From Entropy to Generalizability: Strengthening Automated Essay Scoring Reliability and Sustainability

Yi Gui
University of Iowa
yi-gui@uiowa.edu

Abstract

Generalizability Theory with entropy-derived stratification optimized automated essay scoring reliability. A G-study decomposed variance across 14 encoders and 3 seeds; D-studies identified minimal ensembles achieving $G \geq 0.85$. A hybrid of one medium and one small encoder with two seeds maximized dependability per compute cost. Stratification ensured uniform precision across complexity.

1 Introduction

Automated Essay Scoring (AES) systems based on transformer architectures have transformed large-scale writing assessment by offering both scalability and consistency that rival human raters (Shermis & Burstein, 2013). However, before deployment in high-stakes contexts—such as college admissions or professional licensure—AES models must meet stringent psychometric standards for reliability. In Classical Test Theory, indices like Cronbach's alpha confound multiple error sources into a single coefficient, obscuring the distinct contributions of prompt heterogeneity, model stochasticity, and text complexity (Nunnally & Bernstein, 1994). **Generalizability Theory (G-Theory)** overcomes these limitations by decomposing observed-score variance into multiple facets and interactions, yielding the generalizability coefficient (G-coefficient) as an overall index of dependability (Brennan, 2001; Cronbach et al., 1972).

In any G-Theory study, the object of measurement—the entity whose universe score we seek to estimate—is critical. For AES, as in large-scale writing assessments like the GMAT AWA, these objects are the test-takers themselves (Gao et al., 2015). Here, each essay's observed model score is viewed as an estimate of that test-taker's "universe score" across all combinations of prompts, model seeds, and cross-validation folds.

A persistent obstacle to AES reliability is uneven precision across essays of varying complexity: richly worded, syntactically complex essays tend to yield larger residual errors, thereby reducing the G-coefficient for those subsets (Shermis & Burstein, 2013). **Shannon entropy**—computed from token-frequency distributions—offers a principled, near–zero-cost measure of textual complexity (Shannon, 1948), but raw entropy correlates strongly ($r > .80$) with essay length. To decouple length from unpredictability, we standardize entropy (z-score) and stratify essays into equal-size buckets that contribute uniformly to variance-component estimates.

The present study applies a Generalizability-Theory approach to automated essay scoring by incorporating standardized entropy buckets into both G-Study and D-Study phases. After estimating variance components for prompts, entropy strata, encoder architectures, seed initializations, and folds, we analytically predict G-coefficients for alternative measurement designs—varying the number and combination of transformer models and random seeds—without refitting mixed models. Our aim is to identify minimal ensembles that meet a target reliability ($G \geq 0.85$), thereby guiding more efficient and cost-effective AES deployments.

This study is guided by three primary questions.

1) In a fully crossed G-study of test-takers (essays) × encoders × seeds × folds, what proportions of total score variance are attributable to encoder choice, seed

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 312–328

initialization, and fold assignment, and how practically meaningful are these facets for overall reliability?

2) Can essays stratified into equal-size buckets based on standardized Shannon entropy be seamlessly integrated into a G-Theory design, thereby supporting more precise D-study planning?

3) In the D-study phase, which minimal combinations of encoder architectures and seed replications suffice to reach the predefined reliability threshold ($G \geq 0.85$), and how does this entropy-informed approach reduce unnecessary computation in AES system development?

2 Related Work

Evolution of Scoring Reliability Methods in Writing Assessment.

Early research on essay scoring focused on Classical Test Theory (CTT) concepts of reliability, using measures like inter-rater correlation or Cohen's kappa to judge consistency between human graders. A kappa value of 0.7 is often cited as an acceptable minimum for essay scoring reliability, as it accounts for roughly half of score variance. Traditional writing assessments thus aimed for high inter-rater agreement to ensure reliable scores. However, CTT-based reliability is limited in that it offers a single coefficient (like Cronbach's alpha or inter-rater correlation) for a given test under fixed conditions, without disentangling multiple error sources. This is problematic for essay tasks, where score variance may arise from multiple facets – differences in raters, prompts, or occasions. Researchers recognized that a more nuanced framework was needed beyond what CTT provides.

Generalizability Theory (G-Theory) emerged as that framework, extending CTT by incorporating analysis of variance to parse out various sources of measurement error (persons, raters, tasks, etc.) and to estimate the dependability of scores under varying conditions. Cronbach et al. (1972) introduced the G-Theory model, defining the generalizability coefficient (G-coefficient) as an index analogous to reliability that reflects how well observed scores generalize to the universe of all possible scoring conditions. Unlike a single CTT reliability, the G-coefficient can account for, say, multiple essay prompts or rater inconsistencies simultaneously. Subsequent works (Brennan, 2001; Shavelson & Webb, 1991) further formalized G-Theory and its use in performance assessments. In writing assessment research, G-Theory has been used to examine how many raters or prompts are needed to achieve dependable scores and to diagnose where inconsistencies arise. For instance, Huang (2008) applied G-theory to ESL writing tests and found that using multiple tasks and raters significantly improved score accuracy. These studies demonstrated that classical inter-rater reliability indices could be inadequate, and that G-Theory provides deeper insight into the facets affecting essay score reliability.

Generalizability Theory Applied to AES.

As Automated Essay Scoring (AES) systems have matured—especially with the advent of transformer-based models—researchers have turned to Generalizability Theory (G-Theory) to rigorously evaluate their reliability. Williamson et al. (2012) first argued that an AES engine must demonstrate stable performance not just on a single prompt but across diverse essay tasks and forms. In the high-stakes context of standardized tests—where the objects of measurement are the test-takers—Gao et al. (2015) extended that framework in their GMAC AWA study by modeling facets such as prompts, essay types (fixed), rating engines versus human raters, and occasions. Because a fully crossed design was impractical, they employed overlapping G-Studies and D-Studies to approximate universe-score variance, reporting operational G-coefficients around 0.83.

Subsequent empirical work has confirmed the value of this approach. Han and Sari (2024) applied G-Theory to compare human raters with ETS's e-rater on a set of EFL essays, finding that human raters introduced more score variance than the automated engine. When automated and human scores were combined, overall dependability improved—underscoring how AES can mitigate human-rater inconsistency. Bridgeman et al. (2012) similarly showed that an AES system maintained comparable reliability across gender and ethnic subgroups, suggesting that machine scoring does not exacerbate demographic biases in measurement error.

Together, these studies illustrate two key points: first, that G-Theory provides a nuanced, facet-level

understanding of where scoring variability arises; and second, that modern AES engines can achieve reliability on par with—or even exceeding—that of human raters, provided they are evaluated across a representative range of prompts and test-taker populations.

Entropy-Based Approaches and Score Uncertainty

Information-theoretic metrics—most notably Shannon entropy—have gained traction as tools for characterizing essay complexity and anticipating scoring uncertainty in automated systems. In classical educational measurement, conditional standard errors of measurement (CSEM) acknowledge that precision can vary across score levels or item types; by analogy, essays whose linguistic patterns are highly unpredictable may elicit greater variability in both human and machine-generated scores. Shannon entropy, computed from an essay's token-frequency distribution, quantifies this unpredictability: higher entropy reflects richer vocabulary and structural diversity, which can challenge consistency in scoring. Several studies have extended this concept by measuring relative entropy (Kullback–Leibler divergence) between an essay's word distribution and a reference language model, demonstrating that essays with greater divergence tend to produce more dispersed human ratings (Atkinson & Palma, 2025).

Other work has leveraged entropy of next-token probabilities from pretrained transformers to flag low-confidence segments, finding that these high-entropy regions correspond to larger prediction errors. Such entropy-derived features thus serve as data-driven proxies for CSEM, identifying essays on which automated scorers are likely to be less reliable (Atkinson & Palma, 2025).

Although entropy-informed methods are not yet ubiquitous in production AES pipelines, they offer a promising complement to aggregate reliability indices and G-Theory analyses. Integrating entropy as either a stratification facet or a diagnostic feature enables more nuanced dependability assessments, pinpointing when an individual essay score may warrant caution. As AES systems continue to evolve, embedding information-theoretic insights can enhance both psychometric rigor and operational transparency, ensuring that automated evaluations maintain equitable precision across the full spectrum of student writing complexity.

3 Methods

Data

The PERSUADE 2.0 corpus originally comprised over 25,000 persuasive essays from U.S. students in grades 6–12. We restricted our analysis to ninth through twelfth graders, yielding 13,815 unique essays each annotated with a holistic score and writer demographics (prompt, task type, grade, gender, socioeconomic and ELL status). Because each examinee submitted exactly one essay, essay IDs are fully confounded with test-taker identity, so all essay-level variance components mirror examinee-level differences. To ensure balanced coverage across topics, tasks, and grade levels, we performed a stratified split by prompt × task × grade, sequestering 10% (1,381 essays) as an unlabeled "Holdout Evaluation Set" and retaining 12,428 essays for G-study and D-study modeling (Figures 1–2).

Each essay in the holdout set was then scored according to a fully crossed design of 14 transformer encoders × 3 random seeds × 5 cross-validation folds, for a total of 14 × 3 × 5 = 210 predictions per essay. All encoders shared a common ordinal logistic regression (OLR) head during fine-tuning and inference. Inference proceeded by tokenizing each essay into sliding-window segments matched to model-specific maximum sequence lengths, pooling segment representations into a single fixed-length embedding, and passing that embedding through the OLR head. The resulting 210 predicted scores per essay were collated into one dataset. The wall-clock times for the models are all presented in Table 10.[1]

To examine complexity effects, we computed raw Shannon entropy from token-frequency distributions (Figure 3, r = 0.741 with word count), then standardized these values (z-scores) and applied our equal-variance bucketing algorithm. Standardization inverted the length correlation (r =

1 Wall-clock time refers to the real-world elapsed time measured from the start to the end of each processing step, encompassing all computation, data loading, and any idle or I/O wait times—analogous to timing an event with a stopwatch.

−0.641), equalizing variability across essay lengths. Plotting sample variance against standardized entropy (Figure 4) reveals a flat cloud of points, indicating no systematic rise in inconsistency for more complex texts. Likewise, mean predicted score remains constant across entropy levels (Figure 5), confirming that our stratification yields uniform precision (variance) and fairness (mean score) regardless of textual complexity.

Entropy Standardization and Bucketing Method

To control for the confounding effect of essay length on token-based complexity measures, we first computed each essay's Shannon entropy H_i from its token-frequency distribution:

$$H_i = -\sum_t p_{i,t} log\ (p_{i,t}) \quad (1)$$

where $p_{i,t}$ is the relative frequency of token t in essay i. These raw entropy values were then standardized to z-scores

$$z_i = \frac{H_i - \bar{H}}{SD(H)} \quad (2)$$

centering and scaling the distribution so that z_i has mean zero and unit variance across the corpus.

Because our goal was to ensure that each complexity stratum contributed equally to the D-study's variance estimates, we partitioned the essays into buckets by **equalizing the sum of their observed score variances** rather than by simple quantile splits of z_i. Denoting by v_i the sample variance of the 210 predictions for essay i, we sought a set of cut-points $\{c_1, \ldots, c_{K-1}\}$ that induced buckets $B_k = \{i: c_{k-1} < z_i \leq c_k\}$ satisfying

$$\sum_{i \in B_k} v_i \approx \frac{1}{K} \sum_{i=1}^{N} v_i \ \ for\ k = 1, \ldots, K \quad (3)$$

Starting from a maximum K, we iteratively reduced K until every prompt × bucket cell contained at least the pre-specified minimum number of essays. In practice, we sorted essays by z_i, formed cumulative sums of v_i, and placed cut-points at entropy values corresponding to equal increments of total variance. Essays were then assigned to buckets by thresholding their z_i against these cut-points. This "equal-variance" binning guarantees that each entropy stratum contributes the same total score dispersion—and, by enforcing a minimum cell size per prompt, preserves adequate data in every prompt × bucket combination for reliable variance-component estimation.

G-Study Design

To decompose score variance across measurement facets, we fit a series of linear mixed-effects models in R using the lme4 package. After reshaping the predictions into long format—one record per essay × encoder × seed × fold—and subsetting by entropy bucket, we specified the model

$$Predicted\_score_{ijkl} = \mu + u_{p[i]} + v_{b[i]} + w_{p[i],b[i]} + x_{e[i]} + \varepsilon_{ijkl} \quad (4)$$

Where $u_{p[i]} \sim N(0, \sigma_p^2)$ captures prompt-level variance, $v_{b[i]} \sim N(0, \sigma_b^2)$ the bucket (entropy) effect, $w_{p[i],b[i]} \sim N(0, \sigma_{pb}^2)$ their interaction, $x_{e[i]} \sim N(0, \sigma_e^2)$ and the essay (test-taker) facet, and $\varepsilon_{ijkl} \sim N(0, \sigma_r^2)$ residual error. From each model we extracted the variance components $\sigma_p^2, \sigma_b^2, \sigma_{pb}^2, \sigma_e^2, \sigma_r^2$ via VarCorr(), supplying the inputs for subsequent analytic D-study computations.

D-Study Design

Using those variance components, we predicted the generalizability coefficient G without refitting:

$$G = \frac{\sigma^2_{test-taker}}{\sigma^2_{test-taker} + \frac{\sigma^2_{prompt:bucket}}{n_{buckets}} + \frac{\sigma^2_{residual}}{n_{buckets} \times n_e \times n_s}} \quad (5)$$

Here $n_{buckets}$, n_e, and n_s are the numbers of entropy buckets, encoders, and seeds. Two sweeps were performed:

Small-only ensembles: All $\binom{6}{k} \times \binom{6}{s}$ combinations of k=2–6 small encoders and s=1–3 seeds (399 runs) identified the minimal small-model sets achieving G≥0.85.

Mixed small + medium ensembles: For each subset of 1–4 medium encoders, we incrementally added 1–6 small encoders under 1–3 seeds (5,401 runs), halting further expansion once a medium–seed pair first attained G≥0.95.

Large-model exploration: Finally, we considered ensembles incorporating at least one large encoder (BERT-Large, RoBERTa-Large, GPT-2 Large, DeBERTa-V3 Large), applying a second early-stop rule: after two distinct large-inclusive sets surpassed G≥0.95, no further large-model sweeps were conducted.

These analytic D-study procedures rigorously chart reliability gains against computational expense, guiding selection of AES ensembles that meet dependability targets with minimal overhead.

4 Results

1. Overall Reliability Ceiling

The fully crossed G-study design—incorporating all 14 encoders, three seed initializations, and five-fold cross-validation—yielded near-perfect generalizability coefficients across entropy strata. Substituting $n_e = 14$, $n_s = 3$, and $n_{buckets} = 3$ into the analytic D-study formula produced

Low-entropy bucket: G=0.990

Mid-entropy bucket: G=0.989

High-entropy bucket: G=0.989

An overall average of G≈0.99 confirms that exhaustive model diversity and replication effectively eliminate measurement error, establishing an upper bound against which all reduced-complexity ensembles are benchmarked.

After experimenting with different K values under a minimum cell-size constraint, we determined that only by requiring at least seven essays per prompt–bucket cell could the equal-variance algorithm produce three strata. Higher K values violated this constraint, and K = 2 proved too coarse for reliable variance estimation. With the minimum cell size set to seven, the algorithm converged on K = 3 buckets containing 451, 479, and 457 essays (Table 2), which we label as low, medium, and high complexity based on their mean standardized-entropy (z) scores. Table 3 details each prompt's essay counts within these buckets, confirming an even distribution of texts across topics and complexity levels.

Figure 6 illustrates that the holistic score distributions are virtually identical across low, medium, and high entropy strata. Each bucket peaks in the mid-score range (2–4) and tapers symmetrically toward the extremes (1 and 6), with no stratum showing a disproportionate concentration at any particular score band. This consistency confirms that our equal-variance bucketing preserved the overall score profile: textual complexity, as indexed by standardized entropy, does not systematically bias the distribution of human-assigned scores.

2. G-Study Facet Contributions

Despite this high ceiling, the initial G-study decomposition (Table 4) reveals that genuine test-taker differences remain the predominant source of score variability. Across all three entropy strata, the essay/test-taker component σ_T^2 accounted for approximately 60–65 % of total variance. Residual error σ_R^2contributed another 11–14 %. In contrast, encoder choice σ_E^2 explained only 12–18 %, seed initialization σ_S^2 2–4 %, and fold assignment σ_F^2 less than 1 %. Interaction terms—encoder × seed, encoder × fold, seed × fold—were effectively zero (Table 4).

Thus, although true-score variance dominates AES reliability, the nontrivial contributions of model architecture and random seed justify their explicit modeling in both G- and D-study phases.

3. Small-Only D-Study: Trading Seeds for Model Diversity

A full sweep of the six "small" transformers (ELECTRA Small (Discriminator), DistilBERT-base (uncased), DeBERTa-v3-small, GPT-2 (small), MiniLM-L6-uncased, MobileBERT-uncased) crossed with 1–3 seeds (399 designs with 354 of them obtained G values ≥ 0.85) revealed clear trade-offs between architectural diversity and seed replication (Table 5; Figure 7). Under a single-seed design, two- and three-encoder ensembles fail to reach G=0.85, but four small encoders just meet it (G≈0.851), and adding one or two more models raises reliability to approximately 0.868 and 0.882, respectively. Introducing a second seed yields larger gains: three encoders under two seeds already surpass G=0.85, and six encoders exceed 0.90. With three seed replications, even the minimal two-model pairing (GPT-2-small + MobileBERT) reaches G≈0.921; adding a third model pushes G to roughly 0.942, and four-model ensembles climb to about 0.952. Beyond four encoders, marginal improvements taper off—five- and six-model ensembles with three seeds achieve G≈0.958 and 0.962, respectively (Figure 8).

Figure 9 renders these results as smooth surfaces in the (number of encoders, number of seeds, G-coefficient) space. All three seed planes rise steeply from two to three encoders before plateauing, indicating diminishing returns on adding more models. Conversely, each additional seed shifts the entire surface upward by an almost constant amount, confirming that seed replication is a more

efficient lever for boosting reliability once a modest ensemble is in place. In resource-constrained settings (e.g., limited GPU memory), two or three encoders with three seeds offer the fastest route to G ≥ 0.85; where parallelism is abundant, larger ensembles can edge closer to the reliability ceiling but with diminishing payoff.

4. Mixed D-Study: Leveraging Medium-Sized Models

To explore hybrid configurations, we swept 5,401 designs combining 1–4 medium transformers, 1–6 small transformers, and 1–3 seed replications. As in the small-only study, an early-stop rule was enforced: for each medium–seed pairing, additional small models were added only until G≥0.85 was first reached (see Table 8). We then ranked each ensemble by a simple "resource" metric (number of models + number of seeds) to identify the most cost-effective solutions (Table 9).

Remarkably, a two-model hybrid—one medium encoder (BERT-Base), one small encoder (DistilBERT-base), and two seeds (resource = 4)—achieved G≈0.895, exceeding every small-only design at the same resource level. This single-medium + single-small configuration appears as the top entry in Table 9. Although adding more small models or seeds continued to raise G, the incremental benefit per added compute unit diminished rapidly.

Encouraged by these mid-range gains, we briefly evaluated large transformers under a second early-stop: once two large-inclusive ensembles surpassed G≥0.95, we terminated further exploration on cost-benefit grounds. The reliability uplift from large models (≈1–2 percentage points) did not justify their 5–10× higher FLOPs, memory footprint, and energy cost.

Together, Tables 8 and 9 illustrate that compact mixed ensembles—anchored by a single medium transformer, a single small transformer, and minimal seed replication—deliver the highest generalizability per unit of compute.

5 Discussion

The D-Study sweeps reveal a clear trade-off between reliability and computational cost. Small-only configurations must marshal at least five combined encoder-and-seed resources to surpass a generalizability coefficient of 0.85, whereas a compact hybrid ensemble of one medium transformer, one small transformer, and two seed replications achieves approximately 0.90 with only four resources. This efficiency frontier underscores that thoughtfully chosen model diversity and minimal replication can meet rigorous reliability thresholds while markedly curbing FLOPs, GPU memory, and energy consumption.

Our G-Study also confirms that large transformers—despite reducing encoder-facet variance by just one to two percentage points—incur disproportionately high compute and environmental costs, making medium-sized architectures the practical backbone for high-stakes scoring. By treating each essay (and thus each examinee) as the object of measurement, the mixed-effects framework captures both content and ability variance in a single term, aligning our dependability estimates with established psychometric practice. Moreover, entropy-stratified bucketing ensured that low-, mid-, and high-complexity texts contributed equally to variance-component estimation, guarding against bias from essay length or richness and validating the fairness of our reliability analyses.

Looking ahead, enriching our simple resource metric with actual FLOPs, GPU-hours, and energy use would enable truly multi-objective D-studies. Exploring adaptive or continuous stratification and designs with multiple essays per examinee could further disentangle content from ability variance and broaden applicability beyond holistic scoring.

6 Conclusion

This study integrates standardized Shannon-entropy stratification within a Generalizability-Theory framework to guide efficient AES ensemble design. A fully crossed G-Study (14 encoders × 3 seeds × 5 folds) quantified error sources across prompts, entropy strata, encoder models, and seed initializations. Analytic D-Study formulas then predicted generalizability coefficients for over 5,800 hypothetical ensembles without refitting models, revealing that compact hybrids of medium and small transformers with limited seed replication achieve target reliability at minimal cost. By balancing psychometric rigor with computational pragmatism, our approach offers a principled roadmap for deploying reliable, fair, and sustainable AES systems across diverse writing complexities.

References

Atkinson, J., & Palma, D. (2025). An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports*, *15*(1), 14551. https://doi.org/10.1038/s41598-025-87862-3

Brennan, R. L. (2001). *Generalizability theory* [doi:10.1007/978-1-4757-3456-0]. Springer-Verlag Publishing. https://doi.org/10.1007/978-1-4757-3456-0

Bridgeman, B., Catherine, T., & and Attali, Y. (2012). Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. *Applied Measurement in Education*, *25*(1), 27-40. https://doi.org/10.1080/08957347.2012.635502

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley and Sons, Inc. https://doi.org/10.3102/00028312011001054

Gao, X., Brennan, R. L., & Guo, F. (2015). *Modeling Measurement Facets and Assessing Generalizability in a Large-Scale Writing Assessment* (GMAC ® Research Reports, Issue.

Han, T., & Sari, E. (2024). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning*, *37*(4), 961-985.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach. *Assessing Writing*, *13*(3), 201-218. https://doi.org/https://doi.org/10.1016/j.asw.2008.10.002

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shavelson, R., & Webb, N. (1991). *Generalizability Theory: A Primer*. https://doi.org/10.1002/9781118445112.stat00068

Shermis, M. D., & Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (M. D. Shermis & J. Burstein, Eds.). Routledge.

Williamson, D., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

A Appendices

Category	Model	Hidden Size	# Layers	# Heads	Approx. Parameters	Max Seq. Length
Small	ELECTRA Small (Discriminator)	256	12	4	14 M	512
	DistilBERT-base (uncased)	768	6	12	66 M	512
	DeBERTa-v3-small	768	6	12	86 M	512
	GPT-2 (small)	768	12	12	124 M	1024
	MiniLM-L6-uncased	384	6	12	22 M	512
	MobileBERT-uncased	512	24	4	25 M	512
Medium	BERT-base-uncased	768	12	12	110 M	512
	RoBERTa-base	768	12	12	125 M	512
	Longformer-base-4096	768	12	12	149 M	4 096
	GPT-2 (medium)	1 024	24	16	355 M	1024
Large	BERT-large-uncased	1 024	24	16	340 M	512
	RoBERTa-large	1 024	24	16	355 M	512
	GPT-2 (large)	1 280	36	20	774 M	1024
	DeBERTa-v3-large	1 024	24	16	304 M	512

Table 1 General Information Comparison of All (14) Encoder Models used

Bucket	N	Mean	Range	SD
Low	451	1.131	0.909~1.237	0.048
2	479	1.176	1.100~1.276	0.040
3	457	1.216	1.125~1.331	0.038

Table 2. Entropy Bucketing Result using the Equal Variance Method

Prompt	Entropy Bucket		
	L	M	H
Car-free cities	12	32	152
Distance learning	95	80	43
Does the electoral college work?	90	78	37
Driverless cars	50	81	58
Exploring Venus	38	78	70
Facial action coding system	62	71	84
Summer projects	104	59	13
Total	451	479	457

Table 3 Essay Distribution in Entropy Bucket by Prompt

Bucket	σ^2_E	σ^2_S	σ^2_F	σ^2_T	σ^2_{ExS}	σ^2_{ExF}	σ^2_{SXF}	$\sigma^2_{Residual}$
Low	0.00479	0	1.34×10^{-7}	1.231	0	0	0	0.163
Mid	0.00019	1.7×10^{-10}	0	1.274	0	0	0	0.157
High	0.00051	5.4×10^{-6}	0	0.954	0	0	3.1×10^{-9}	0.164

Table 4 Variance Components and Ceiling-G Coefficients by Entropy Bucket

# Encoders	# Seeds	Mean G
2	3	0.872
3	2	0.860
4	1	0.851

Table 5. Minimal small-only configurations achieving G ≥ 0.85

Seeds	Encoders	Encoder Set	G-Coefficient
1	3	DistilBERT-base , GPT-2 (small) , MobileBERT-uncased	0.866
	4	DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.893
	5	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.909
	6	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MiniLM-L6-uncased, MobileBERT-uncased	0.920
2	2	DistilBERT-base , MobileBERT-uncased	0.893
	3	DistilBERT-base , GPT-2 (small) , MobileBERT-uncased	0.922
	4	DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.937
	5	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.946
	6	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MiniLM-L6-uncased, MobileBERT-uncased	0.951
3	2	GPT-2 (small) , MobileBERT-uncased	0.921
	3	DistilBERT-base , GPT-2 (small) , MobileBERT-uncased	0.942
	4	DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.952
	5	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MobileBERT-uncased	0.958
	6	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small) , MiniLM-L6-uncased, MobileBERT-uncased	0.962

Table 6. Best G ⩾ 0.85 Designs By Seed And Encoder Count in Small-encoder Ensemble

N of Encoders	Small Encoders	Seeds	G
1	DistilBERT-base	3	0.872
2	GPT-2 small; MobileBERT-uncased	3	0.921
3	DistilBERT-base; GPT-2 small; MobileBERT-uncased	3	0.942
4	DistilBERT-base; DeBERTa-V3 small; GPT-2 small; MobileBERT-uncased	3	0.952
5	ELECTRA Small (Discriminator) ; DistilBERT-base; DeBERTa-V3 small; GPT-2 small; MobileBERT-uncased	3	0.958
6	ELECTRA Small (Discriminator) ; DistilBERT-base; DeBERTa-V3 small; GPT-2 small; MiniLM-L6-uncased; MobileBERT-uncased	3	0.962

Table 7 Best Small-encoder Ensemble Designs by Number of Models.

Medium Encoder	# of M Encoders	Small Encoder	# of S Encoders	# of seeds	G
BERT-base-uncased	1	DistilBERT-base (uncased)	1	2	0.895
BERT-base-uncased	1	DistilBERT-base (uncased)	1	2	0.895
BERT-base-uncased	1	DistilBERT-base (uncased)	1	2	0.895
BERT-base-uncased	1	GPT-2 (small)	1	2	0.894
BERT-base-uncased	1	GPT-2 (small)	1	2	0.894

Table 8. Top 5 G ≥ 0.85 Mixed Ensemble Designs with Fewest Resources

Medium Encoder	# of M Encoders	Small Encoder	# of S Encoders	# of seeds	G
BERT-base-uncased , RoBERTa-base , Longformer-base-4096 , GPT-2 (medium)	4	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.968
BERT-base-uncased , RoBERTa-base , Longformer-base-4096	3	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.968
BERT-base-uncased , RoBERTa-base , GPT-2 (medium)	3	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.968
BERT-base-uncased , Longformer-base-4096 , GPT-2 (medium)	3	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.967
BERT-base-uncased , RoBERTa-base	2	ELECTRA Small (Discriminator) , DistilBERT-base , DeBERTa-v3-small , GPT-2 (small), MiniLM-L6-uncased , MobileBERT-uncased	6	3	0.967

Table 9. Mixed Designs with Top 5 G-coefficients

Category	Model	Train Embedding	Test Embedding	Full 3×5 CV & Merge
Small	ELECTRA Small (Discriminator)	0 h 9 m 29 s	0 h 1 m 3 s	0 h 12 m 45 s
	DistilBERT-base	0 h 10 m 18 s	0 h 1 m 10 s	0 h 17 m 42 s
	DeBERTa-v3-small	0 h 31 m 53 s	0 h 3 m 30 s	0 h 41 m 14 s
	GPT-2 (small)	0 h 41 m 54 s	0 h 4 m 31 s	0 h 52 m 46 s
	MiniLM-L6-uncased	0 h 4 m 50 s	0 h 0 m 32 s	0 h 8 m 49 s
	MobileBERT-uncased	0 h 21 m 12 s	0 h 2 m 15 s	0 h 28 m 10 s
Medium	BERT-base-uncased	0 h 21 m 6 s	0 h 2 m 27 s	0 h 29 m 48 s
	RoBERTa-base	0 h 21 m 18 s	0 h 2 m 25 s	0 h 30 m 8 s
	Longformer-base-4096	5 h 40 m 48 s	0 h 38 m 3 s	6 h 24 m 32 s
	GPT-2 (medium)	2 h 2 m 50 s	0 h 13 m 14 s	2 h 24 m 24 s
Large	BERT-large-uncased	1 h 4 m 41 s	0 h 7 m 17 s	1 h 20 m 10 s
	RoBERTa-large	1 h 10 m 46 s	0 h 7 m 58 s	1 h 27 m 23 s
	GPT-2 (large)	4 h 29 m 16 s	0 h 31 m 11 s	5 h 11 m 33 s
	DeBERTa-v3-large	2 h 58 m 33 s	0 h 20 m 48 s	3 h 28 m 12 s

Table 10. Wall-Clock Times for Embedding and 3×5 Cross-Validation Scoring of 14 Transformer Encoders

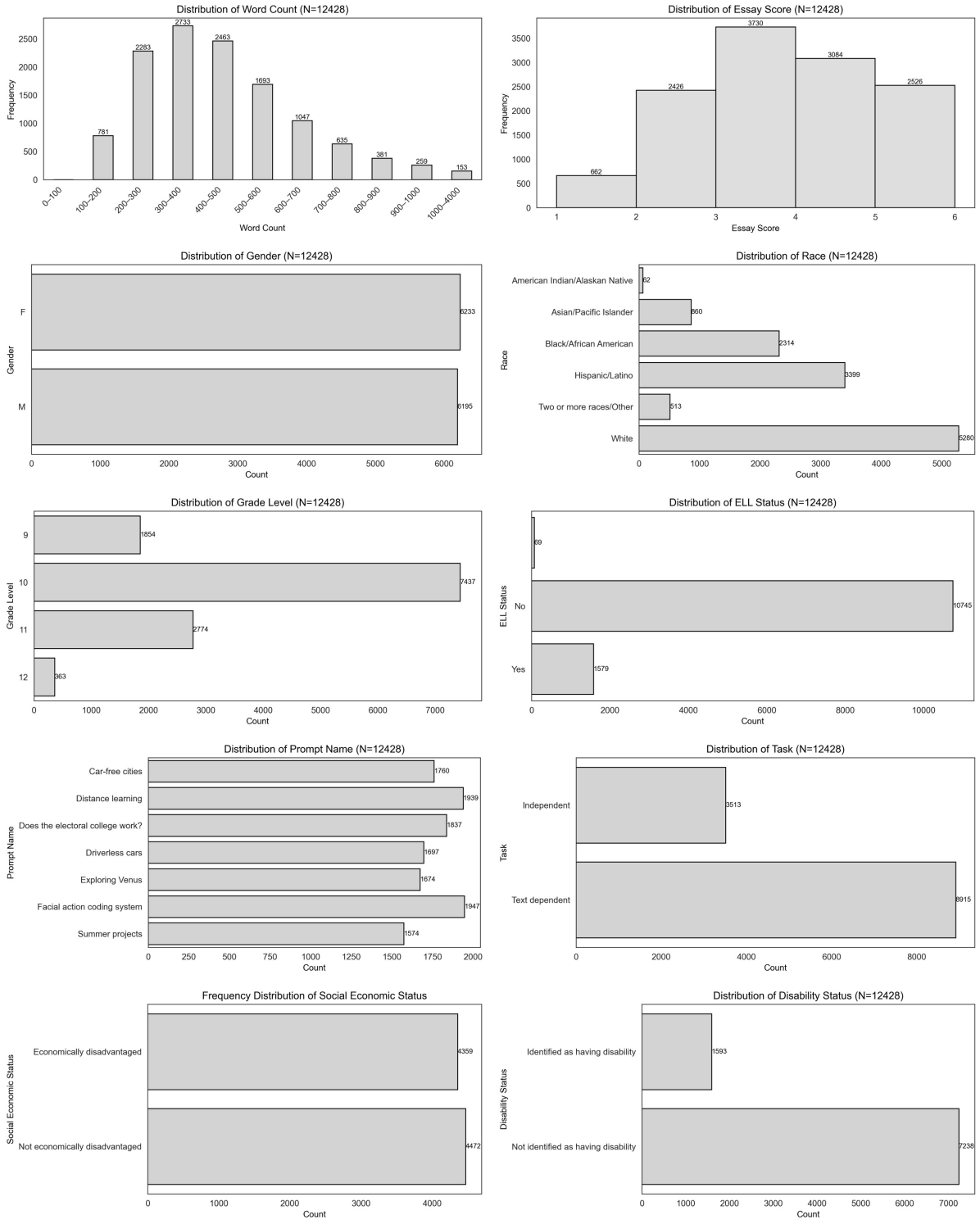

Figure 1. Descriptive Statistics of Essays in Training Set and the Test-takers' Demographic Information

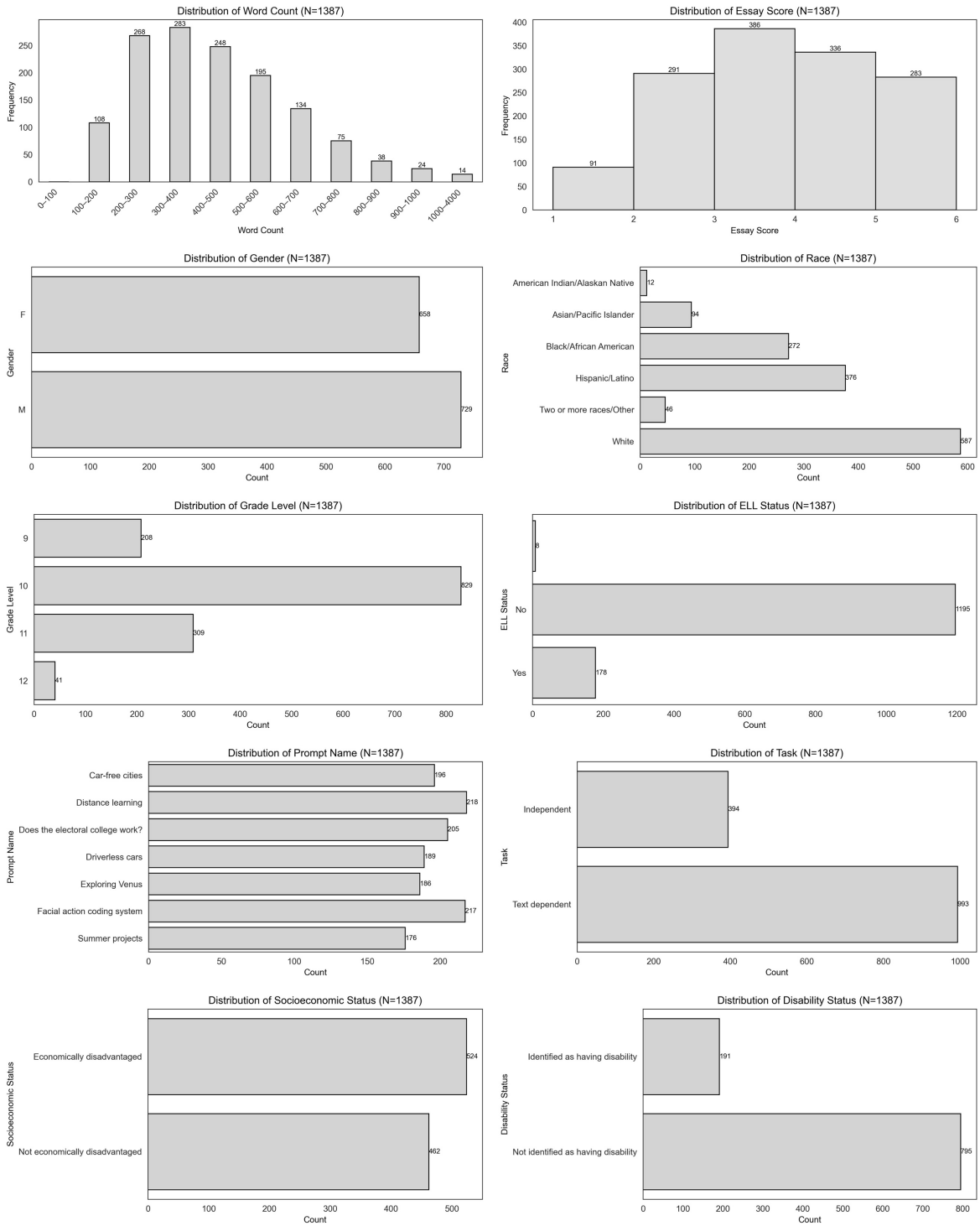

Figure 2. Descriptive Statistics of Essays in “Holdout_evaluation_set” and the Test-takers’ Demographic Information

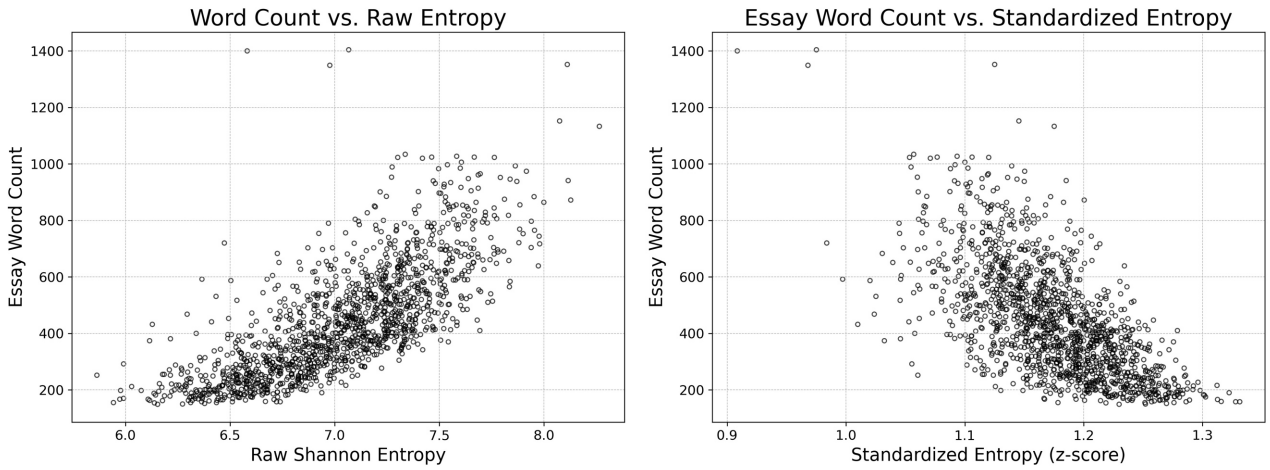

Figure 3. Comparison of Scatterplots of Word Count vs. Raw Shannon Entropy and Standardized Entropy

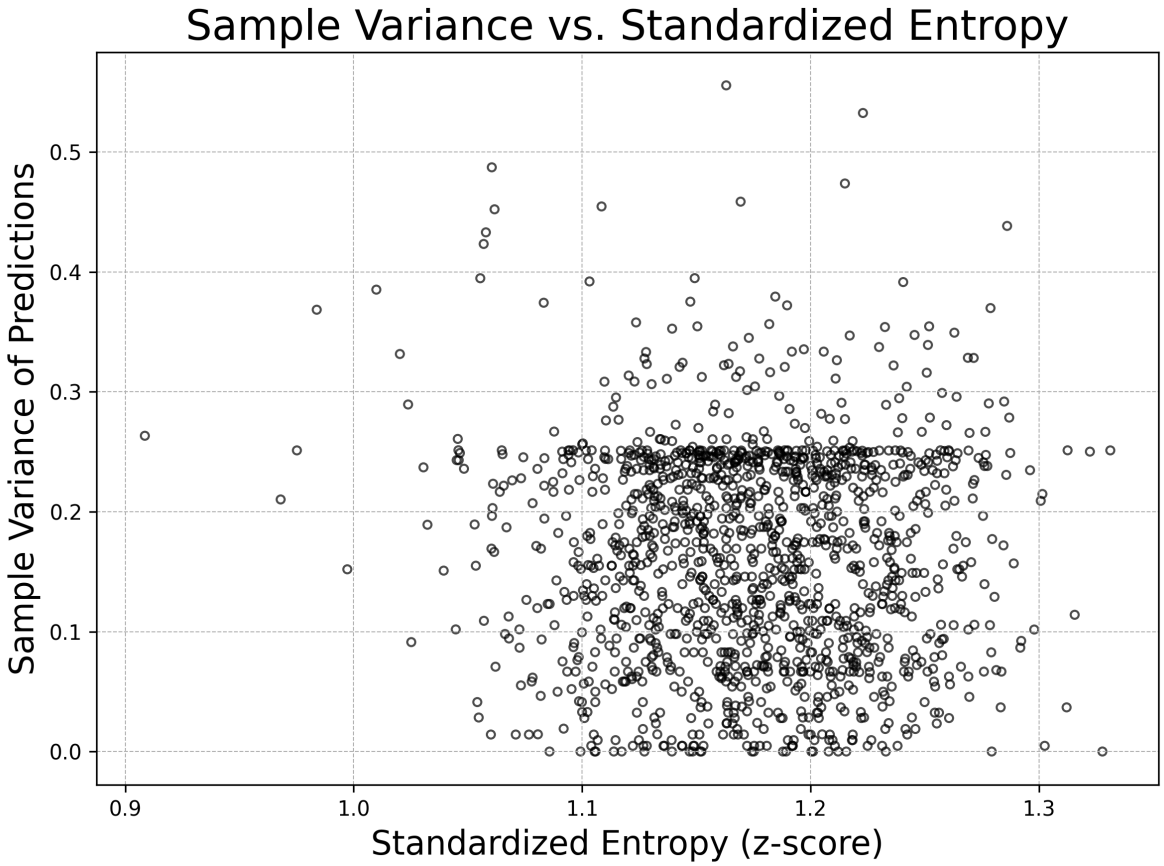

Figure 4. Scatterplot of Score Sample Variance and Standardized Entropy

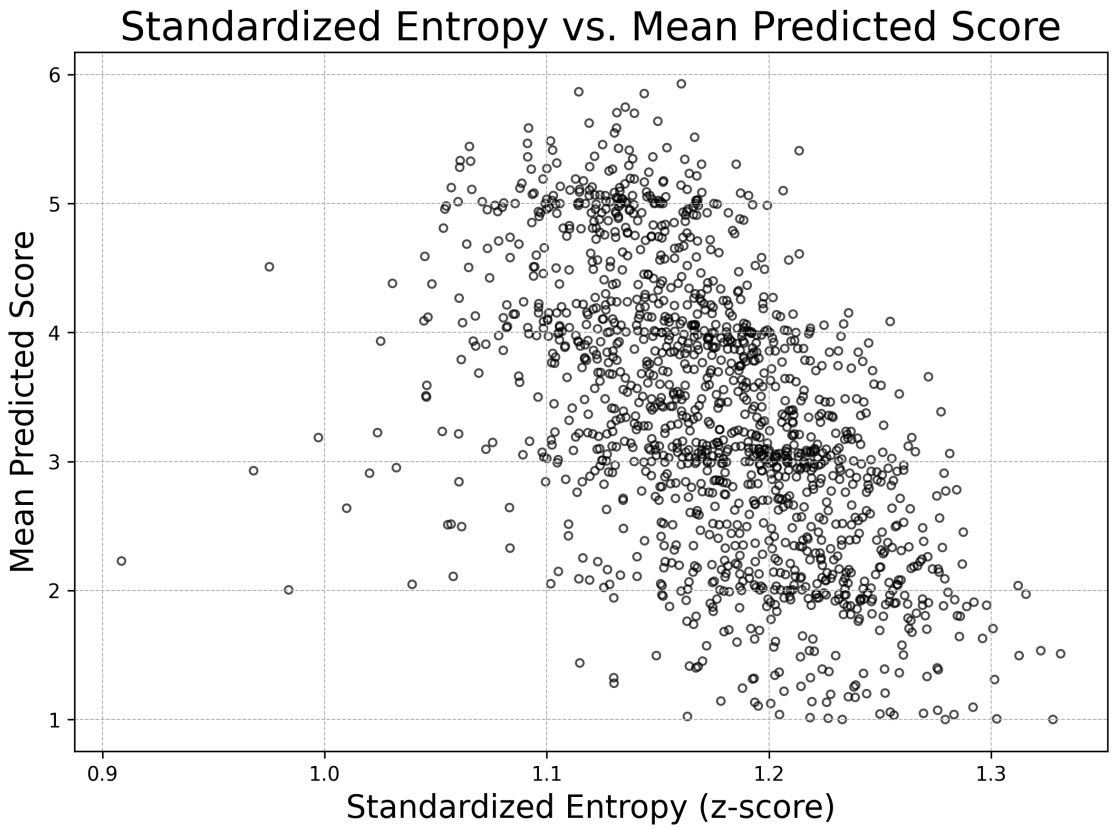

Figure 5. Scatterplot of Standardized Entropy vs. Mean Predicted Scores

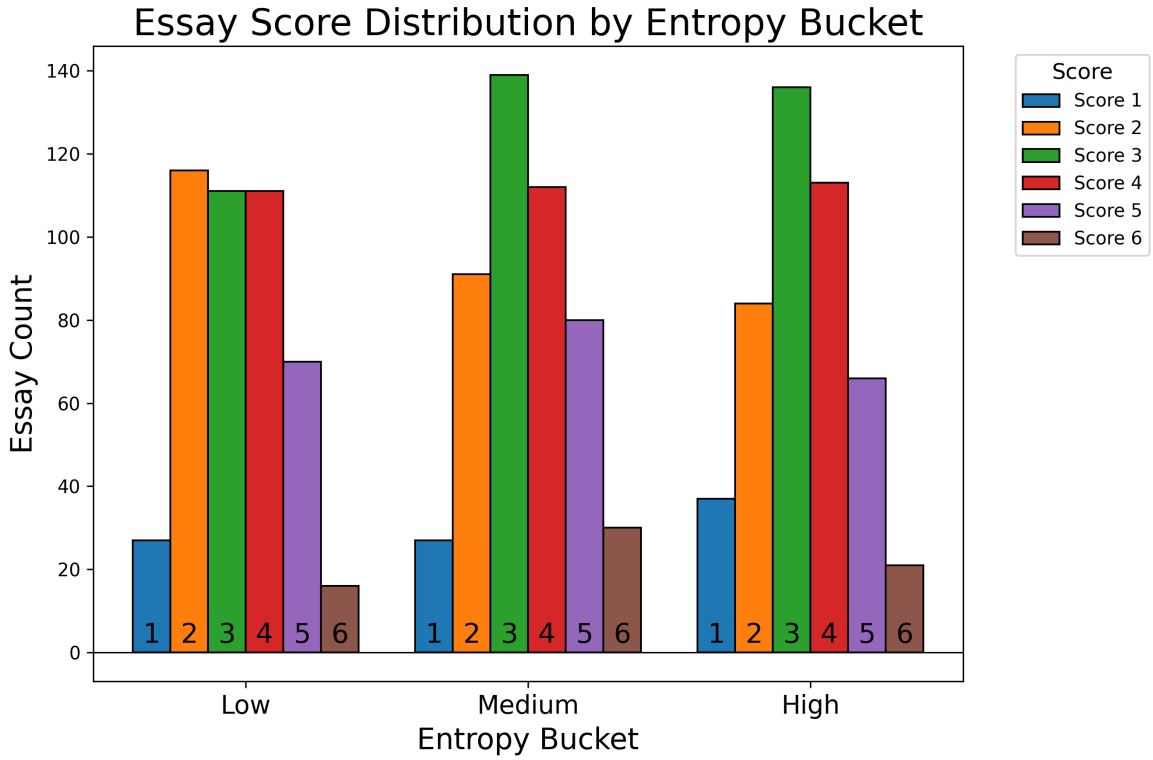

Figure 6. Histogram of Essay Score Distribution by Entropy Bucket

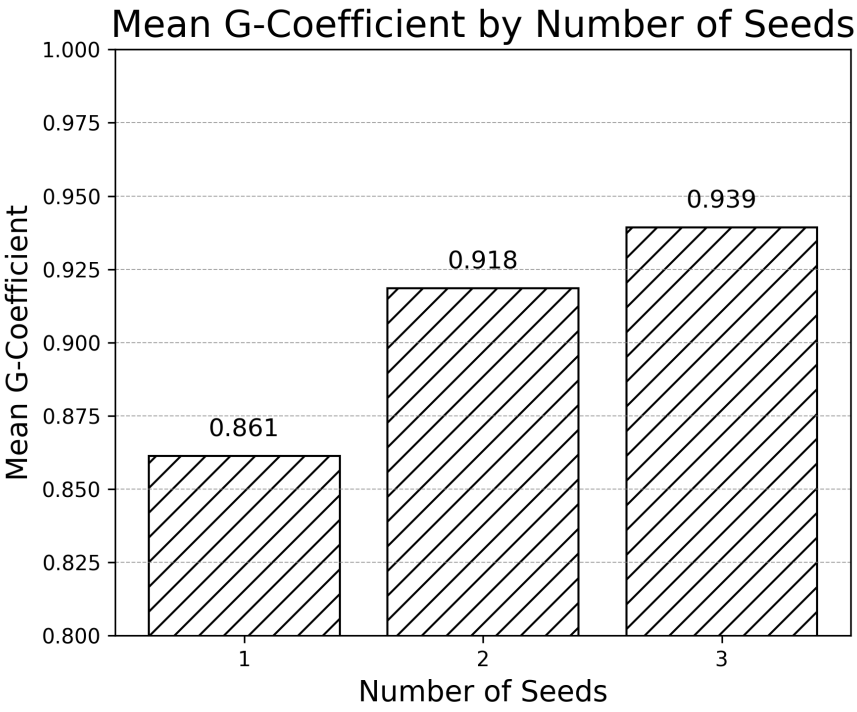

Figure 7. Mean G-coefficients by Number of Seeds in Small-encoder Ensemble Designs

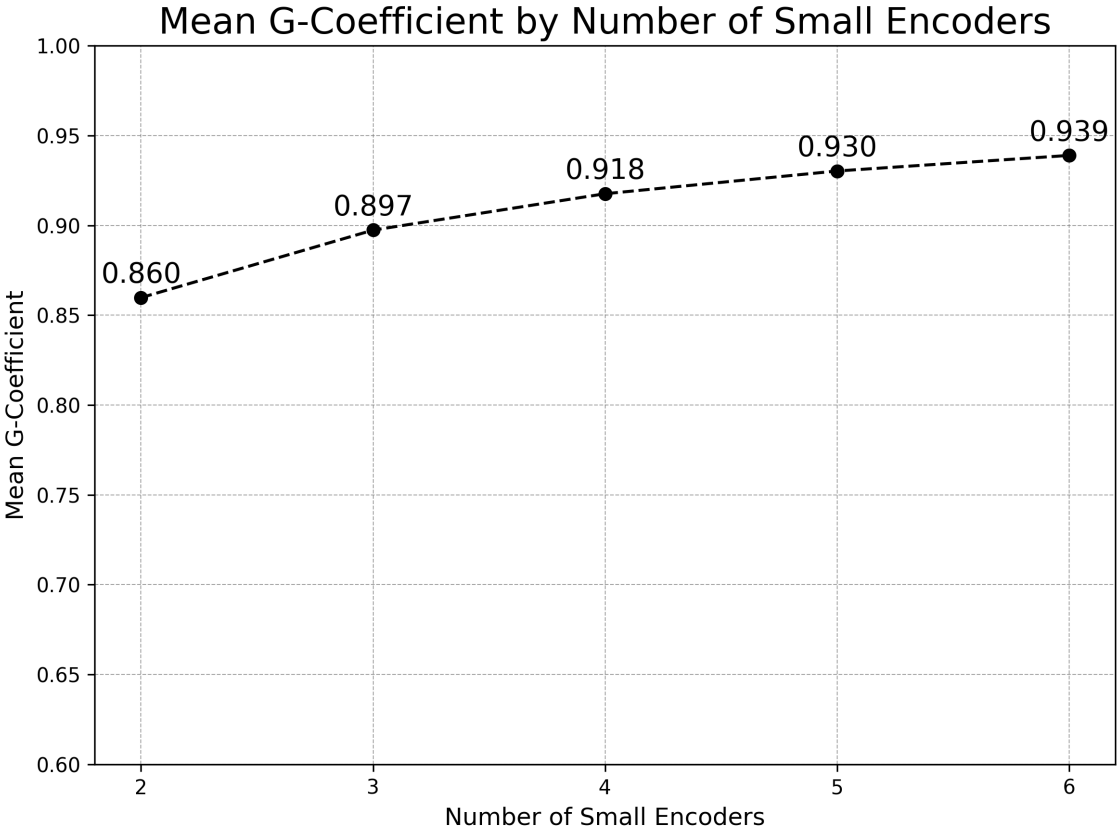

Figure 8. Mean G-coefficients by Number of Small-encoders in Small-encoder Ensemble Designs

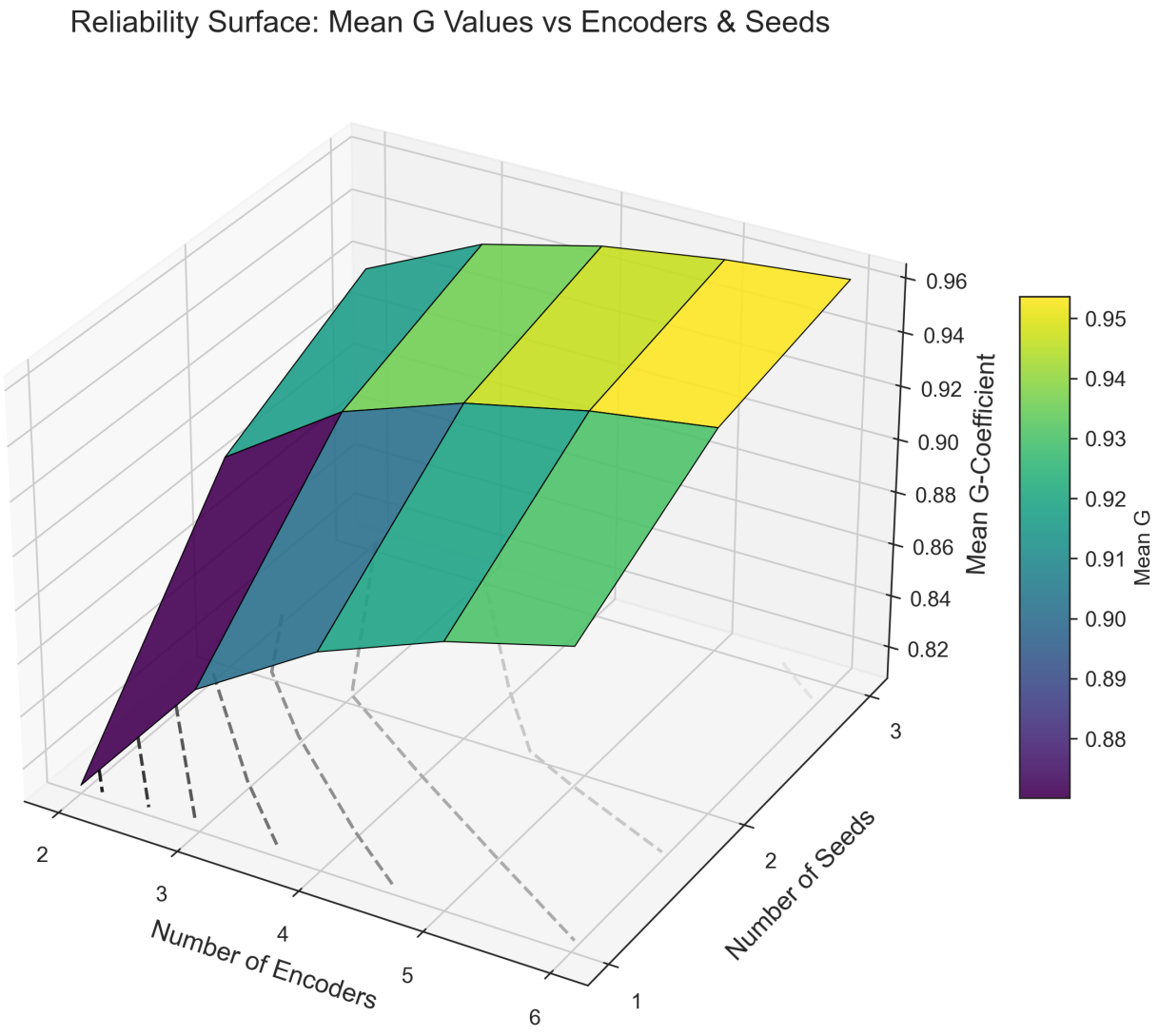

Figure 9. 3-D Plot of Mean C-cefficients vs. Number of Encoders and Seeds