Spectral Ratings: Recovering Semantic Capability from Biased Benchmarks

Michael Kaisers
Google DeepMind
Paris, France
mkaisers@google.com

Ian Gemp
Google DeepMind
New York, NY, USA
imgemp@google.com

Marc Lanctot
Google DeepMind
Montreal, Canada
lanctot@google.com

Kate Larson
Google DeepMind, Montreal
University of Waterloo, Waterloo
Canada
katelarson@google.com

Georgios Piliouras
Google DeepMind
London, UK
gpil@google.com

ABSTRACT

Standard aggregate metrics for foundation model evaluation, such as mean accuracy or Elo, implicitly conflate the frequency of each prompt with its overall importance: They can be corrupted by introducing redundant prompts that favor a specific model. To recover how well models span different semantic capabilities, we propose **Spectral Ratings**: we reframe evaluation through spectral geometry where model performance is represented as a quadratic form in the embedding space (a latent feature space where proximity corresponds to semantically similar text). We theoretically prove that Spectral Ratings are robust (change in a bounded way) to exact clones (prompt duplication), and improve robustness to approximate clones (minor prompt variations). Experiments on the HELM-MMLU benchmark expose approximate clones 'in the wild', and illustrate Spectral Ratings and corresponding ranking differences to standard metrics. By penalizing redundancy, the *Spectral Average Rating* favors models with broader semantic coverage, while the full spectral decomposition—including *Spectral Minimum* and *Maximum*—diagnoses worst-case failures and peak skills, offering a holistic characterization of model reliability.

KEYWORDS

Foundation Model Evaluation, Spectral Analysis, Rating, Ranking

1 INTRODUCTION

The rapid ascent of foundation models has shifted the central challenge of the field from architecture design to rigorous evaluation [3, 4]. Research and deployment decisions now rely heavily on aggregate leaderboards based on static benchmarks like MMLU or GSM8K [10, 15]. These leaderboards typically rank models using scalar statistics– most commonly mean accuracy or head-to-head comparisons using Elo [5]– both of which implicitly assume that the benchmark's distribution of input data (evaluation questions) is a valid proxy for real-world utility and general capability.

However, this assumption rarely holds in practice. Benchmarks are often imperfect collections (e.g., MMLU [6]), containing non-uniform clusters of prompts alongside sparse examples of rarer concepts. Standard metrics conflate *frequency* with *importance*: a model can artificially inflate its score by overfitting to dense clusters while failing to generalize [13, 16, 17]. As illustrated in Figure 1, adding redundant prompts can arbitrarily permute rankings, rewarding narrow specialists over robust generalists. This fragility suggests current leaderboards measure alignment of model skills with the arbitrary sampling density of prompts in benchmarks, and improved statistics are needed to assess emergent capabilities [20].

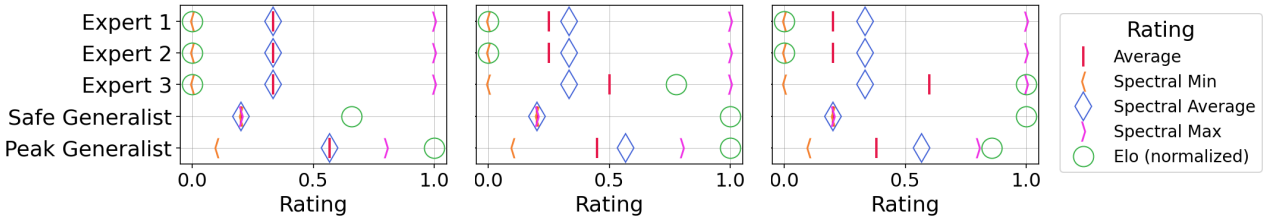
We argue that correcting for this requires reframing evaluation as a geometric problem. Since language models are trained to input and output text, and evaluation datasets are textual questions with textual answers, their semantic capabilities should be compared directly in the appropriate abstract space of language understanding. A standard way to study semantic similarity in natural language processing is by learning **embeddings**: vectors of latent values that codify a piece of text into distinct semantic dimensions [12, 14]. Given an embedding function pre-learned from a large corpus of textual data, we assess each model using a quadratic form involving the model's reported performance across semantic directions. Through this spectral lens, standard average accuracy is revealed to be a density-biased trace operator, which rewards memorizing high-frequency components of the test distribution more than improving on sparser, yet equally critical, directions of capability.

To compensate for this geometric bias in the standard aggregate ratings, we introduce *Spectral Ratings*, a framework that normalizes performance by the *density* (second moment) of the dataset, effectively down-weighting redundant directions to recover a uniform evaluation over the latent semantic manifold. We further introduce a regularization term to interpolate between the raw data distribution and density normalisation, ensuring stability even when the benchmark is low-rank or sparse.

We advance the methodology of foundation model evaluation through the following contributions. **Geometric Formalism (Sections 3.1–3.2):** We reframe model evaluation as a spectral problem, proving that standard mean accuracy is a density-biased estimator, equivalent to the trace of a performance matrix skewed by the dataset's geometry. **Spectral Ratings Framework (Section 3.3):** We introduce *Spectral Ratings*, a family of metrics derived from the Generalized Rayleigh Quotient. By applying a regularized whitening transformation to the embedding space, we derive the *Spectral Average* to estimate capability under a uniform semantic prior, alongside *Spectral Min/Max* ratings to diagnose worst-case

Proc. of the Adaptive and Learning Agents Workshop (ALA 2026), Aydeniz, Delgrange, Mohammedalamen, Yang (eds.), May 25 – 26, 2026, Paphos, Cyprus, https://alaworkshop2026.github.io/. 2026.

Game	Embeddings			Model Evaluation Scores				
	1	2	3	Expert 1	Expert 2	Expert 3	Safe Generalist	Peak Generalist
a	1.0	0.0	0.0	1.0	0.0	0.0	0.2	0.8
b	0.0	1.0	0.0	0.0	1.0	0.0	0.2	0.8
c	0.0	0.0	1.0	0.0	0.0	1.0	0.2	0.1
	0.0	0.0	1.0	0.0	0.0	1.0	0.2	0.1
	0.0	0.0	1.0	0.0	0.0	1.0	0.2	0.1



(a) Original Ratings (b) Ratings with one clone (c) Ratings with two clones

Figure 1: The table (top) shows three game variants, with 3 (a), 4 (b) and 5 (c) prompt embeddings and model scores. Ratings by Average and Elo change drastically, while Spectral Ratings remain stable despite exact clones (prompt duplicates). Spectral Average top-ranks the Peak Generalist (highest performance when averaging over latent directions), and Spectral Min top-ranks the Safe Generalist with the highest worst-case performance. Spectral Max identifies the niche peak performance of experts.

blind spots and peak specialized skills. **Theoretical Robustness (Section 4):** We prove that exact clones (prompt repetition) have bounded effect on Spectral Average, and explicitly express robustness limits to approximate clones. **Empirical Analysis (Section 5):** We compute Spectral Ratings on synthetic data and the HELM-MMLU leaderboard [15], demonstrating they successfully penalize redundancy of clones, stabilizing rankings to favor models with broader semantic coverage.

2 BACKGROUND

This section reviews the evaluation setting for Spectral Ratings, standard aggregation baselines, and the spectral and geometric tools—specifically the Generalized Rayleigh Quotient and embedding spaces—that underpin our method. Finally, we define rank correlation metrics used to quantify leaderboard stability.

2.1 Benchmark Datasets and Rewards

A benchmark dataset $\mathcal{D} = \{p_i\}_{i=1}^N$ consists of N evaluation prompts, such as 14079 test questions in MMLU [10]. Each prompt p_i implies a ground-truth reward function, provided with the dataset: When a model m is evaluated on prompt p_i, the reward function associates its answer with a scalar reward $r_{mi} \in \mathbb{R}$. This reward may be binary ($r_{mi} \in \{0, 1\}$) for multiple-choice questions (as in MMLU) or a continuous preference score for open-ended generation.

2.2 Standard Aggregate Ratings

Leaderboards complement benchmark datasets with achieved reward vectors $r_m = [r_{m1}, \ldots, r_{mN}]^T \in \mathbb{R}^N$ for models m, and traditionally compress the vector r_m into a scalar rating $R(m)$ to rank models by utility. The HELM framework [15] for example contains, among others, a complete evaluation of 76 models on 14042 ($> 99\%$) of the MMLU questions, and publishes ranked ratings online[1].

[1]https://crfm.stanford.edu/helm/mmlu/latest/#/leaderboard

Arithmetic Mean. The ubiquitous arithmetic mean rating weights rewards across all prompts equally: $R_{\text{avg}}(m) = \frac{1}{N}\sum_{i=1}^N r_{mi}$. Consequently, this metric directly varies with sampling frequency; oversampled prompts increasingly influence the average reward.

Elo Rating System. For pairwise comparisons—exemplified by the Chatbot Arena (lmarena.ai) [5]—the Elo system models the probability that model A defeats model B as a logistic function of their rating difference, $P(A \succ B) = \sigma(R_B - R_A)$, where σ denotes the sigmoid function (typically $\sigma(x) = (1 + 10^{-x/400})^{-1}$). Ratings are updated iteratively to minimize prediction error. However, because the underlying objective minimizes loss over the observed set of comparisons, Elo remains frequency-dependent: a model that excels in a high-density cluster of queries will accumulate a higher rating solely due to the volume of comparisons in that domain.

2.3 The Generalized Rayleigh Quotient

Our framework rests on the *Rayleigh Quotient*, a functional used to characterize the spectrum of linear operators (for a comprehensive review, see [7–9]). In Euclidean space, for a symmetric matrix A, it is defined as $\rho(A, x) = \frac{x^T A x}{x^T x}$. When the geometry is distorted by a positive-definite metric or mass matrix C, this extends to the *Generalized Rayleigh Quotient* defined as $\rho(A, C, x) = \frac{x^T A x}{x^T C x}$.

The stationary points of this quotient correspond to the solutions of the generalized eigenvalue problem $Ax = \lambda Cx$. In our context, this ratio compares the model's accumulated reward in a specific direction ($x^T Ax$) against the density of prompts sampled in that direction ($x^T Cx$), effectively normalizing the accumulated reward by the volume of available evidence. Essentially, the quotient estimates performance in any continuous semantic direction x from the finite set of discrete observations, performing local averaging over similar prompts.

2.4 Embeddings and Latent Geometry

Embedding functions map prompts into a geometric space, and are trained such that semantic similarity (or redundancy) manifests as collinearity among vectors (see Chapter 5 in [12]). We assume an embedding function $\phi : \mathcal{X} \to \mathbb{R}^d$ maps each discrete prompt $p_i \in \mathcal{X}$ to a continuous vector $v_i \in \mathbb{R}^d$, e.g. Gecko embeddings [14] map textual prompts to 768-dimensional vectors ($d = 768$).

$$v_i = \frac{\phi(p_i)}{\|\phi(p_i)\|_2} \tag{1}$$

We enforce normalization such that all vectors lie on the unit hypersphere—a common constraint for efficient similarity search [18, 19]. We collect v_i^T as rows of the **Embedding Matrix** $V \in \mathbb{R}^{N \times d}$.

2.5 Comparing Rankings

To quantify how different evaluation metrics permute the leaderboard, we employ **Kendall's** τ, a non-parametric measure of rank correlation. Given two rankings of M models, let n_c be the number of concordant pairs (ordered identically) and n_d be the number of discordant pairs (ordered differently). The coefficient is:

$$\tau = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{M(M-1)/2} \tag{2}$$

We also report the **Kendall Tau Distance** K_τ, which counts the number of pairwise swaps required to transform one ranking into the other ($K_\tau = n_d$). A distance of 0 implies identical rankings, while higher values indicate increasing disagreement between metrics.

3 METHOD: SPECTRAL RATINGS

To rigorously disentangle a model's semantic capability from the prompt distribution biases inherent in benchmark datasets, we re-examine evaluation through the lens of spectral geometry. We first establish that the standard arithmetic mean is effectively a trace operator biased by data density. We then propose Spectral Ratings, which apply a regularized whitening transformation to normalize by density, recovering an estimate of capability over the latent semantic manifold.

3.1 The Geometric View of Averaging

The standard evaluation metric is the arithmetic mean of rewards $\bar{r}_m = \frac{1}{N}\sum r_{mi}$. While traditionally viewed as a scalar statistic, this mean possesses a direct geometric interpretation. Given that embeddings lie on the unit hypersphere, $1 = v_i^T v_i = \mathrm{Tr}(v_i v_i^T)$, we can exploit the cyclic property of the trace operator to inject geometry into the sum:

$$\bar{r}_m = \frac{1}{N}\sum_{i=1}^{N} r_{mi}\mathrm{Tr}(v_i v_i^T) = \frac{1}{N}\mathrm{Tr}\left(\sum_{i=1}^{N} r_{mi} v_i v_i^T\right) \tag{3}$$

We define the term inside the trace as the **Performance Matrix** $A_m \in \mathbb{R}^{d \times d}$, representing the reward-weighted second moment of the prompt embedding distribution:

$$A_m = V^T \mathrm{diag}(r_m) V \tag{4}$$

Since the trace is the sum of eigenvalues, this identity reveals that the standard leaderboard average is the mean eigenvalue of the performance matrix, scaled by the dataset sparsity factor d/N:

$$\bar{r}_m = \frac{d}{N}\left(\frac{1}{d}\mathrm{Tr}(A_m)\right) \tag{5}$$

For model ratings on the same dataset this factor is constant across models, and the average eigenvalue would induce the same ranking as the average; however, such standard spectral statistics implicitly assume dataset density were uniform ($C = I$, compare Section 2.3).

3.2 De-biasing via Regularized Whitening

If our objective is to measure capability as a model's ability to achieve high reward across the latent semantic directions, independent of sampling density, then we must transition from the standard eigenvalue problem to the *Generalized Rayleigh Quotient* (see Section 2.3). The normalization in the generalized quotient relies on the **second moment matrix** $C = V^T V \in \mathbb{R}^{d \times d}$, which encodes the uncentered covariance of the semantic embedding dimensions. Ideally, we would employ the quotient $x^T A_m x/(x^T C x)$ to measure density-normalized performance. However, real-world benchmarks are often low-rank or contain sparse voids where $x^T C x \approx 0$. In these directions, the inverse C^{-1} is undefined or unstable, and naive "whitening" would amplify noise.

To resolve this, we introduce a regularization term $\gamma > 0$, interpreting evaluation as a shrinkage estimation problem. We blend the empirical data density with a uniform prior (the identity matrix), defining the **regularized second moment matrix**:

$$C_\gamma = V^T V + \gamma I \tag{6}$$

This allows us to define a robust whitening transformation. The parameter γ controls the interpolation between the raw data geometry and the uniform prior.

- As $\gamma \to \infty$, $C_\gamma \propto I$, and the Spectral Average Rating introduced in the next section reverts to the standard, density-biased arithmetic mean.
- As $\gamma \to 0$, the metric approaches pure spectral whitening, where rare prompts are up-weighted to possess equal influence to dense clusters.

Intermediate values of γ yield a rating based on *empirically proven performance*, where a model is credited for success in a direction only if the benchmark provides sufficient evidence.

3.3 Spectral Ratings and Extrema

We define the **Spectral Average Rating** as the mean eigenvalue of the whitened system, computed from the trace of the performance matrix normalized by the regularized second moment matrix:

$$\bar{\lambda}_\gamma(m) = \frac{1}{d}\mathrm{Tr}(C_\gamma^{-1} A_m) \tag{7}$$

This metric favors generalist models that maintain consistent performance across the embedding manifold of the benchmark, penalizing strategies that overfit to frequent prompt variations. Figure 1 illustrates this stability on a toy problem: even as prompts are cloned (Games b, c), the Spectral Average correctly maintains the rank of the "Peak Generalist," whereas the average and Elo succumb to bias.

While the Spectral Average provides a robust summary, the full spectrum of generalized eigenvalues $\Lambda = \{\lambda_1, \ldots, \lambda_d\}$ encodes the

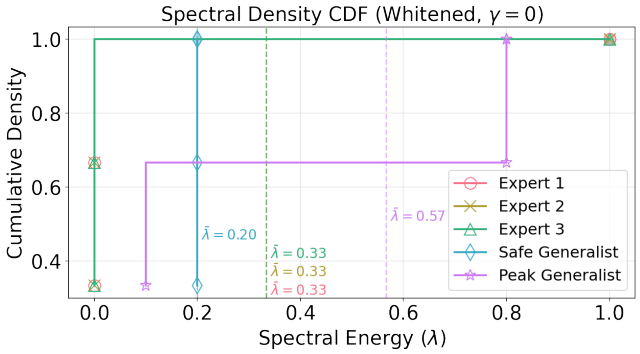


Figure 2: Cumulative spectral density plots of the original game (Figure 1a) with $\gamma = 0$. Dashed vertical lines and labels show Spectral Average Ratings. Markers at specific eigenvalues reveal the distribution of semantic capabilities.

complete capability profile (see Figure 2). We extract two complementary extremal metrics (also shown in Figure 1):

- **Spectral Max ($\lambda_{\max}$):** Identifies the model's "peak skill"—the specific semantic direction where it maximizes reward relative to data density.
- **Spectral Min ($\lambda_{\min}$):** Identifies the "blind spot"—the direction of worst-case alignment. A high $\bar{\lambda}_\gamma$ with a near-zero $\lambda_{\min}$ reveals a model that is generally competent but catastrophically fragile in specific sub-domains.

4 THEORETICAL ANALYSIS

In this section, we analyze the statistical properties of the Spectral Average Rating. We first establish that the rating is equivalent to a weighted average where weights correspond to statistical leverage scores. We then examine the metric's sensitivity to exact and approximate prompt repetitions.

4.1 Spectral De-biasing via Statistical Leverage

To understand how the spectral rating corrects for distribution bias, we decompose the trace formulation into individual sample contributions. This decomposition is governed by the **projection matrix** (or hat matrix) $H = VC_\gamma^{-1}V^T$, the operator that maps the full data space onto the subspace spanned by the embeddings.

Definition 4.1 (Statistical Leverage [2]). *Given a dataset with regularized second moment matrix C_γ, the statistical leverage of the i-th sample with embedding v_i is defined as the i-th diagonal element of the projection matrix, $h_i = v_i^T C_\gamma^{-1} v_i$.*

Recall the definition of the Spectral Average $\bar{\lambda}_\gamma = \frac{1}{d}\mathrm{Tr}(C_\gamma^{-1}A_m)$. Substituting the performance matrix $A_m = \sum_{i=1}^N r_{mi} v_i v_i^T$, we obtain the following characterization.

Proposition 4.2 (Leverage-Weighted Estimation). *The Spectral Average Rating $\bar{\lambda}_\gamma$ is a leverage-weighted sum of the per-sample rewards r_{mi}, where the weights are proportional to the statistical leverage of the samples:*

$$\bar{\lambda}_\gamma = \sum_{i=1}^N w_i r_{mi}, \quad \text{where} \quad w_i = \frac{h_i}{d}. \tag{8}$$

Proof. By the linearity and cyclic property of the trace operator: $\bar{\lambda}_\gamma = \frac{1}{d}\mathrm{Tr}\left(C_\gamma^{-1}\sum_{i=1}^N r_{mi} v_i v_i^T\right) = \frac{1}{d}\sum_{i=1}^N r_{mi}\mathrm{Tr}(C_\gamma^{-1}v_i v_i^T) = \frac{1}{d}\sum_{i=1}^N r_{mi}(v_i^T C_\gamma^{-1} v_i) = \sum_{i=1}^N \frac{h_i}{d} r_{mi}$. □

The weights $\{w_i\}$ form a probability distribution in the unregularized limit ($\gamma \to 0$), where $\sum h_i = d$. For $\gamma > 0$, the weights sum to a total mass $S_\gamma = 1 - \frac{\gamma}{d}\mathrm{Tr}(C_\gamma^{-1})$. The deficit $1 - S_\gamma = \frac{\gamma}{d}\mathrm{Tr}(C_\gamma^{-1})$ reflects the shrinkage property: the metric reserves weight for the uniform prior. In the isotropic limit ($C_\gamma \propto I$)—due to either a perfectly uniform dataset or high regularization ($\gamma \to \infty$)-the leverage becomes constant, implying $w_i = S_\gamma/N$ and thereby recovering the arithmetic mean (scaled by S_γ). In general, however, this formulation reveals the geometric mechanism of de-biasing: samples within dense clusters ($h_i \ll S_\gamma d/N$) are down-weighted, while unique samples in sparse regions ($h_i \gg S_\gamma d/N$) are up-weighted. Notably, these weights can be pre-computed for a given dataset, making model rating efficient—a simple dot product—when new rewards are observed for the same embeddings.

4.2 Sensitivity to Exact Clones

We now analyze the robustness of the metric when the benchmark is augmented by M identical copies ("exact clones") of a specific prompt with embedding v and reward r. The second moment matrix updates as $C'_\gamma = C_\gamma + Mvv^T$. To quantify the impact on existing data, we employ the Sherman-Morrison inverse update. The leverage score h'_i of an existing sample i in the augmented dataset becomes:

$$h'_i = v_i^T (C'_\gamma)^{-1} v_i = h_i - \frac{M(v_i^T C_\gamma^{-1} v)^2}{1 + M h_v}, \tag{9}$$

where $h_v = v^T C_\gamma^{-1} v$ is the leverage of the cloned direction.

Corollary 4.3 (Spectral Cannibalism). *The introduction of clones causes a reduction in the weight of existing samples proportional to their squared cosine similarity with the clone in the whitened space. Specifically, if prompt embedding v_i is C_γ^{-1}-orthogonal to the clone, i.e., $v_i^T C_\gamma^{-1} v = 0$, the corresponding weight remains invariant ($h'_i = h_i$).*

This implies weights penalize redundancy, while preserving the evaluation of skills that are statistically independent of the duplicated prompt given the benchmark's geometry.

4.3 Asymptotic Saturation under Exact Clones

The standard arithmetic mean is arbitrarily sensitive to redundancy: as $M \to \infty$, the average rating converges to the performance on the repeated prompt, r. We prove that the Spectral Average Rating satisfies a **Bounded Influence** property, preventing any single semantic direction from dominating the evaluation.

Theorem 4.4 (Asymptotic Saturation). *Let $\bar{\lambda}_\gamma(M)$ denote the Spectral Rating of a model on a dataset augmented with M identical clones (v, r). As the redundancy $M \to \infty$, the rating converges to:*

$$\lim_{M\to\infty} \bar{\lambda}_\gamma(M) = \bar{\lambda}_{orig} + \frac{1}{d}\left(r - \frac{\beta}{h_v}\right) \tag{10}$$

where $\bar{\lambda}_{orig}$ is the original rating, $h_v = v^T C_\gamma^{-1} v$ is the leverage of the clone direction, and $\beta = v^T C_\gamma^{-1} A C_\gamma^{-1} v$ represents the prior weighted performance in that direction.

Proof. Let the regularized second moment matrix of the augmented dataset be $C'_\gamma = (C + Mvv^T) + \gamma I$. By commutativity, this is equivalent to updating the prior regularized matrix: $C'_\gamma = C_\gamma + Mvv^T$. We apply the Sherman-Morrison formula to invert C'_γ, and substitute $u = C_\gamma^{-1} v$:

$$(C'_\gamma)^{-1} = C_\gamma^{-1} - \frac{MC_\gamma^{-1} vv^T C_\gamma^{-1}}{1 + Mh_v} = C_\gamma^{-1} - \frac{M}{1 + Mh_v} uu^T \tag{11}$$

We substitute this inverse into the Spectral Average Rating, noting $A' = A + Mrvv^T$, and expand the product $(C'_\gamma)^{-1} A'$ into four terms:

$$\begin{aligned} d\bar{\lambda}(M) = \mathrm{Tr}((C'_\gamma)^{-1} A') &= \mathrm{Tr}(C_\gamma^{-1} A) \quad \text{(Pre-clone Trace)} \\ &+ Mr\,\mathrm{Tr}(C_\gamma^{-1} vv^T) \\ &- \frac{M}{1 + Mh_v} \mathrm{Tr}(uu^T A) \\ &- \frac{M^2 r}{1 + Mh_v} \mathrm{Tr}(uu^T vv^T) \end{aligned} \tag{12}$$

Using the cyclic property $\mathrm{Tr}(xy^T) = x^T y$, we resolve the traces to scalars: $\mathrm{Tr}(C_\gamma^{-1} vv^T) = h_v$, $\mathrm{Tr}(uu^T A) = \beta$, and $\mathrm{Tr}(uu^T vv^T) = (u^T v)^2 = h_v^2$. Grouping the terms containing r:

$$Mrh_v - \frac{M^2 r h_v^2}{1 + Mh_v} = Mrh_v \left(1 - \frac{Mh_v}{1 + Mh_v}\right) = \frac{Mrh_v}{1 + Mh_v} \tag{13}$$

Substituting back, the total trace becomes:

$$d\bar{\lambda}_\gamma(M) = d\bar{\lambda}_{\text{orig}} + \frac{M}{1 + Mh_v}(rh_v - \beta) \tag{14}$$

Finally, we take the limit as $M \to \infty$. The fraction $\frac{M}{1+Mh_v}$ converges to $\frac{1}{h_v}$.

$$\lim_{M\to\infty} d\bar{\lambda}_\gamma(M) = d\bar{\lambda}_{\text{orig}} + \frac{1}{h_v}(rh_v - \beta) = d\bar{\lambda}_{\text{orig}} + r - \frac{\beta}{h_v} \tag{15}$$

Dividing by d yields the theorem statement. □

The term β/h_v corresponds to the Generalized Rayleigh Quotient of the prior data along direction v, and as such is bounded by the reward range. Thus, $(r - \beta/h_v)$ represents the *prediction error*—the difference between the clone's reward and the model's expected performance based on prior evidence. The theorem confirms that even an infinite cluster of clones contributes no more to the global rating than this error term scaled by $1/d$, effectively bounding the impact of exact clones on ratings.

4.4 Sensitivity to Approximate Clones

While Theorem 4.4 covers exact duplicates, real-world clusters exhibit *approximate redundancy*. We model such a cluster as M unit vectors $\{z_k\}_{k=1}^M$ drawn from a rotationally symmetric distribution around a central concept v. We quantify the intra-cluster variation via the **Orthogonal Noise Variance** $\sigma^2 = \mathbb{E}[\|(I - vv^T)z_k\|^2]$, implying a signal variance along v of $1 - \sigma^2$.

Assuming the noise is isotropic (i.e., directionally uniform) across the orthogonal dimensions, the cluster's expected additive contribution to the second moment matrix is full-rank:

$$\mathbb{E}[\Delta C] = \mathbb{E}[C'_\gamma - C_\gamma] = \underbrace{M(1 - \sigma^2)}_{\alpha_M} vv^T + \underbrace{\frac{M\sigma^2}{d-1}}_{\epsilon_M}(I - vv^T) \tag{16}$$

Here, the cluster distribution explicitly dictates the spectral update: the mass is split between a concentrated signal term (α_M) and a dispersed noise floor (ϵ_M). The following theorem establishes that the rating update is strictly bounded by this Signal-to-Noise ratio.

Theorem 4.5 (Robustness to Approximate Redundancy). *Let $\Delta\bar{\lambda}$ be the change in Spectral Rating after adding a cluster of M approximate clones with noise variance σ^2 and constant reward r. Let δ be the maximum absolute difference between r and the model's prior generalized Rayleigh quotient in any direction. The magnitude of the rating shift is bounded by:*

$$|\Delta\bar{\lambda}| \le \frac{\delta}{d}\left[\underbrace{\frac{\tilde{\alpha}_M h_v}{1 + \tilde{\alpha}_M h_v}}_{\text{Signal Saturation}} + \underbrace{\frac{(d-1)\epsilon_M \bar{h}}{1 + \epsilon_M \bar{h}}}_{\text{Max Noise Leakage}}\right] \tag{17}$$

where $\tilde{\alpha}_M = \alpha_M - \epsilon_M = M(1 - \frac{d}{d-1}\sigma^2)$ is the excess signal mass, h_v is the prior leverage of the signal direction, and $\bar{h}$ is the average prior leverage of the noise subspace.

The proof is given in Appendix A, and mirrors the proof in Section 4.3 with additional considerations for the noise. The theorem bounds the worst-case instability, with a crucial dependence on σ^2:

- **Tight Limit ($\sigma \to 0$):** The noise leakage vanishes ($\epsilon_M \to 0$). The bound tightens to the signal term, recovering the saturation result of Theorem 4.4 when restricting δ to the local error $|r - \beta/h_v|$.
- **Noise Leakage ($\sigma > 0$):** As cluster variance increases, weight shifts to the second term. The rating remains robust as long as the total noise energy is small relative to the prior density ($M\sigma^2\bar{h} \ll 1$).

5 EXPERIMENTS

We validate the proposed framework through a progressive experimental design. First, we empirically verify the **Bounded Influence** property (Theorem 4.4) in a controlled synthetic environment, demonstrating that Spectral Ratings remain invariant to exact row duplication where standard metrics fail. Second, we test the limits of this robustness under **Approximate Redundancy** (Section 4.4), quantifying the signal-to-noise threshold at which noise aliasing occurs. Finally, we apply the method to the **HELM-MMLU** leaderboard [15], analyzing the intrinsic geometry of approximately 14,000 prompts to show how spectral de-biasing reorganizes the leaderboard to favor robust generalization.

5.1 Clone Robustness of Spectral Ratings in A Minimal Example

To demonstrate the robustness of Spectral ratings against data redundancy, we introduce a minimal experimental setup involving

"cloned" prompts. A robust evaluation metric remains invariant when identical prompts are added to the evaluation set.

We define three game variants, denoted as (a), (b), and (c) in Figure 1. Game (a) consists of three semantically distinct prompts, represented by one-hot orthogonal embeddings. We evaluate five model archetypes: three "Experts" (specialized in a single prompt), a "Safe Generalist" (low but consistent reward), and a "Peak Generalist" (high reward on two prompts, low on the third). To model redundancy, Game (b) introduces a clone of the third prompt, and Game (c) introduces two clones.

As shown in the table, the "Peak Generalist" performs poorly on the third prompt (reward 0.1). Consequently, as we introduce clones of this specific weakness in variants (b) and (c), traditional metrics punish this model disproportionately. The average for the Peak Generalist drops significantly as the test set becomes saturated with the prompt it fails at. Similarly, Expert 3, which specializes in the cloned prompt, sees an artificial inflation in its average and Elo rating.

In contrast, the Spectral ratings demonstrated in the subfigures remain remarkably stable across all three variants. By leveraging the geometric structure of the prompt embeddings, the spectral method identifies that the cloned rows are collinear and do not contribute new information about semantic model capability. As a result, the Spectral Average maintains the "Peak Generalist" as the top-ranking model, correctly disregarding the redundancy. Furthermore, the decomposition allows for targeted insights: Spectral Minimum Rating (=smallest eigenvalue) correctly identifies the "Safe Generalist" as the most robust option (highest worst-case performance), while Spectral Maximum (=highest eigenvalue) highlights the specialized capabilities of the Experts, regardless of how many times their specific prompt is repeated in the benchmark.

5.2 Robustness to Approximate Redundancy

While the theoretical analysis in Section 4.3 guarantees saturation for exact duplicates, real-world benchmarks often exhibit *approximate redundancy*—clusters of distinct prompts that share a core semantic intent but differ in phrasing or parameters. As discussed in Section 4.4, this introduces noise that can compromise the metric if the cluster variance becomes sufficiently large.

To quantify this "Noise Aliasing" limit, we conducted a controlled "Clustered Attack" simulation. We generated a base dataset of $N_{\text{base}} = 100$ prompt embeddings in $d = 64$ dimensions. To ensure a representative unstructured distribution, the base embeddings were sampled from an isotropic Gaussian distribution and projected onto the unit hypersphere. We then injected a redundant cluster of M approximate clones generated via a Projected Normal distribution centered on the first embedding (v_1). Specifically, we perturbed v_1 with i.i.d. Gaussian noise with standard deviation $s = 0.01$ and re-normalized the resulting vectors to unit length, resulting in an angular separation between approximate clones and the cluster center of 4.0-5.0 degrees (10th - 90th percentile). We scaled the cluster size M logarithmically from 1 to $10^{4.2}$. We evaluated the stability of the Spectral Average Rating (with regularization $\gamma = 1.0$) against the standard Arithmetic Mean.

Figure 3 evaluates two synthetic models with reward functions defined by their cosine alignment with the cluster center v_1:

- **The Specialist:** Reward is proportional to its alignment with the cluster: $r(x) = \max(0, x^T v_1)$.
- **The Generalist:** High general reward with a penalty proportional to cluster alignment: $r(x) = 0.8 - 0.6|x^T v_1|$.

We can make two key observations: **1. Arithmetic Mean Failure:** The standard average (dashed lines) is linearly sensitive to the sampling density. As the cluster size approaches the base dataset size ($M \approx 10^2$), the Specialist's high density of rewards allows it to overtake the Generalist, falsely crowning the narrow expert as the superior model. **2. Spectral Robustness:** The Spectral Rating (solid lines) exhibits high resilience to this distribution shift. By effectively measuring the volume of the capability ellipsoid rather than counting successful samples, the metric down-weights the redundant cluster. The "crossover point"—where the accumulated variance of the noise finally overwhelms the semantic signal of the base dataset—is pushed to $M \approx 10^4$, which confirms our theoretical stability condition $M\sigma^2\bar{h} \approx 1$. The injected noise had an effective noise variance per dimension $\sigma^2 = s^2$. With $\bar{h} \approx 1$, the predicted saturation threshold is $M \approx s^{-2} = 10^4$, matching the observed crossover.

This result demonstrates that while Spectral Ratings are not immune to infinite noise accumulation, they provide a robustness margin compared to standard averaging, effectively handling the levels of approximate redundancy typical in large-scale benchmarks.

5.3 Spectral Ratings in HELM-MMLU

While the synthetic experiments in previous subsections confirm that Spectral Ratings possess the theoretical property of robustness to cloning, the critical empirical question is whether such redundancies exist in widely adopted public benchmarks. This section introduces the public benchmark dataset MMLU, analyses the existence of exact and approximate clones, and applies our Spectral Ratings to it.

5.3.1 Massive Multitask Language Understanding (MMLU). We base our experiments on the public *Massive Multitask Language Understanding (MMLU)* benchmark [10], with evaluation scores provided within the Holistic Evaluation of Language Models (HELM) framework [15]. MMLU is the de facto standard for assessing general

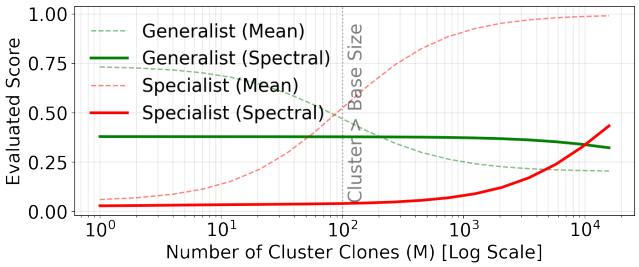


Figure 3: The Arithmetic Mean (dashed) allows the Specialist (red) to overtake the Generalist (green) as soon as the M clones outnumber the 100 base observations. The Spectral Rating (solid) maintains the correct ranking up to extreme redundancy levels ($M \approx 10^4$).

knowledge, comprising 57 distinct "subjects" ranging from *Elementary Mathematics* to *Professional Law*, with 100-1534 multiple choice questions per subject. After pre-processing the public dataset to remove models with ambiguous or incomplete evaluations, we retained 76 models evaluated on 14042 questions each. While prompts in HELM may be formatted in model-specific ways, our evaluation is concerned with embedding the *multiple-choice question* implied by each prompt. Hence, we concatenate the question and answer options of each multiple-choice question into a canonical string, and use Gecko-1b to compute 768-dimensional embeddings [14], which we cache as a 14042×768 matrix (see Appendix B for details).

5.3.2 Clones in MMLU. There are 212 exact clones in MMLU, arising from 106 multiple choice questions each appearing twice: 78 questions overlap between 'clinical knowledge' and 'college medicine', and 28 questions repeat within subjects: college physics (11), elementary mathematics (1), high school psychology (11), prehistory (1), professional psychology (1), public relations (2), US foreign policy (1).

To provide a principled reference for **approximate clones**, we employ a heuristic inspired by statistical anomaly detection. We analyze the distribution of nearest-neighbor distances by plotting the **empirical quantile function** (Figure 4), i.e. sorted angular distances against their cumulative fraction. While the global geometry of embeddings on a hypersphere is naturally non-linear, we observe a predictable (for similarities on a high-dimensional sphere), linear-like "bulk" regime.

To model this typical background density, we fit a linear regression to the central 90% of the data. We then identify the "clone tail"—prompts significantly more similar to their neighbors than expected—by calculating the residual standard deviation (σ) of the fit. Adopting a standard 4σ outlier threshold, we define the boundary where empirical distances deviate precipitously below the trend. This yields a cut-off of 27.4°, capturing the bottom 5.3% of prompts. By this metric, there are 740 approximate semantic clones in the MMLU benchmark, of which 212 are exact string matches. As detailed in Appendix C, qualitative inspection confirms that this threshold effectively identifies semantic redundancy, ranging from minor typographical differences (1-15° angular distance in embedding space) to broad rephrasing of related concepts (15-27°).

5.3.3 Spectral Ratings in MMLU. Standard reporting on MMLU typically aggregates scores via a micro-average (treating all questions as equal) or a macro-average (averaging over subject groupings). Both approaches rely on the assumption that the discrete elements, i.e. prompts or subject labels ("Math", "History") respectively, accurately partition the semantic space into equal-importance regions.

By contrast, the Spectral Rating discards these manual labels entirely, and instead embeds each multiple choice question. We compute the Mass Matrix C_γ directly from the embeddings of the 14,042 canonical prompts, allowing the geometry of the embedding space to reveal the true density of the evaluation.

Intrinsic Dimensionality and Skew. To assess the geometric distortion of the MMLU benchmark, we analyzed the spectrum of its empirical Mass Matrix C. The raw condition number (at $\gamma = 0$) is approximately 3.8×10^4, indicating extreme anisotropy; the dataset

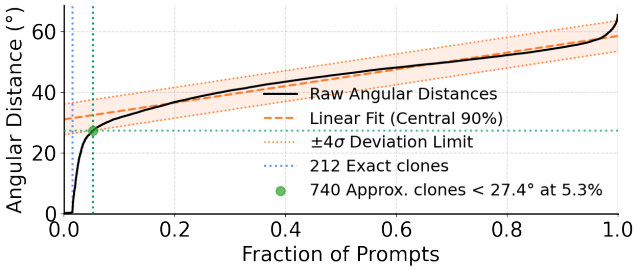


Figure 4: Angular distance to the closest neighboring question in the MMLU benchmark, using Gecko-1b embeddings. We fit a linear regression (orange dashed line) to the central 90% of the raw empirical distances sorted by rank (solid black line). The green circle marks the point where the left tail drops more than 4σ (orange dotted line) below the linear prediction, identifying the threshold of 27.4° and 740 approximate clones, of which 212 are exact clones.

Table 1: Pairwise Kendall Tau Distances and Correlations (τ) for ranking by Average $\bar{r}$, Spectral Minimum $\lambda_{\min}$, Spectral Average $\bar{\lambda}$, Spectral Maximum $\lambda_{\max}$, and Subject Average $\bar{r}_s$.

	$\bar{r}$	$\lambda_{\min}^{\gamma=10}$	$\bar{\lambda}^{\gamma=10}$	$\lambda_{\max}^{\gamma=10}$	$\bar{r}_s$
$\lambda_{\min}^{\gamma=10}$	48 (.97)				
$\bar{\lambda}^{\gamma=10}$	21 (.99)	35 (.98)			
$\lambda_{\max}^{\gamma=10}$	143 (.90)	123 (.91)	134 (.91)		
$\bar{r}_s$	55 (.96)	51 (.96)	36 (.97)	136 (.90)	
$\bar{\lambda}^{\gamma=0}$	22 (.98)	34 (.98)	1 (1.00)	133 (.91)	37 (.97)

variance is compressed into a few dominant directions while leaving vast subspaces of the embedding manifold nearly empty. This confirms that the "natural" distribution of MMLU is highly degenerate. Applying regularization effectively dampens this skew: setting $\gamma = 1.0$ reduces the condition number to $\sim 3.5 \times 10^3$, and $\gamma = 10$ further stabilizes it to $\sim 3.8 \times 10^2$. We adopt $\gamma = 10$ for all Spectral Ratings in experiments.

Model rating comparison. Figure 5 provides a visual overview of the rankings resulting from different rating metrics. Ratings across all 76 models correlate globally, but may change order locally (see also Appendix D, Figure 7). The pairwise ranking distances are given in Table 1. This table shows that the geometric Spectral Average approach results in rankings that are closer to the Average and Subject Average rankings than they are to each other. The Spectral Average ranking between $\gamma = 0$ and $\gamma = 10$ only differs by one pair. In additional experiments, we scaled to γ to 100 and 1000, resulting in 3 (1.00) and 12 (0.99) Kendall-Tau Distance and correlation (τ) respectively, compared to $\bar{\lambda}^{\gamma=0}$. These results confirm that Spectral Ratings are stable in practice, providing a principled alternative to traditional distribution-biased metrics.

Figure 6 shows a detailed comparison of rankings for models that rank in the top ten in any of the ratings. Within this top ten, the Spectral Average agrees with the plain average, while both disagree with the subject average—all but two models in this subset

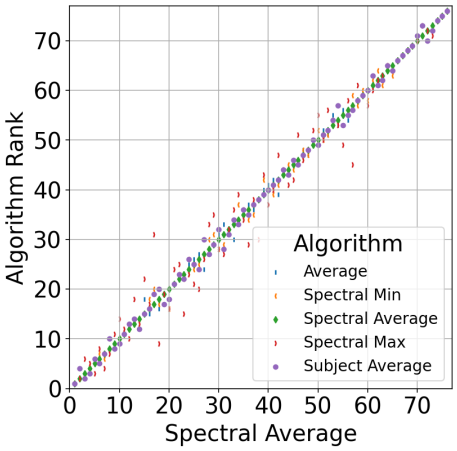


Figure 5: Rank-rank plot in MMLU ($\gamma = 10$). Ranks are globally consistent and vary locally across algorithms.

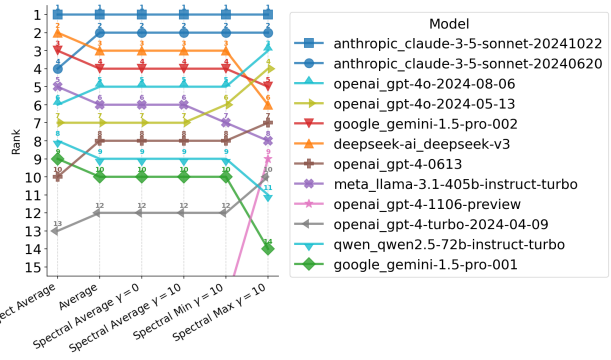


Figure 6: Bump chart of models that appear in top ten of any rating, based on the HELM-MMLU dataset.

change rank, highlighting the potential impact of dataset biases on ranking.

6 DISCUSSION AND CONCLUSIONS

The Geometry of Evaluation. The central thesis of this work is that model evaluation is fundamentally a geometric problem, not an arithmetic one. Standard metrics implicitly accept sampling frequency as a proxy for prompt importance, rendering leaderboards unstable under redundancy. By mapping evaluation into the latent semantic space, Spectral Ratings decouple the *measurement* of performance from the *distribution* of the test set. The whitening transformation ($C_\gamma^{-1} A_m$) effectively creates a "semantic effective sample size," ensuring that a dense cluster of variations contributes no more to the global score than a single unique concept.

Kernel Function Bias. By adopting this geometric framework, we have fundamentally replaced the inductive bias of the prompt distribution with the inductive bias arising from the combination of the embedding function and the chosen similarity metric. Specifically, because the Generalized Rayleigh Quotient evaluates performance through a ratio of quadratic forms, it scores linear projections based on squared cosine similarities—structurally behaving as a squared inner-product kernel. To relax the global and linear geometric assumptions inherent to this specific formulation, future work may explore kernelized extensions. Substituting the standard inner product with Radial Basis Functions (RBF) or learned similarity metrics would map discrete observations to continuous performance estimates using localized, non-linear inductive biases.

Dependency on Embedding Quality. A critical consideration is the dependence on the embedding function ϕ. The metric assumes that orthogonality in the embedding space reflects true semantic distinctness. If the embedding model collapses distinct concepts or artificially separates synonyms, the Spectral Rating will inherit these biases. However, this dependency offers a path for improvement: as representation models improve (e.g., via instruction-tuned embeddings), the resolution and fairness of the spectral evaluation will naturally increase. Furthermore, the framework is agnostic to the source of embeddings, allowing for domain-specific representations (e.g., code or medical) to derive specialized scores, or to apply our methodology to other modalities (e.g., images, sound, robotic control policies).

Relation to Game-Theoretic Evaluation. Our approach parallels game-theoretic efforts to establish clone-invariant metrics, such as Nash Averaging [1] and Deviation Ratings [17]. However, a critical distinction remains: game-theoretic methods operate in *payoff space*, potentially interpreting noisy samples within a cluster as distinct strategies. In contrast, Spectral Ratings operate in the *latent semantic space*, where the continuous covariance manifold naturally smooths out local irregularities. This allows us to quantify the effective "volume" of a model's capabilities—a robust signal that pure payoff-based rankings do not capture.

Implications for Benchmark Design. Historically, benchmark creation required painstaking curation to ensure balance. Our findings suggest a paradigm shift: if metrics can mathematically correct for sampling bias post-hoc, the pressure to curate perfectly balanced datasets is alleviated. Researchers can aggregate "messy," unstructured data from diverse sources—logs, user queries, scraped exams—and rely on the Regularized Mass Matrix C_γ to automatically down-weight redundancies. This enables the use of significantly larger, naturally occurring datasets without skewing the leaderboard toward the most frequent data modes.

Conclusion. We have introduced **Spectral Ratings**, a theoretically grounded framework that recovers intrinsic model capability from biased benchmarks. By modeling performance as a quadratic form and applying a whitening transformation, we prove that this metric exhibits Bounded Influence, saturating the impact of redundant prompts where standard averages fail. Empirical validation confirms that this geometric correction penalizes overfitting to prompt clusters and rewards consistent generalization. As foundation models continue to scale, moving from simple arithmetic averages to geometrically aware metrics will be essential for measuring true progress in general semantic capabilities.

REFERENCES

[1] David Balduzzi, Karl Tuyls, Julien Perolat, and Thore Graepel. 2018. Re-evaluating Evaluation. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc., 3272–3283.

[2] David A. Belsley, Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley & Sons, New York.

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[4] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* 15, 3 (2024), 1–45.

[5] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *International Conference on Machine Learning*. PMLR, 8359–8388.

[6] Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. 2025. Are we done with mmlu?. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 5069–5096.

[7] Benyamin Ghojogh, Fakhri Karray, and Mark Crowley. 2019. Eigenvalue and generalized eigenvalue problems: Tutorial. *arXiv preprint arXiv:1903.11240* (2019).

[8] Gene H Golub and Henk A Van der Vorst. 2000. Eigenvalue computation in the 20th century. *J. Comput. Appl. Math.* 123, 1-2 (2000), 35–65.

[9] Gene H Golub and Charles F Van Loan. 2013. *Matrix computations*. JHU press.

[10] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

[11] Roger A Horn and Charles R Johnson. 2012. *Matrix analysis*. Cambridge university press.

[12] Daniel Jurafsky and James H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/ Online manuscript released January 6, 2026.

[13] Marc Lanctot, Kate Larson, Yoram Bachrach, Luke Marris, Zun Li, Avishkar Bhoopchand, Thomas Anthony, Brian Tanner, and Anna Koop. 2025. Evaluating Agents using Social Choice Theory. arXiv:2312.03121 [cs.AI] https://arxiv.org/abs/2312.03121

[14] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. 2024. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327* (2024).

[15] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=iO4LZibEqW

[16] Siqi Liu, Ian Gemp, Luke Marris, Georgios Piliouras, Nicolas Heess, and Marc Lanctot. 2025. Re-evaluating Open-ended Evaluation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=c113

[17] Luke Marris, Siqi Liu, Ian Gemp, Georgios Piliouras, and Marc Lanctot. 2025. Deviation Ratings: A General, Clone-Invariant Rating Method. *arXiv preprint arXiv:2502.11645* (2025). https://doi.org/10.48550/arXiv.2502.11645

[18] Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. 2019. Spherical text embedding. *Advances in neural information processing systems* 32 (2019).

[19] Simon Roburin, Yann de Mont-Marin, Andrei Bursuc, Renaud Marlet, Patrick Perez, and Mathieu Aubry. 2022. Spherical perspective on learning with normalization layers. *Neurocomputing* 487 (2022), 66–74.

[20] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems* 36 (2023), 55565–55581.

A DERIVATION OF APPROXIMATE REDUNDANCY BOUNDS

In this appendix, we prove Theorem 4.5. We first confirm the second moment matrix update, derive an exact expansion of the rating update, distinguishing between signal and noise components, and finally apply Jensen's Inequality to establish the bound.

A.1 Second Moment Matrix Update

Let $\{z_k\}_{k=1}^M$ be random unit vectors drawn from a distribution on the hypersphere $\mathbb{S}^{d-1}$ that is rotationally symmetric around a central concept direction v (where $\|v\| = 1$). Any realization z can be uniquely decomposed into a component along v and a component in the orthogonal complement:

$$z = c \cdot v + s \cdot u \tag{18}$$

where u is a unit vector orthogonal to v ($u \perp v, \|u\| = 1$), $c = v^T z$ is the cosine similarity, and $s = \sqrt{1-c^2}$ is the magnitude of the orthogonal projection.

The **Orthogonal Noise Variance** is defined as the expected squared magnitude of the projection onto the complement:

$$\sigma^2 = \mathbb{E}[\|(I - vv^T)z_k\|^2] = \mathbb{E}[s^2] \tag{19}$$

Because the vectors lie on the unit hypersphere ($c^2 + s^2 = 1$), the variance along the signal direction is strictly determined by this noise: $\mathbb{E}[c^2] = 1 - \sigma^2$.

The clones expected contribution to the second moment matrix is $\mathbb{E}[\Delta C] = M\mathbb{E}[zz^T]$. Note that v is not random, hence expanding the outer product yields:

$$\mathbb{E}[zz^T] = \mathbb{E}[c^2]vv^T + \mathbb{E}[s^2uu^T] + \mathbb{E}[cs(vu^T + uv^T)] \tag{20}$$

Because the distribution is rotationally symmetric around v, the density of z depends strictly on the cosine similarity c. Conditioned on c, the orthogonal direction u is uniformly distributed on the unit hypersphere in the $(d-1)$-dimensional orthogonal complement. Consequently, the conditional expectation of u is zero ($\mathbb{E}[u \mid c] = 0$), which causes the cross-terms $\mathbb{E}[cs(vu^T + uv^T)]$ to vanish entirely. Furthermore, the uniform distribution of u implies isotropic noise across the orthogonal dimensions, yielding $\mathbb{E}[uu^T] = \frac{1}{d-1}(I - vv^T)$. Substituting these expectations yields the full covariance update:

$$\mathbb{E}[\Delta C] = \underbrace{M(1-\sigma^2)}_{\alpha_M} vv^T + \underbrace{\frac{M\sigma^2}{d-1}}_{\epsilon_M}(I - vv^T) \tag{21}$$

Here, α_M represents the **Effective Signal Mass**, and ϵ_M represents the **Dispersed Noise Mass** per dimension.

A.2 Exact Expansion of the Rating Update

To perform the inversion, we treat the dispersed noise $\epsilon_M(I - vv^T)$ as a global regularization update. We define the **Noise-Augmented Prior** matrices:

$$\tilde{C} = C_\gamma + \epsilon_M I \tag{22}$$

$$\tilde{A} = A + r\epsilon_M I \tag{23}$$

The fully perturbed system is then a rank-1 perturbation of this augmented base: $C' = \tilde{C} + \tilde{\alpha}_M vv^T$ and $A' = \tilde{A} + r\tilde{\alpha}_M vv^T$, where $\tilde{\alpha}_M = \alpha_M - \epsilon_M$.

We compute the new rating $d\bar{\lambda}_{\text{new}} = \text{Tr}((C')^{-1}A')$. Using the Sherman-Morrison formula for $(C')^{-1}$ and expanding the product yields four terms, identical in structure to the exact clone derivation in Theorem 4.4:

$$\begin{aligned} d\bar{\lambda}_{\text{new}} = \text{Tr}(\tilde{C}^{-1}\tilde{A}) \quad &\text{(Augmented Base)} \\ &+ r\tilde{\alpha}_M \text{Tr}(\tilde{C}^{-1}vv^T) \\ &- \frac{\tilde{\alpha}_M}{1 + \tilde{\alpha}_M \tilde{h}_v} \text{Tr}(\tilde{u}\tilde{u}^T\tilde{A}) \\ &- \frac{r\tilde{\alpha}_M^2}{1 + \tilde{\alpha}_M \tilde{h}_v} \text{Tr}(\tilde{u}\tilde{u}^T vv^T) \end{aligned} \tag{24}$$

where $\tilde{u} = \tilde{C}^{-1}v$ and $\tilde{h}_v = v^T\tilde{C}^{-1}v$. Using the cyclic property of the trace to resolve the matrix products to scalars (e.g., $\text{Tr}(\tilde{u}\tilde{u}^T\tilde{A}) = \tilde{u}^T\tilde{A}\tilde{u} = \tilde{\beta}$), and grouping the terms:

$$d\bar{\lambda}_{\text{new}} = \text{Tr}(\tilde{C}^{-1}\tilde{A}) + \frac{\tilde{\alpha}_M}{1 + \tilde{\alpha}_M \tilde{h}_v}(r\tilde{h}_v - \tilde{\beta}) \tag{25}$$

Subtracting the original rating $d\bar{\lambda}_{\text{old}} = \text{Tr}(C_\gamma^{-1}A)$ gives the exact net update:

$$d\Delta\bar{\lambda} = \underbrace{\left[\text{Tr}(\tilde{C}^{-1}\tilde{A}) - \text{Tr}(C_\gamma^{-1}A)\right]}_{\Delta_{\text{noise}}} + \underbrace{\left[\frac{\tilde{\alpha}_M}{1 + \tilde{\alpha}_M \tilde{h}_v}(r\tilde{h}_v - \tilde{\beta})\right]}_{\Delta_{\text{signal}}} \tag{26}$$

A.3 Proof of Absolute Bounds

We seek to bound the magnitude of the total update $|\Delta\bar{\lambda}|$. By the Triangle Inequality, this is bounded by the sum of the magnitudes of the signal and noise components:

$$|\Delta\bar{\lambda}| \le \frac{1}{d}\left(|\Delta_{\text{signal}}| + |\Delta_{\text{noise}}|\right) \tag{27}$$

Let $\rho(x) = \frac{x^TAx}{x^TC_\gamma x}$ be the Generalized Rayleigh Quotient of the model on the prior distribution (without clones). We define the global prediction error bound δ as:

$$\delta = \sup_{x \neq 0} |r - \rho(x)| \tag{28}$$

1. Bounding the Signal Term. Recall the signal update derived in C.2:

$$\Delta_{\text{signal}} = \frac{\tilde{\alpha}_M}{1 + \tilde{\alpha}_M \tilde{h}_v}(r\tilde{h}_v - \tilde{\beta}) \tag{29}$$

The term $(r\tilde{h}_v - \tilde{\beta})$ can be rewritten as $\tilde{h}_v(r - \tilde{\beta}/\tilde{h}_v)$, where $\tilde{\beta}/\tilde{h}_v = \rho(v)$ is the Rayleigh quotient along the signal direction (evaluated against the augmented prior). This is bounded by $\delta\tilde{h}_v$. Next, we observe that the leverage is monotonically decreasing with respect to dataset mass. Since $\tilde{C} = C_\gamma + \epsilon_M I \succeq C_\gamma$, the inverse satisfies $\tilde{C}^{-1} \preceq C_\gamma^{-1}$, implying:

$$\tilde{h}_v = v^T\tilde{C}^{-1}v \le v^TC_\gamma^{-1}v = h_v \tag{30}$$

The saturation function $g(x) = \frac{\tilde{\alpha}_M x}{1 + \tilde{\alpha}_M x}$ is monotonically increasing for $x > 0$. Therefore, replacing the augmented leverage $\tilde{h}_v$ with the

larger prior leverage h_v yields a conservative upper bound:

$$|\Delta_{\text{signal}}| \leq \delta \frac{\tilde{\alpha}_M \tilde{h}_v}{1+\tilde{\alpha}_M \tilde{h}_v} \leq \delta \underbrace{\frac{\tilde{\alpha}_M h_v}{1+\tilde{\alpha}_M h_v}}_{\text{Signal Saturation}} \tag{31}$$

Interlude on the trace as a sum of Rayleigh quotients. Since the trace is the sum of diagonal elements in any orthonormal basis $\{u_j\}$ (see e.g. 2.4.1 in [11]), we can write $\text{Tr}(C^{-1}A) = \sum_{j=1}^{d} u_j^T C^{-1} A u_j$. If we specifically choose $\{u_j\}$ to be the eigenbasis of the symmetric matrix C, with corresponding eigenvalues μ_j, we apply the property $u_j^T C^{-1} = \frac{1}{\mu_j} u_j^T$ to obtain:

$$\text{Tr}(C^{-1}A) = \sum_{j=1}^{d} \frac{u_j^T A u_j}{\mu_j} \tag{32}$$

This identity allows us to cleanly map the global performance estimate onto our specific spectral components.

2. Bounding the Noise Term. Note that C_γ and the regularized matrix $\tilde{C} = C_\gamma + \epsilon_M I$ share the same eigenvectors u_j, while the eigenvalues shift: $\tilde{\mu}_j = \mu_j + \epsilon_M$. Using the eigenbasis of C_γ and applying the trace expansion above, we can directly evaluate the trace difference:

$$\begin{aligned}\Delta_{\text{noise}} &= \text{Tr}(\tilde{C}^{-1}\tilde{A}) - \text{Tr}(C_\gamma^{-1}A) \\ &= \sum_{j=1}^{d} \left(\frac{u_j^T (A + r\epsilon_M I) u_j}{\mu_j + \epsilon_M} - \frac{u_j^T A u_j}{\mu_j} \right)\end{aligned} \tag{33}$$

Let $\beta_j = u_j^T A u_j$ be the performance projected onto the j-th direction. The summand simplifies algebraically by isolating the common denominator:

$$\frac{\beta_j + r\epsilon_M}{\mu_j + \epsilon_M} - \frac{\beta_j}{\mu_j} = \frac{\mu_j(\beta_j + r\epsilon_M) - \beta_j(\mu_j + \epsilon_M)}{\mu_j(\mu_j + \epsilon_M)} = \frac{\epsilon_M}{\mu_j + \epsilon_M}\left(r - \frac{\beta_j}{\mu_j}\right) \tag{34}$$

Recall that for any direction x, leverage is defined as $h(x) = x^T C_\gamma^{-1} x$. For the orthogonal eigenvectors u_j, this quadratic form evaluates exactly to the inverse eigenvalue: $h_j = u_j^T C_\gamma^{-1} u_j = 1/\mu_j$. We identify $\rho_j = \beta_j h_j$ as the prior Generalized Rayleigh Quotient in direction j.[2] Substituting these into the coefficient: $\frac{\epsilon_M}{\mu_j+\epsilon_M}\left(r - \frac{\beta_j}{\mu_j}\right) = \frac{\epsilon_M h_j}{1+\epsilon_M h_j}(r - \rho_j) \geq 0$.

Using δ, the magnitude of the noise term is thus:

$$|\Delta_{\text{noise}}| \leq \sum_{j=1}^{d} \frac{\epsilon_M h_j}{1+\epsilon_M h_j} |r - \rho_j| \leq \delta \sum_{j=1}^{d} \frac{\epsilon_M h_j}{1+\epsilon_M h_j} \tag{35}$$

The function $f(x) = \frac{\epsilon_M x}{1+\epsilon_M x}$ is concave for $x \geq 0$. By **Jensen's Inequality**, the average of the function values is less than or equal to the function of the average value. Summing over the $d-1$ orthogonal noise dimensions (where $\bar{h}$ is the average leverage):

$$|\Delta_{\text{noise}}| \leq \delta(d-1)\frac{\epsilon_M \bar{h}}{1+\epsilon_M \bar{h}} \tag{36}$$

Conclusion. Combining the bounds for $|\Delta_{\text{signal}}|$ and $|\Delta_{\text{noise}}|$ and dividing by d yields the inequality stated in Theorem 4.5.

[2]Unlike Section 4.3, where β was defined in a doubly whitened space and required division by h_v, here β_j has not been whitened yet, and requires multiplication by h_j.

B HELM MMLU DATA DOWNLOADING AND PRE-PROCESSING

The public dataset was obtained from Holistic Evaluation of Language Models (HELM, 15) by following instructions[3] for downloading the MMLU data subset[4] on 30 September, 2025. Runs included versions `1.0.0` until `1.13.0`. For data cleaning, we dropped any model with ambiguous data due to inconsistent evaluation counts or missing values, resulting in the removal of 5 models, with 76 models remaining, evaluated on 57 subjects.
Removed due to evaluation counts being inconsistent w.r.t. remaining models, in particular having duplicate subject evaluations:

- `anthropic_claude-3-opus-20240229`
- `anthropic_claude-3-sonnet-20240229`
- `nvidia_nemotron-4-340b-instruct`

Removed due to missing data:

- `google_gemini-pro` missing most subject evaluations,
- `ai21_jamba-1.5-mini` missing prompts, required for computing embeddings.

The subject ratings computed from this data approximate the published rankings[5] with Kendall Tau distance of 4, and inspection showed these four pairs swapped neighbouring ranks.

B.1 Composing canonical inputs

The dataset contains prompts and inputs, as well as choices for the multiple-choice administered MMLU questions. Since prompts may vary across models, e.g. adhering to specific required formatting of model prompts, the concatenation of inputs and choices was used to uniquely identify the evaluation context.

Each evaluation run for a (model, subject) pair yields a file `instances.json` with a list of instances. Each instance dict has a field 'input', mapping to a dict with key 'text' mapping to the value we shall refer to as **query**, e.g. "Find the degree for the given field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q.". In addition, each instance dict has a field 'references', mapping to a list of answers, each characterised as a dict with 'output' mapping to a dict with 'text' mapping to the corresponding answer string. We collect the sorted answer strings into an **answer ballot** `"\n".join(f"- {z}" for z in sorted(x))`, where x are all answer strings for the instance. The canonical **prompt** is composed by concatenating the query with the answer ballot, resulting in 13936 distinct prompts `query + "\n" + answer_ballot`. The total of 14042 evaluations thus includes 106 duplicate pairs.

Example canonical prompt

```
Find the degree for the given field extension Q(sqrt(2), sqrt(3),
↪ sqrt(18)) over Q.
- 0
- 2
- 4
- 6
```

[3]https://crfm-helm.readthedocs.io/en/latest/downloading_raw_results/

[4] `gs:// crfm-helm-public/mmlu/benchmark_output`

[5]https://crfm.stanford.edu/helm/mmlu/latest/#/leaderboard

The subject assignments were kept for later re-grouping of prompts into subject-level metrics.

C QUALITATIVE ANALYSIS OF APPROXIMATE CLONES

To qualitatively assess the nature of semantic redundancy beyond trivial exact string matches, Table 2 presents a representative sample of approximate clone pairs identified in the HELM-MMLU benchmark. To ensure a diverse and unbiased cross-section, these examples were systematically selected by sampling seven evenly spaced ranks across the identified tail of semantic clones. Specifically, the sampling spans from the first approximate clone (the pair with the smallest non-zero angular distance) up to the 740th rank, which corresponds to the data-driven anomaly threshold established in Section 4 (Figure 4).

As demonstrated in the table, these approximate clones capture a wide spectrum of dataset redundancy that standard exact-match deduplication scripts fundamentally fail to detect. The highlighted textual differences reveal several distinct categories of approximate clones:

- **Minor Typographical Deviations:** Pairs separated by trivial formatting differences, such as missing punctuation, errant whitespace, or slight notation shifts (e.g., 1.1°).
- **Direct Paraphrasing and Swaps:** Questions that test identical underlying concepts using slightly altered vocabulary or inverted scenarios, such as swapping "ceiling" for "floor" or substituting a specific mathematical constant (e.g., 22.7°).
- **Shared Contextual Boilerplate:** Completely distinct multiple-choice questions that are appended to identical, extensive reading comprehension passages or document excerpts. Because the massive shared context dominates the string length, these prompts project to highly localized regions in the embedding space, correctly identifying them as structurally dependent evaluations (e.g., 24.8°).

By visualizing the raw text alongside the angular distances, it becomes evident that while these pairs are not strict identical strings, they represent heavily concentrated evaluation mass. This reinforces the necessity of utilizing continuous embedding distances, rather than discrete string matching, to accurately quantify and correct for dataset bias.

Dist.	Subject	Prompt 1	Prompt 2
1.1°	college_physics	Protons used in cancer therapy are typically accelerated to about 0.6c. How much work must be done on a particle of mass m in order for it to reach this speed, assuming it starts at rest? - $0.25mc^2$ - $0.60mc^2$ - $0.67mc^2$ - $1.25mc^2$	Protons used in cancer therapy are typically accelerated to about 0.6c . How much work must be done on a particle of mass m in order for it to reach this speed, assuming it starts at rest? - $0.25mc^2$ - $0.60mc^2$ - $0.67mc^2$ - $1.25mc^2$
11.2°	global_facts	Which of the following countries emitted the most CO2 per capita in 2017? - Canada - Iran - Japan - Russia	Which of the following countries emitted the most CO2 in 2017? - Canada - Iran - Japan - Russia
18.6°	conceptual_physics	A principal source of the Earth's internal energy is - gravitational pressure. - radioactivity. - solar radiation. - tidal friction.	The principal source of Earth's internal energy is - geothermal heat - gravitational pressure - radioactivity - tidal friction
22.7°	high_school_microeconomics	Which of the following is true about a price ceiling? - It is used to correct government policy. - It is used when equilibrium prices are too low. - It will be located above the equilibrium price. - It will be located below the equilibrium price.	Which of the following is true about a price floor? - It is used to correct government policy. - It is used when the equilibrium price is too high. - It will be located above the equilibrium price. - It will be located below the equilibrium price.
24.8°	high_school_world_history	This question refers to the following information. The passage below is the Chinese emperor's response to English King George III's diplomatic envoys, who were seeking expanded trading privileges (1793). Strange and costly objects do not interest me. If I have commanded that the tribute offerings sent by you, O King, are to be accepted, this was solely in consideration for the spirit which prompted you to dispatch them from afar. . . . As your Ambassador can see for himself, we possess all things. I set no value on objects strange or ingenious, and have no use for your country's manufactures. It behooves you, O King, to display even greater devotion and loyalty in future, so that, by perpetual submission to our Throne, you may secure peace and prosperity. According to the passage, what was the Chinese reaction to the British goods? - Awe at their technological superiority - Fascination with their strangeness - Interpreting them as an act of submission - Offense at a perceived bribe	This question refers to the following information. The passage below is the Chinese emperor's response to English King George III's diplomatic envoys, who were seeking expanded trading privileges (1793). Strange and costly objects do not interest me. If I have commanded that the tribute offerings sent by you, O King, are to be accepted, this was solely in consideration for the spirit which prompted you to dispatch them from afar. . . . As your Ambassador can see for himself, we possess all things. I set no value on objects strange or ingenious, and have no use for your country's manufactures. It behooves you, O King, to display even greater devotion and loyalty in future, so that, by perpetual submission to our Throne, you may secure peace and prosperity. Why were the Chinese not interested in expanding trading rights with Great Britain? - The Chinese had a preexisting exclusive trade agreement with the Dutch. - The Chinese were angry over Britain's interference in their foreign affairs. - They had no interest in the products that Great Britain could provide. - They were afraid that the British would gain too much influence within China.
26.3°	abstract_algebra	Statement 1| Every group of order p^2 where p is prime is Abelian. Statement 2 | For a fixed prime p a Sylow p-subgroup of a group G is a normal subgroup of G if and only if it is the only Sylow p-subgroup of G. - False, False - False, True - True, False - True, True	Statement 1 | If H and K are subgroups of G and one of H or K is normal subgroup of G, then HK is a subgroup of G. Statement 2 | All groups of order p^2 where p is prime are Abelian. - False, False - False, True - True, False - True, True
27.4°	high_school_microeconomics	Marginal cost (MC) is equal to average variable cost (AVC) and average total cost (ATC) when: - AVC and ATC intersect MC at its maximum point. - AVC and ATC intersect MC at its minimum point. - MC intersects AVC and ATC at their minimum points. - marginal cost (MC) intersects AVC and ATC at their maximum points.	Which of the following is true about the relationship of the average total cost (ATC) curve and the marginal cost (MC) curve? - ATC and MC are always equal. - ATC and MC are never equal. - The ATC curve intersects the MC curve at the minimum point of the MC curve. - The MC curve intersects the ATC curve at the minimum point of the ATC curve.

Table 2: Examples of approximate clone pairs identified in the MMLU dataset, drawn at equally spaced quantiles. Prompts are reproduced exactly as they appear in the benchmark, with textual differences highlighted.

D MODEL RATINGS IN MMLU

Figure 7 provides a complete list of model ratings. Regularisation biases spectral ratings towards zero, but avoids fitting to noise in the null space. The Spectral Minimum and Maximum ratings provide an envelop of performance in the worst and best embedding direction respectively.

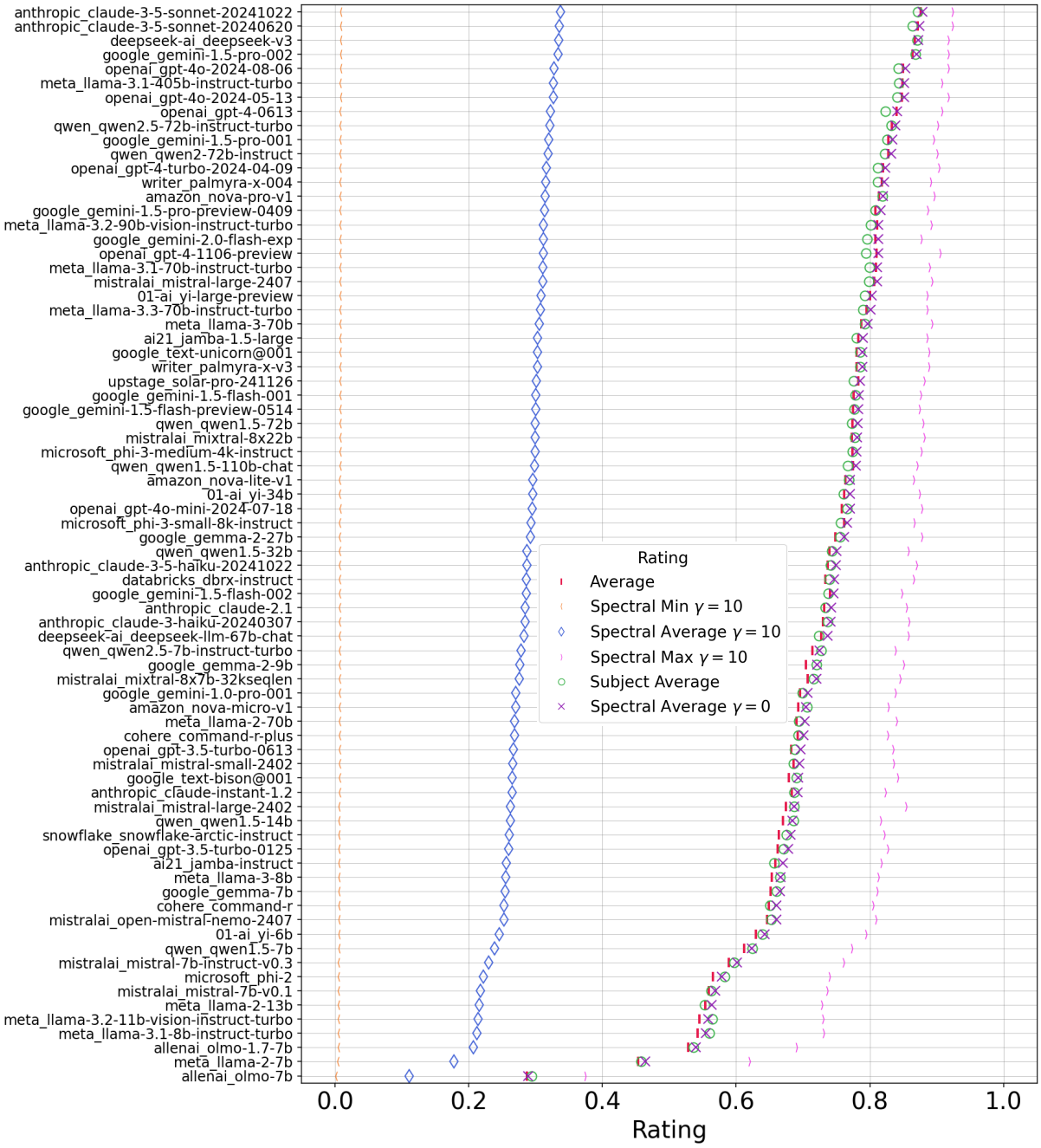


Figure 7: Model ratings in HELM-MMLU, sorted by Spectral Average ($\gamma = 10$).