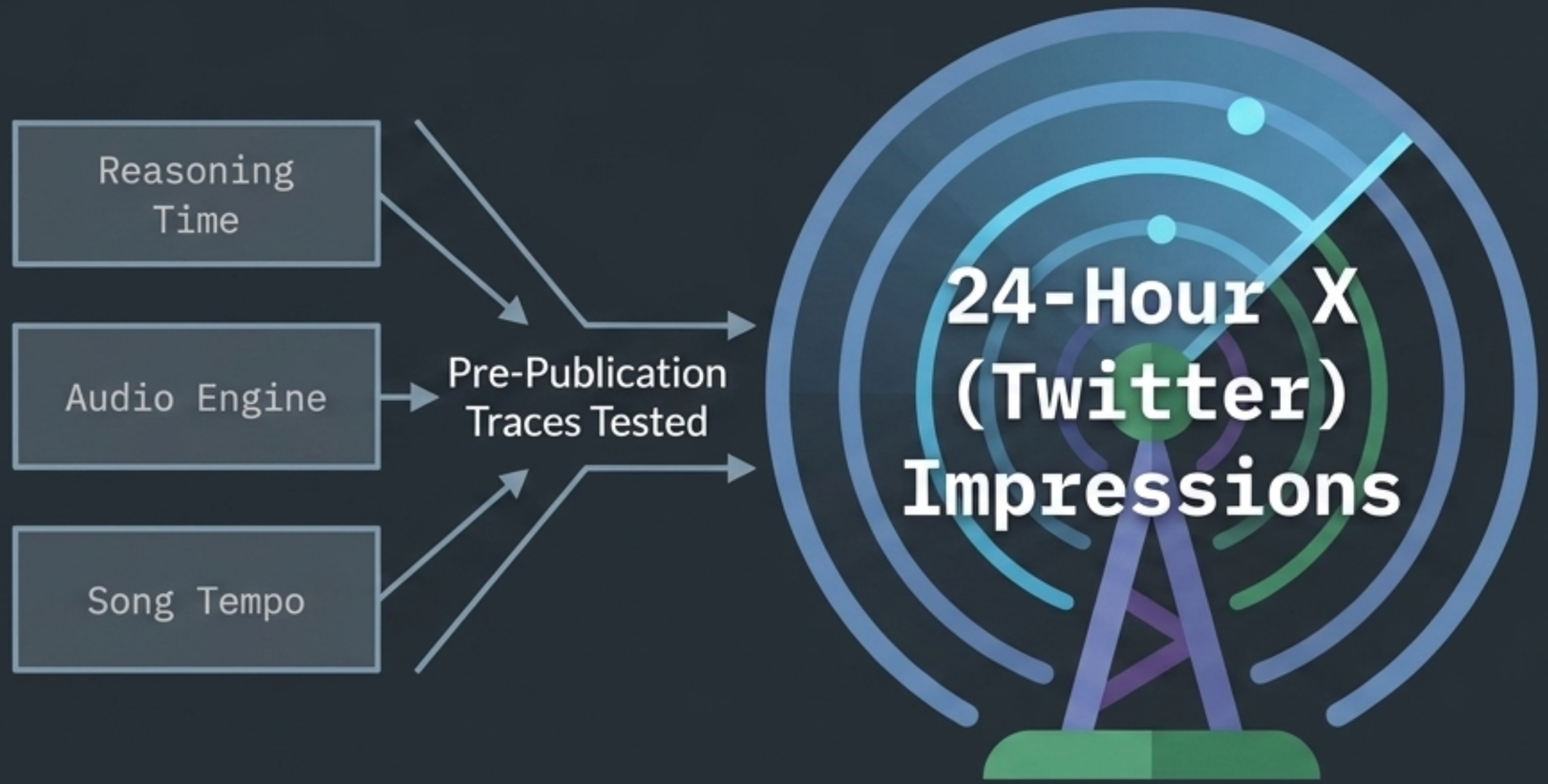
Mission Success

Model Identity Trumps Prompt Length

The model-family-only baseline performed best. Adding structured prompt length (Input Word Count) or TF-IDF prompt text to the **Ridge Regression** did not improve performance under grouped validation. The length of generated HTML is primarily driven by **WHO** is writing it, not **WHAT** was asked.

MAE	RMSE	R^2
Evaluated	Evaluated	Evaluated

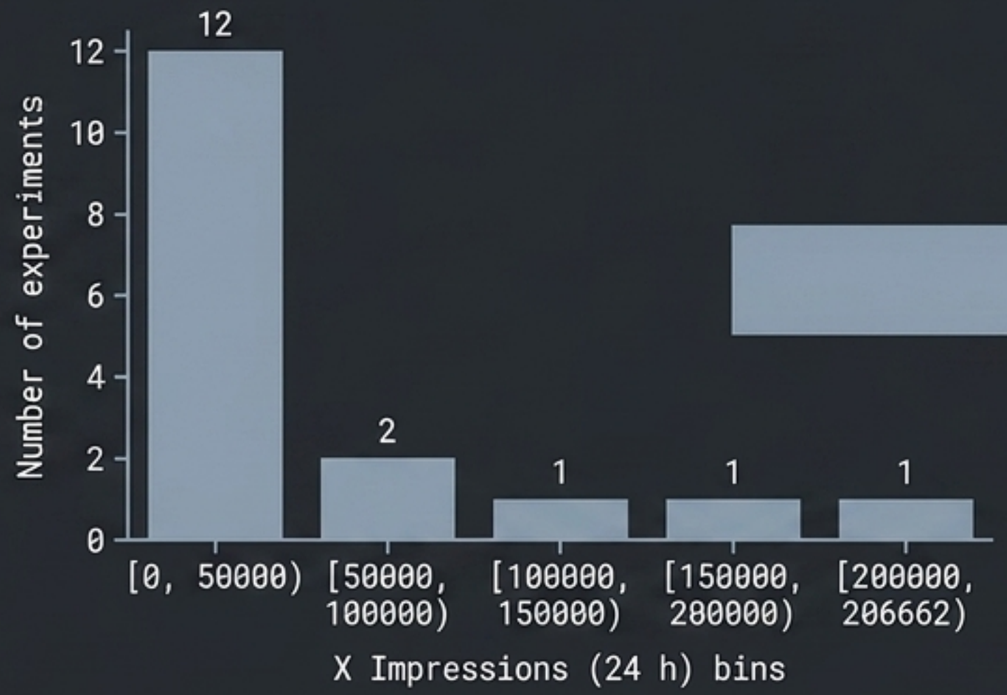
Target Acquisition: Social Reach



DATA LEVEL: N=17 (Collapsed Experiment Level)
CAVEAT: Exploratory and statistically underpowered.
CORE QUESTION: Can pre-publication technical generation variables predict post-publication social reach?

THE HEAVY-TAIL SKEW

Original Distribution: X Impressions (24 h) by Experiment



Transformed Distribution: Normal Distribution

$$\log(1 + X\ \text{Impressions})$$

This log-scale regression was strictly necessary to stabilize the heavy-tailed distribution caused by a few massive reach outliers.

Signal Detection

Full-Sample Lasso Screen (Narrowing) → Final Ridge Model → Leave-One-Out Cross-Validation (LOOCV)

Final 4 Features Evaluated

- ReasonTime_mean
- Reasoning_Ratio_mean
- Suno_version (v5 vs v4.5)
- Song_BPM

Weak Predictive Signal.

The resulting error estimates are unstable descriptive evidence. Technical generation traces and AI audio specs do not provide a reliable, generalizable forecast for social virality in this dataset.

The Predictive Engines Matrix

The Verbosity Model	The Virality Model
Target: Lines of HTML	Target: Log(1 + X Impressions)
Scale: N=68 (Model Level)	Scale: N=17 (Experiment Level)
Methodology: Ridge (Leave-One-Experiment-Out)	Methodology: Lasso Screen -> Ridge (LOOCV)
Conclusion: **STRONG SIGNAL.** Output length is dictated by model identity, not not the prompt.	**Conclusion:** **WEAK SIGNAL.** Pre-publication telemetry fails to reliably predict post-publication reach.

The First-Output Reality

Under strict zero-shot constraints via public UIs, model behaviors diverge wildly. “The Wild” breaks the parity seen in sterile benchmarks.

The Identity Fingerprint

If you want verbose, complex architectures, Claude and GPT naturally default to to it. Gemini and and Grok bias heavily toward terseness, regardless of prompt length.

The Telemetry Disconnect

Backend AI friction (Reasoning time, thinking ratios) does not correlate with frontend human engagement. Viral reach remains a social phenomenon, not a mathematically predictable output of telemetry.