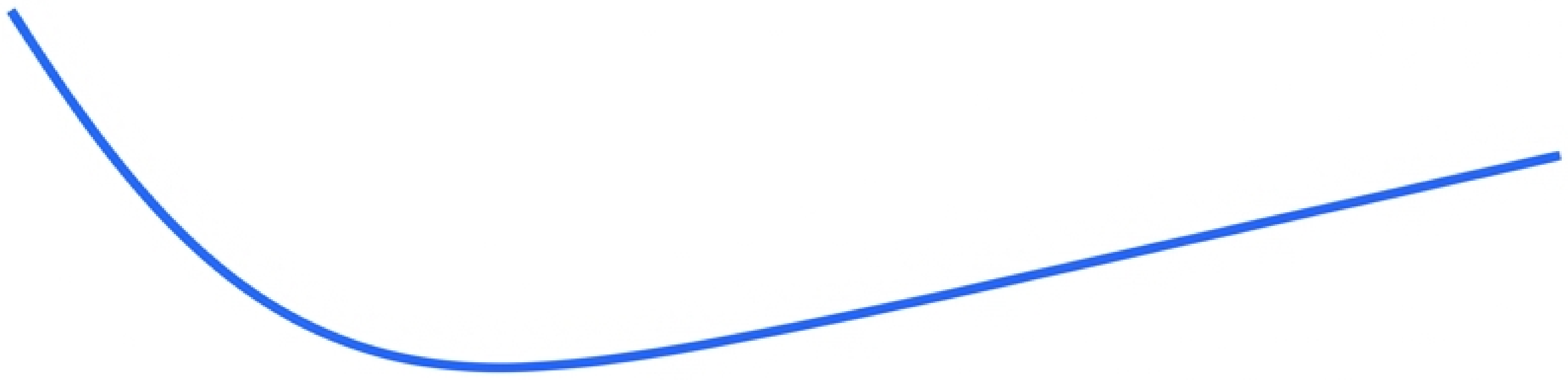
Characterizing Data Consumption *E*(*S*) in the Intermediate Regime



A rigorous evaluation of closed-form expressions for the WSD Stable Phase.

WSD Schedules | Scaling Laws | Optimization

Executive Summary

We found a principled replacement for the ad-hoc quadratic approximation.

The 'Intermediate Regime' of WSD training (where S is between minimum steps and infinity) currently lacks a derived formula for data consumption. This forces engineers to rely on messy piecewise approximations.

After evaluating six candidate functions against asymptotic constraints and noise, the Hyperbolic Blend emerges as the optimal model.

Recommendation: Adopt the Hyperbolic form:

$$E(S) = aS + \frac{b * S_{min}}{S - S_{min}} + c$$

0.9986

R-Squared Accuracy

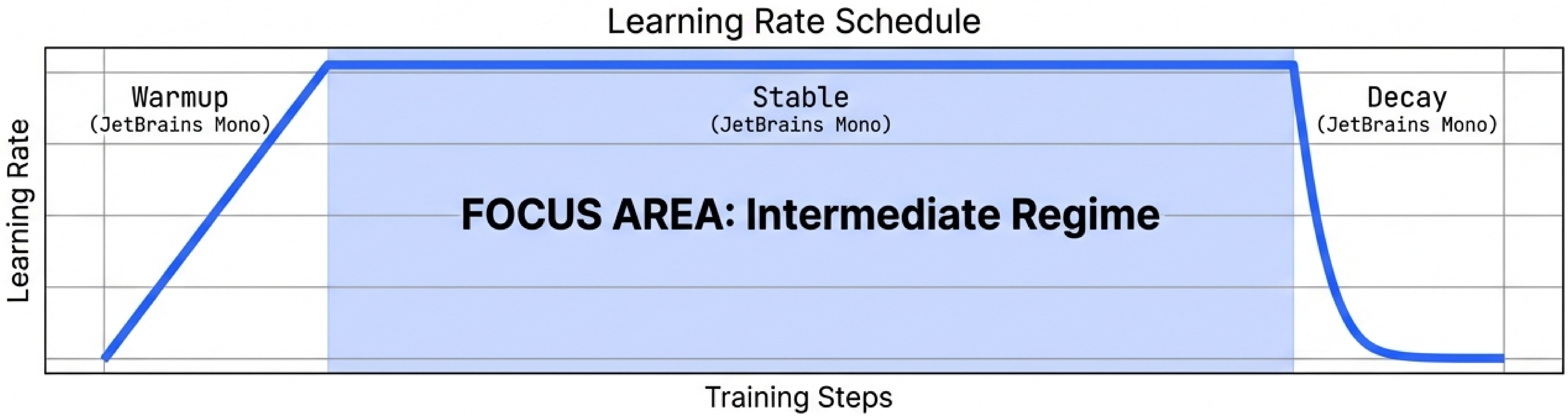
4968

BIC Score - Lowest Complexity

20%

Noise Robustness

Classical Critical Batch Size relationships break down in the WSD Stable Phase.



Context: The Warmup-Stable-Decay (WSD) schedule is standard for modern LLM pre-training.

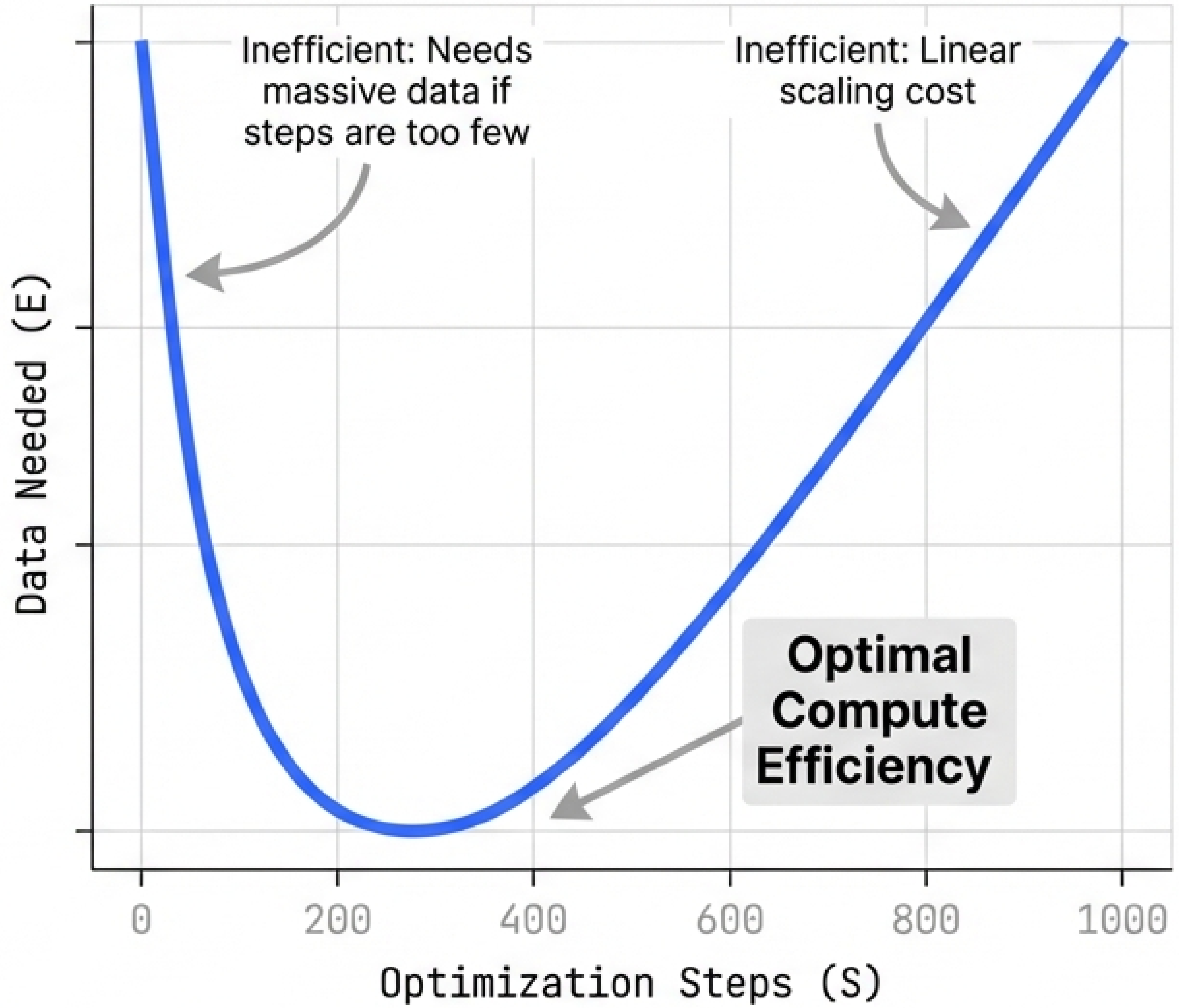
Problem: Zhou et al. [5] established that while we understand the edges of the curve (Warmup and Decay), the classical scaling relationships do not hold during the Stable phase.

The Data Consumption Function E(S) dictates training efficiency.

E(S) = Total tokens required to reach a target loss.

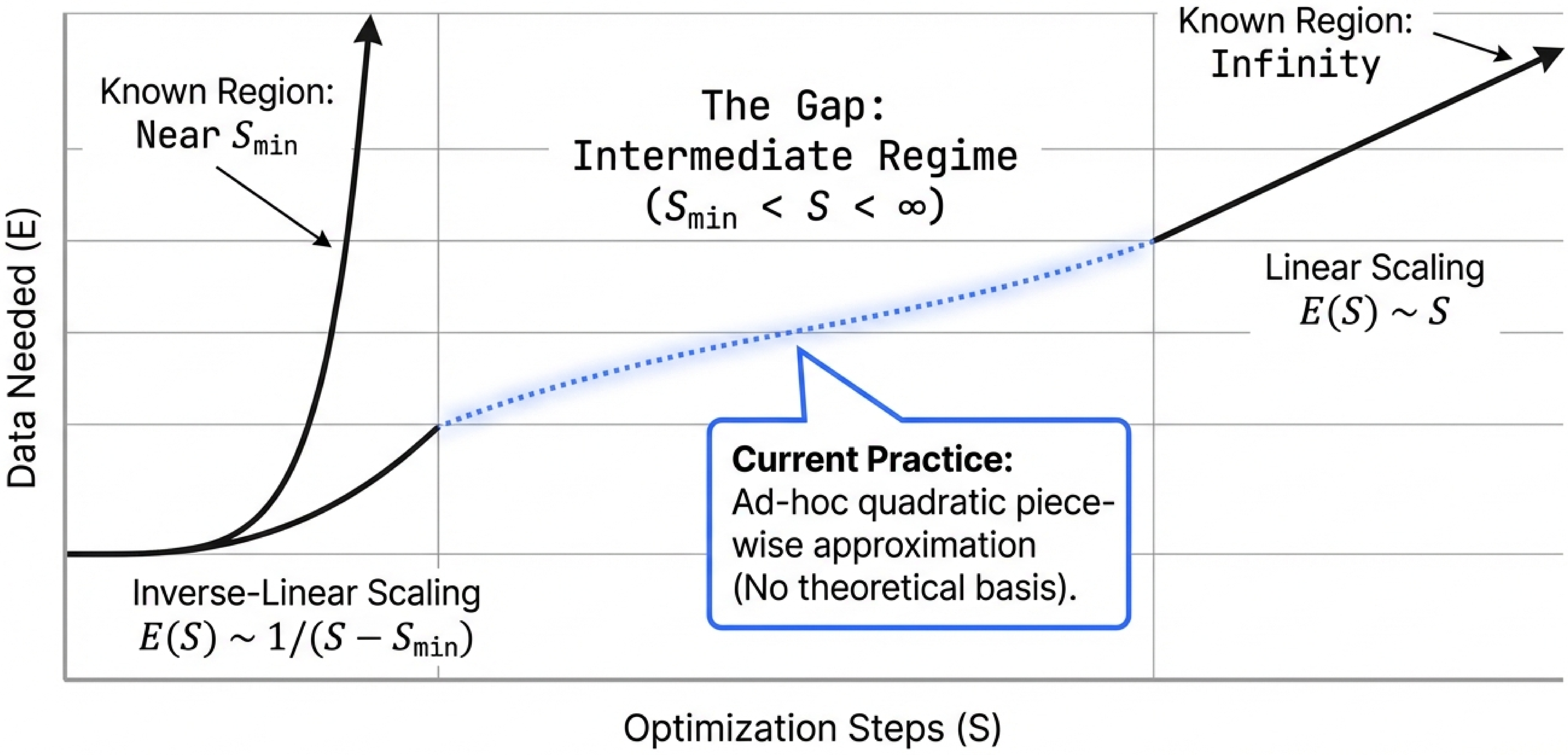
Constraint: Given fixed optimization steps **S**.

The Engineering Question: “How much data do I need to process if I am constrained to **S** steps?”

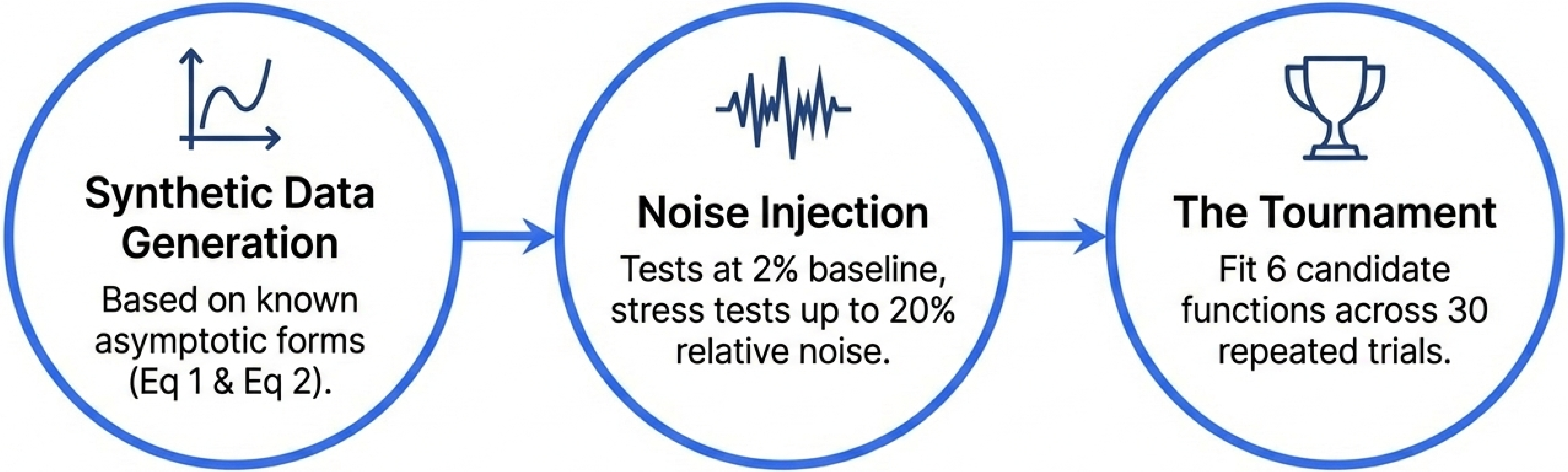


Source: Based on scaling laws by Kaplan et al. [3].

The "Intermediate Regime" has remained mathematically uncharacterized



Evaluating candidates through synthetic generation and stress testing.



Evaluation Metrics: **R-Squared, RMSE, MAPE, BIC (Bayesian Info Criterion), AIC.**

Six closed-form candidates were evaluated.

1. Quadratic

$$E = a(S - S_{min})^2 + b(S - S_{min}) + \frac{c}{S - S_{min}}$$

4. Logistic Blend

$$E = \text{sigma}(\ldots) * aS + (1 - \text{sigma})(\ldots)$$

2. Rational

$$E = \frac{aS^2 + bS + c}{S - S_{min} + d}$$

5. Power-Rational

$$E = aS^p + \frac{b * S_{min}{}^p}{(S - S_{min})^p}$$

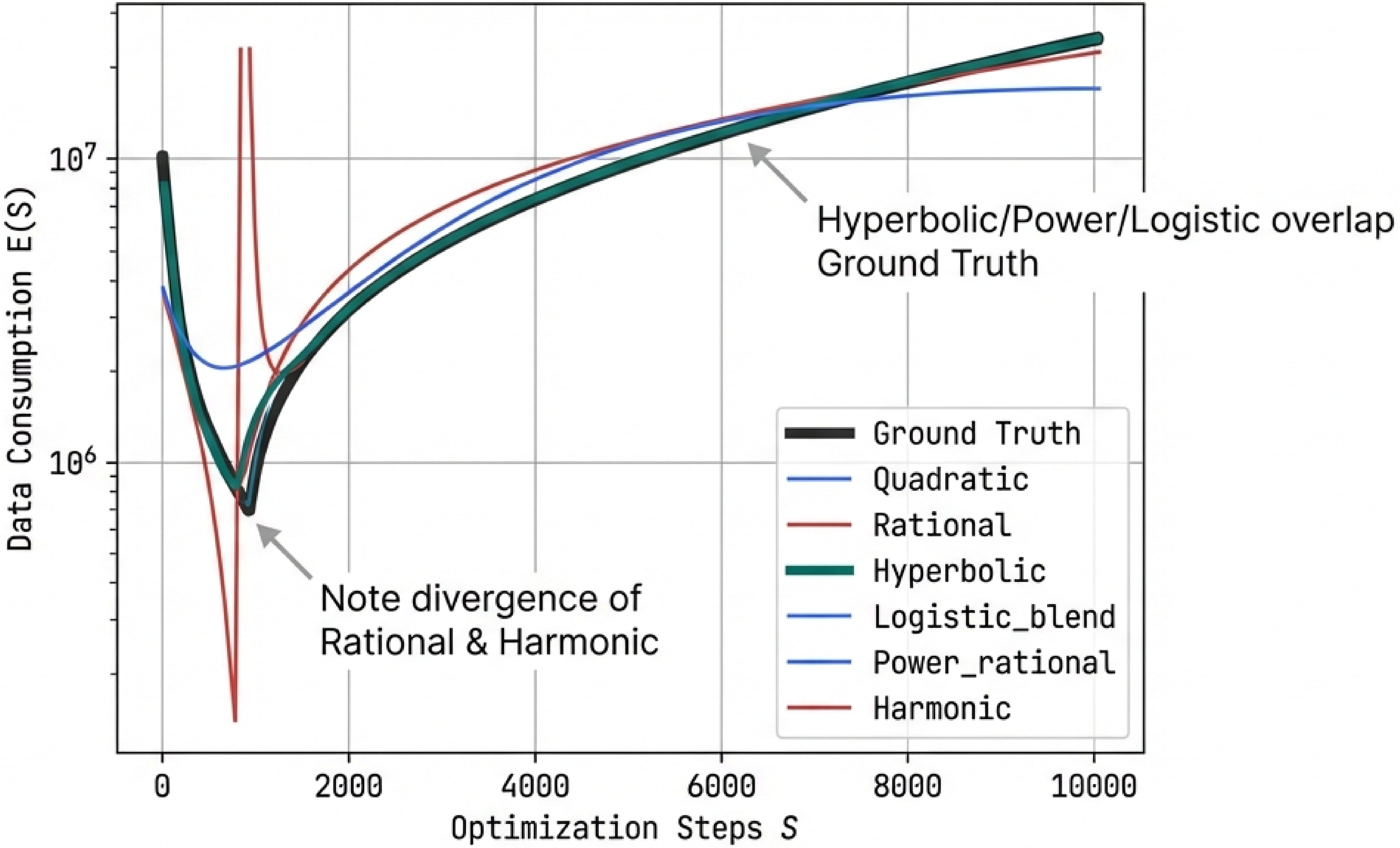
3. Hyperbolic (Protagonist)

$$E = aS + \frac{b * S_{min}}{S - S_{min}} + c$$

6. Harmonic

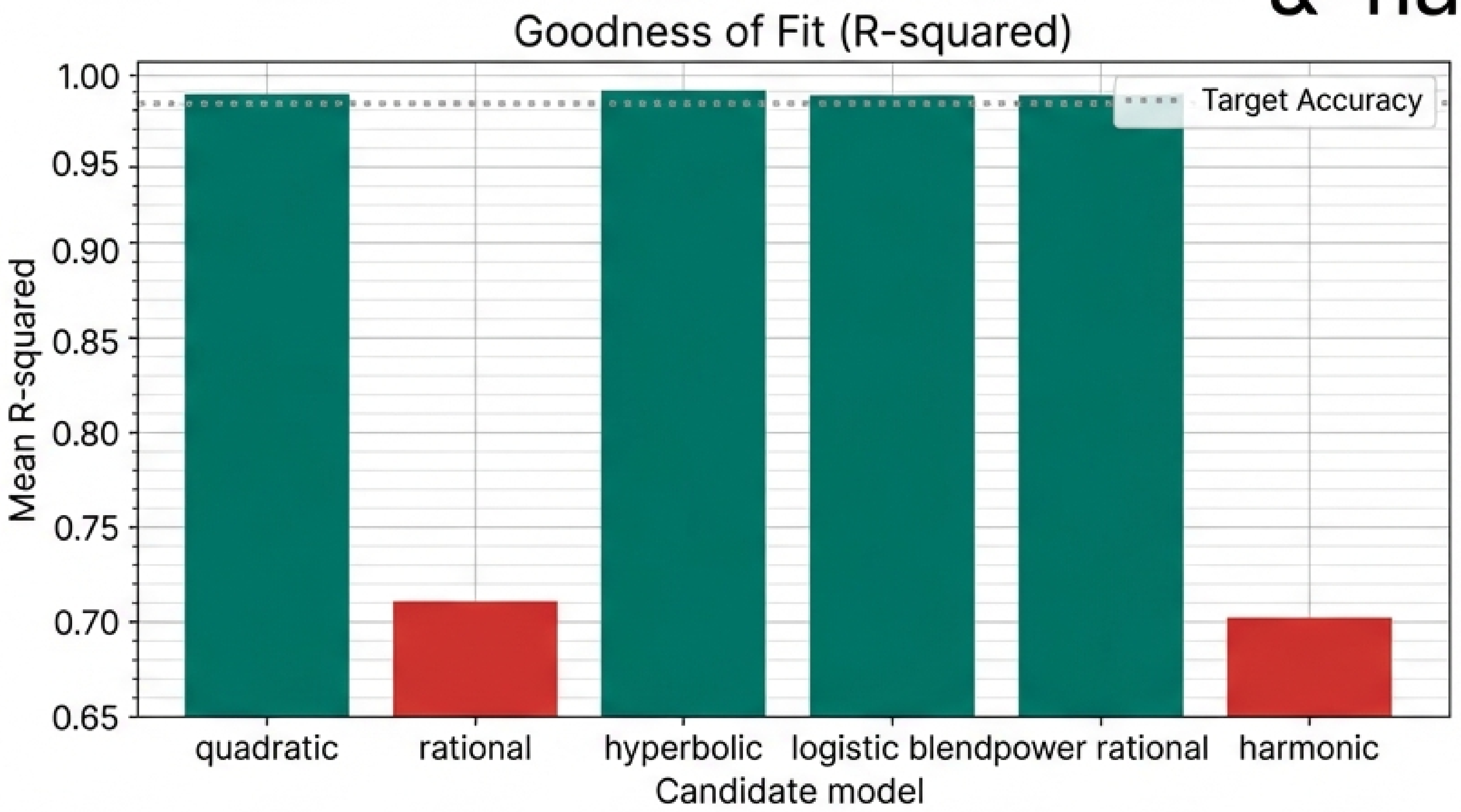
$$E = \frac{1}{1/aS + (S-S_{min})/b} + cS$$

Visualizing the fit against Ground Truth.

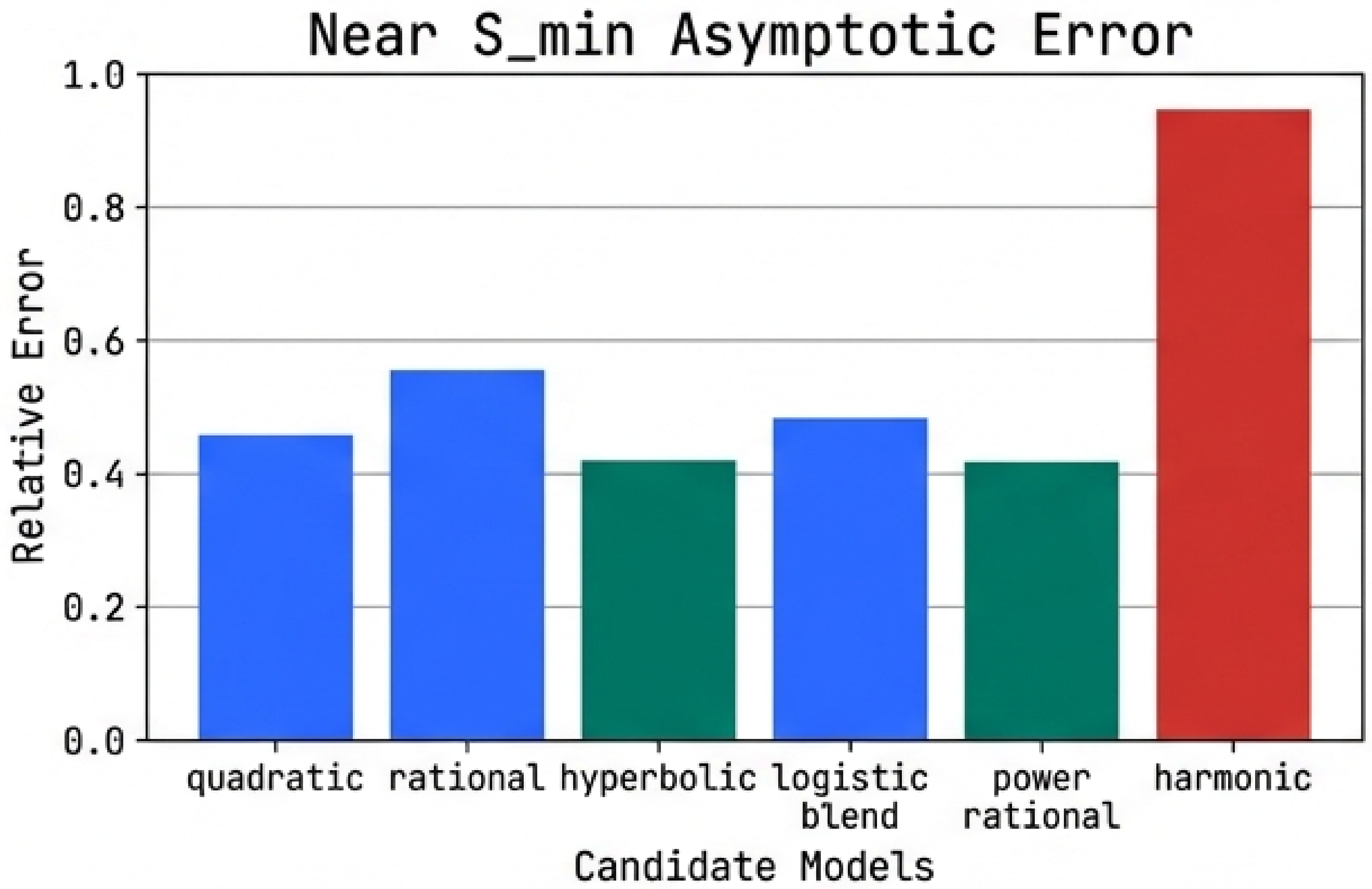


Round 1: Rational and Harmonic forms fail to capture the data structure.

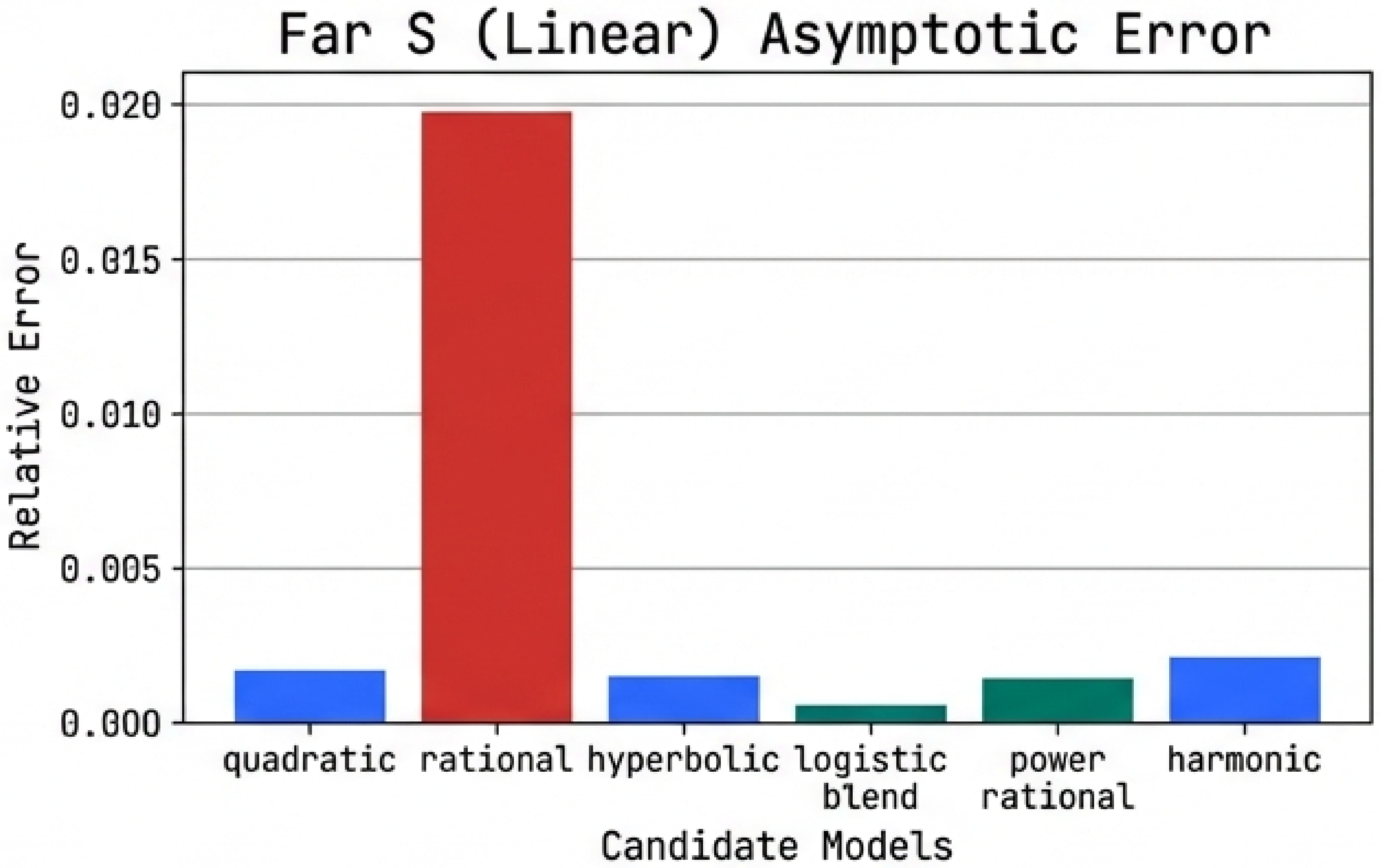
Eliminated: Rational (R^2 ~ 0.71)
& Harmonic (R^2 ~ 0.70)



Round 2: Hyperbolic and Power-Rational forms offer consistent boundary behavior.



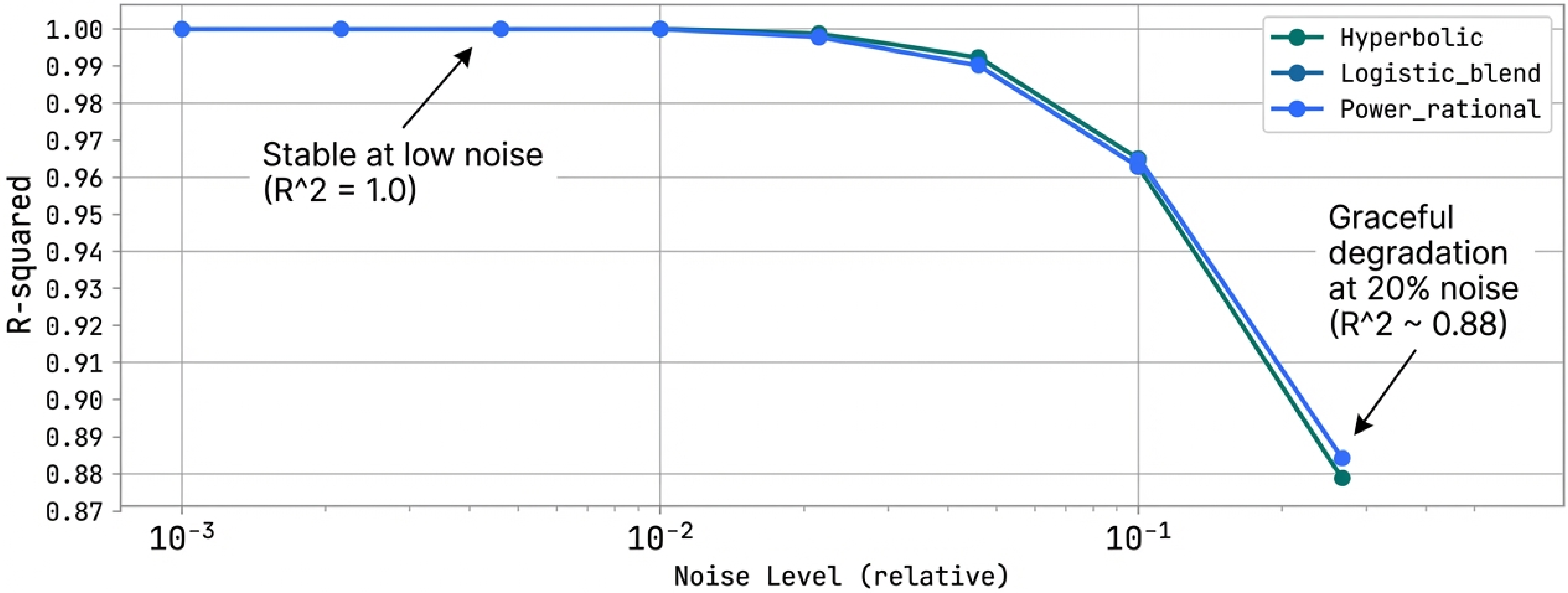
Near S_min: Harmonic fails. Hyperbolic & Power-Rational low error.



Far S: Rational fails massively. Hyperbolic & Power-Rational lowest error.

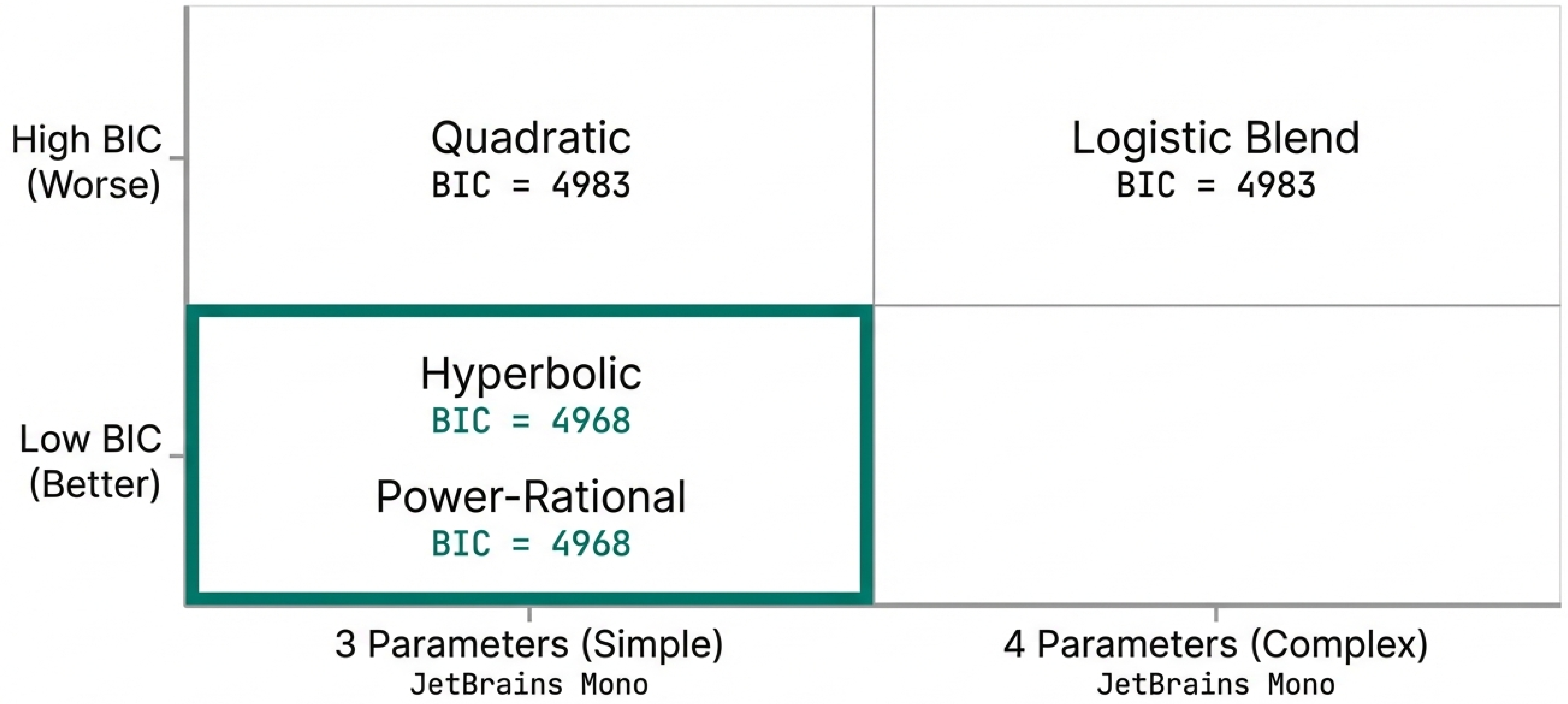
We require a model that naturally satisfies edges without forcing.

Round 3: Top candidates maintain stability up to 20% relative noise.



Real-world training data is noisy. The model must be robust.

Round 4: Simplicity breaks the tie (Occam's Razor).



Logistic Blend eliminated due to unnecessary complexity.

The Final Verdict: Hyperbolic vs. Power-Rational.

Power-Rational Form

R^2: 0.9986

BIC: 4968

Con: Relies on abstract exponent ‘p’. Harder to interpret physically.

Hyperbolic Form

R^2: 0.9986

BIC: 4968

WINNER

Pro: Structural Transparency. Terms map directly to scaling laws.

Deconstructing the Hyperbolic Solution

Linear Regime: Captures behavior at large S (matches alpha * B_crit * S).

$$E(S) = aS + \frac{b * S_{min}}{S - S_{min}} + c$$

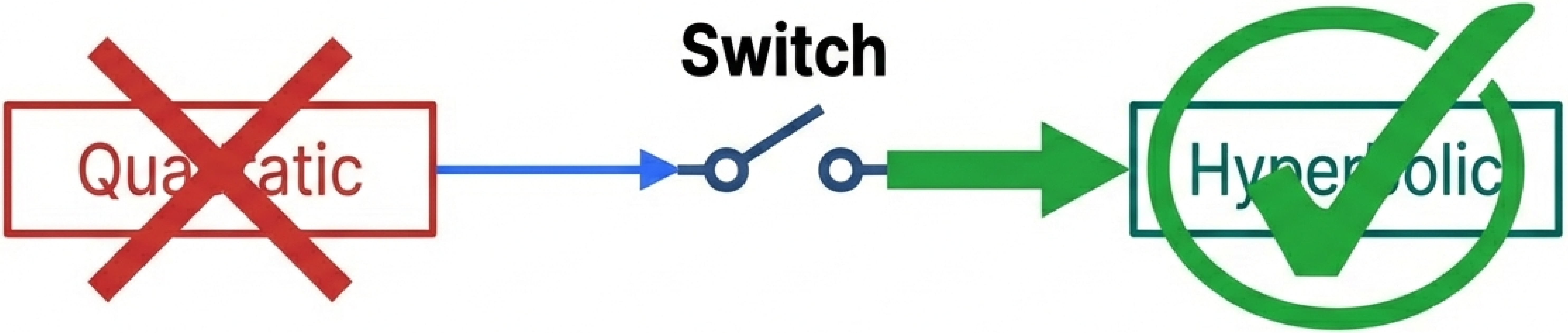
Divergence Regime: Captures inverse-linear explosion near S_min.

Constant Offset.

Statistical Performance Summary (30-Trial Means).

Candidate	R-Squared	BIC	Parameters	Status
Quadratic	0.9985	4983	3	--
Rational	0.7123	6043	4	Poor
Hyperbolic	0.9986	4968	3	**BEST**
Logistic Blend	0.9986	4983	4	--
Power-Rational	0.9986	4968	3	Strong
Harmonic	0.7012	6045	3	Poor

Conclusion & Recommendation.



- We evaluated six forms for the WSD Stable phase intermediate regime.
- **Result:** The Hyperbolic form matches the best fits in accuracy ($R^2 > 0.99$) while maintaining structural simplicity (3 params).
- **Action:** Replace the current ad-hoc quadratic piecewise approximation with the Hyperbolic form.
- **Benefit:** A principled, closed-form basis for scaling laws that naturally satisfies asymptotic constraints.

References

1. Hoffmann et al. (2022) - "Training compute-optimal large language models."
2. Hu et al. (2024) - "MiniCPM: Unveiling the potential of small language models..."
3. Kaplan et al. (2020) - "Scaling laws for neural language models."
4. McCandlish et al. (2018) - "An empirical model of large-batch training."
5. Zhou et al. (2026) - "How to Set the Batch Size for Large-Scale Pre-training?"