The Selector Remains Human

A Negative Result on Compression-Based Value Signals for Creative Text

Eduardo Cortes Working manuscript · July 2026

Abstract

We ask whether compression-based readouts can provide a mechanical proxy for human creative appraisal. Across a human-rated poetry corpus and LitBench, the strongest label-free versions fail. Compression progress against generic literary domains does not show resolved positive alignment with poetic ratings and can anti-align under matched controls; unconditional language-model surprise likewise does not explain appraisal beyond simple surface features. On a favorable repeated-prompt LitBench subset, however, preference-labeled responses to the same prompt form a strong transductive class boundary. A conditional language-model readout reaches 79.7% pairwise accuracy, but an operator-matched directional MiniLM probe reaches 87.7% on the identical 1,155 rows, showing that the effect is not compression-specific and that aggregation choice is decisive. Centroid pooling collapses to 48.9%. Random same-prompt domains also collapse to chance, while preference-labeled domains drawn from different prompts reach only about 53%, whether sourced from train or test, and add no resolved value beyond a 60.8% surface-format baseline. The positive effect is therefore prompt-local and label-defined, not a transferable label-free value function. We conclude that the selector remains human: mechanical readouts can amortize human-constructed comparison domains, but they do not replace the human source of value.

Keywords: creative writing evaluation; compression progress; preference learning; sentence embeddings; transductive inference; human judgment

1. Introduction

Creative systems have a generator problem and a selector problem. Modern language models can generate a large and varied candidate set, but open-ended creative work has no single ground-truth target and no generally accepted automatic verifier. The selector must distinguish work that is merely fluent or typical from work that a reader values. Because people often find recognition easier than explicit specification, it is tempting to search for a mechanical signal that could stand in for that selector.

Compression progress offers one such hypothesis. If a candidate helps a bounded observer predict or compress a relevant domain, then the candidate may have exposed a useful regularity rather than merely repeating the observer's prior. This idea has been proposed as a computational account of subjective beauty, novelty, surprise, and interestingness [1]. Applied to creative text, it suggests that an artifact may be valuable when conditioning on it improves compression of a held-out literary domain.

The central ambiguity is the domain itself. A readout over a generic literary corpus asks whether a candidate makes conventional literature easier to predict. A readout over a human-preference-shaped corpus asks whether a candidate resembles a contrast already encoded by human judgments. These are not equivalent claims. The first would be evidence for a label-free selector. The second may be useful, but it is supervised by construction.

We study this distinction in two stages. First, we evaluate compression-based readouts on a small corpus of human-rated poems. This study tests the original generic-domain hypothesis and establishes a boundary between human appraisal and raw language-model surprise. Second, we use LitBench [2], a large paired-preference benchmark for creative stories, to identify the mechanism behind an initially strong same-prompt result. A sequence of controls separates representation, aggregation operator, prompt locality, and label construction.

Our contributions are:

- We show that compression progress against generic literary domains does not recover human poetic appraisal and can resolve in the opposite direction under matched controls.
- We show that unconditional language-model surprise is not equivalent to human-rated surprise or creativity, and adds little beyond surface-format features in LitBench.
- We identify a strong LitBench signal under transductive, same-prompt, preference-labeled domains.
- We show that the signal is not compression-specific: a directional per-instance MiniLM operator exceeds the conditional language-model readout on identical rows, while centroid pooling erases the effect.
- We show that the boundary is jointly prompt-local and label-defined. Same-prompt random domains, cross-prompt train-label domains, and cross-prompt test-label domains all collapse near chance or add no resolved value beyond surface formatting.
- We conclude with a negative-with-relocation result: no transferable label-free selector is established. Human judgment remains load-bearing and is relocated into constructing the comparison domain D.

2. Related Work

Compression-progress accounts treat learning progress, rather than static surprise alone, as a source of subjective interestingness [1]. The distinction matters because high entropy can be random and uninformative, whereas improved compression indicates newly captured structure. Our experiments operationalize a text-domain version of this idea and test whether it aligns with human appraisal.

Automatic text evaluation has increasingly moved from lexical overlap to learned

representations. Sentence-BERT enables efficient semantic comparison through cosine similarity in a sentence-embedding space [3], and BERTScore demonstrates the broader utility of directional, per-instance contextual matching for text generation evaluation [4]. Our operator controls are conceptually related: the decisive comparison is not a single centroid but a candidate's relation to individual members of preferred and rejected pools.

Pairwise preference data are commonly modeled with Bradley-Terry-style objectives [5], and human comparisons have become a central source of reward signals in machine learning [6]. LitBench extends this paradigm to creative writing, providing 43,827 training pairs and a held-out test set of 2,480 debiased human-labeled comparisons [2]. Fein et al. report 78% inductive accuracy for trained reward models and 73% for their strongest off-the-shelf judge. Our 87.7% result is not a competing reward-model number: it is a transductive same-prompt probe whose domains contain labels from other test items.

Recent work also emphasizes the difficulty and subjectivity of poetry and creative-writing evaluation. Chaudhuri et al. collected ratings of 36 poems across clarity, aesthetic appeal, valence, arousal, surprise, and creativity, finding substantial reader-level variation [7]. Porter and Machery found that non-expert readers often preferred and misidentified AI-generated poetry, illustrating how accessibility and conventionality can shape judgments [8]. These findings motivate explicit controls for surface form and caution against equating model predictability with artistic value.

3. Formal Setup

Let a candidate artifact be a, a prior context be H, and a held-out domain be D. The motivating compression-progress functional is:

$$V(a \mid H, D) = \sum_i q_i [L(D \mid O_i \oplus H) - L(D \mid O_i \oplus H \oplus a)] - \text{cost}(a).$$

where L is a description length or negative log-likelihood and q_i weights possible observations. The crucial property is that the candidate is evaluated by its effect on D, not by the likelihood of the candidate itself.

For contrastive domains D+ and D-, we define a candidate specificity score:

$$S(a) = s(a, D^+) - s(a, D^-).$$

where s(a,D) is a candidate-to-domain readout. For a chosen/rejected pair (c+, c-), the pairwise contrast is:

$$\Delta = S(c^+) - S(c^-).$$

and the sign rule predicts the human-chosen item when $\Delta > 0$.

We instantiate s in two ways. Conditional cross-predictability uses a language model to measure how well a candidate predicts the stories in a pool. Directional embedding similarity computes cosine similarity between the candidate embedding and every pool item, then averages the top k values. The centroid control instead compares the candidate with the mean pool embedding. This distinction turns out to be load-bearing.

4. Study 1: Human-Rated Poetry

4.1 Data. We use the 36 English poems released by Chaudhuri et al. [7]. The source study selected 18 low-surprise and 18 high-surprise poems from an initial set of 108. Participants rated each poem on seven-point scales for clarity, aesthetic appeal, felt valence, felt arousal, surprise, and overall creativity. We aggregate the public ratings at the poem level. Because the inferential unit is the poem, the effective sample size for item-level claims is n=36, not the number of individual ratings.

4.2 Domains and readouts. Generic domains are sampled from Gutenberg-derived poetry-like text, with accessible and formal variants. Preference-shaped domains are constructed from high- and low-rated poems inside held-out folds. Compression readouts use DistilGPT-2, GPT-2, and GPT-2-medium observers. TF-IDF and sentence-embedding similarities provide non-compression controls. Matched-other and word-shuffle controls estimate whether apparent gains reflect generic length, form, or variance normalization rather than target-specific appraisal structure.

4.3 Statistical treatment. We report poem-level Spearman associations and nonparametric bootstrap intervals over poems. K-fold and seed repetitions reuse the same 36 items and are therefore treated as stability diagnostics, not independent inferential samples. For comparisons between human-shaped and control domains, paired bootstrap intervals are the primary uncertainty estimate.

4.4 Generic domains fail. Compression progress against generic Gutenberg domains does not show a stable positive relationship with human ratings. In the formal Gutenberg matched-other condition, the structural association resolves negative (ρ=-0.443, 95% CI [-0.672, -0.157]). Other generic-domain conditions are near zero and unresolved; for example, accessible Gutenberg with a word-shuffle control yields ρ=0.076, 95% CI [-0.283, 0.417]. The strongest label-free interpretation is therefore rejected in this setting.

4.5 Human-shaped domains are positive but not isolated from contrastive construction. Preference-shaped domains produce positive point estimates across compression, TF-IDF, and embedding readouts. However, same-form comparisons against surface-matched contrastive pools remain unresolved at n=36. For DistilGPT-2, the human-shaped correlation is 0.518 and the surface-pool correlation is 0.481; their paired difference is only 0.037, 95% CI [-0.179, 0.254].

GPT-2 and GPT-2-medium show similarly small unresolved differences. Thus the poetry study supports domain-relative supervised probing but cannot isolate a uniquely semantic or compression-specific mechanism.

4.6 Appraisal is not language-model surprise. Higher-rated poems are often more predictable under unconditional language models. Associations between average token NLL and aesthetic appeal, creativity, and felt valence are generally negative; human-rated surprise is also not a monotonic proxy for model confusion. This is a boundary result: artistic surprise is not reducible to low probability under a generic language model.

Table 1. Selected poetry controls

Analysis	Estimate	95% bootstrap interval	Interpretation
Formal Gutenberg, matched-other structural association	-0.443	[-0.672, -0.157]	Resolved anti-alignment
Accessible Gutenberg, word-shuffle structural association	+0.076	[-0.283, +0.417]	Unresolved
Human-shaped vs surface-pool same-form contrast, DistilGPT-2	+0.037	[-0.179, +0.254]	No isolated human-domain advantage at n=36

5. Study 2: LitBench Mechanism Tests

5.1 Dataset. LitBench contains 43,827 training pairs and a held-out test set of 2,480 human-labeled story comparisons [2]. Each pair contains a prompt, a chosen story, and a rejected story. The original benchmark was designed for inductive evaluation of judges and reward models. Our main same-prompt analysis is different: it is transductive because other labeled test responses are used to construct comparison domains.

5.2 Surface baseline and raw NLL. We compute pairwise differences in character count, word count, type-token ratio, mean word length, punctuation count, newline count, and paragraph count. A five-fold cross-validated logistic model reaches 60.78% accuracy on the exact repeated-prompt overlap (n=1,155; 95% CI [57.92%, 63.55%]) and 60.16% on the full test set. This is a substantial

confound. A raw average-NLL rule is much weaker: preferring the lower-NLL story reaches approximately 54.3%, and adding NLL to surface features does not provide a resolved increment over surface alone.

5.3 Same-prompt transductive domains. We retain test pairs whose prompts have enough other labeled responses to build both preferred and rejected pools, using a minimum domain size of two and a maximum of ten. For each target pair, the pair itself is excluded from its domains. D+ contains other chosen stories for the same prompt and D- contains other rejected stories. This produces an exact-overlap analysis of 1,155 pairs.

5.4 Conditional cross-predictability. For each candidate, we measure its relative ability to predict D+ versus D- with DistilGPT-2. The sign rule reaches 79.74% accuracy (95% CI [77.40%, 82.08%]); a logistic readout reaches 79.57% (95% CI [77.23%, 81.90%]). The mean continuous pair contrast is +0.4287. Taken alone, this looks like a strong compression-based selector.

5.5 Random same-prompt domains. To test whether repeated prompts alone create the effect, we randomly split same-prompt stories into two domains while preserving the prompt ecology. Accuracy collapses to 48.92% (n=1,161; 95% CI [46.08%, 51.85%]) and the mean continuous contrast is effectively zero. Prompt matching alone is not sufficient; the preference labels are load-bearing.

5.6 Centroid controls. We first replaced the language-model readout with TF-IDF and MiniLM centroid cosine. Both fail. On the exact overlap, MiniLM centroid logistic reaches 48.92% (95% CI [46.15%, 51.77%]) and adds no value beyond surface formatting. This null initially appeared to support compression-specificity, but it changed both the representation and the aggregation operator.

5.7 Crossed operator control. We therefore retain MiniLM embeddings but replace centroid pooling with directional per-instance comparisons: mean, maximum, and top-k cosine across pool items. The top-3 sign rule reaches 87.71% on the identical 1,155 rows (95% CI [85.80%, 89.61%]); adding surface features reaches 88.66% (95% CI [86.84%, 90.39%]). The top-2, top-5, and mean operators are similarly strong. This result ends compression-specificity: the language model was one kernel for reading the same labeled boundary, and not the best one.

5.8 Cross-prompt train-label domains. We next construct D+ and D- from LitBench training labels. Test prompts retrieve nearby training prompts using MiniLM prompt embeddings; the nearest stories populate preferred and rejected pools. On the full test set, the best domain-only result is 53.55% and adds no resolved value beyond surface. On the exact 1,155-row overlap, the best domain-only result is 53.07% (95% CI [50.22%, 55.93%]); top-3 plus surface reaches 61.30%, only +0.52 percentage points over surface with a paired 95% CI of [-0.87, 1.90].

5.9 Cross-prompt test-label domains. The train-domain experiment changed two factors at once: label source and prompt relation. To isolate prompt locality, we

retrieve domains from other test prompts while preserving test chosen/rejected labels and excluding the literal same prompt. This condition also collapses. The best domain-only result is 52.99% (95% CI [50.13%, 55.85%]); the best surface-plus-domain result is 61.39%, with no resolved gain over the 60.78% surface baseline.

5.10 Statistical treatment. Accuracy intervals use 5,000 item-level bootstrap samples. Surface and learned single-feature readouts use five-fold cross-validation on pairwise feature differences. Paired deltas are bootstrapped over item-level correctness differences. The sign-rule results require no fitted classifier and are emphasized when they agree with logistic readouts.

6. Results Synthesis

Figure 1A isolates the operator confound. Conditional cross-predictability is strong, but a directional MiniLM operator is stronger, while MiniLM centroid pooling is at chance. The representation family is therefore not the mechanism. The candidate-to-pool operator and the labeled pool construction are.

Figure 1B isolates the domain ecology. Same-prompt preference labels produce the strong transductive boundary. Same-prompt random domains collapse, showing that prompt match alone is insufficient. Cross-prompt domains also collapse whether their labels come from train or test, showing that preference labels alone are insufficient when the prompt ecology changes.

The most precise mechanism statement is therefore conjunctive: the positive LitBench result is a transductive, same-prompt, preference-labeled class-structure effect. Neither component alone reproduces it.

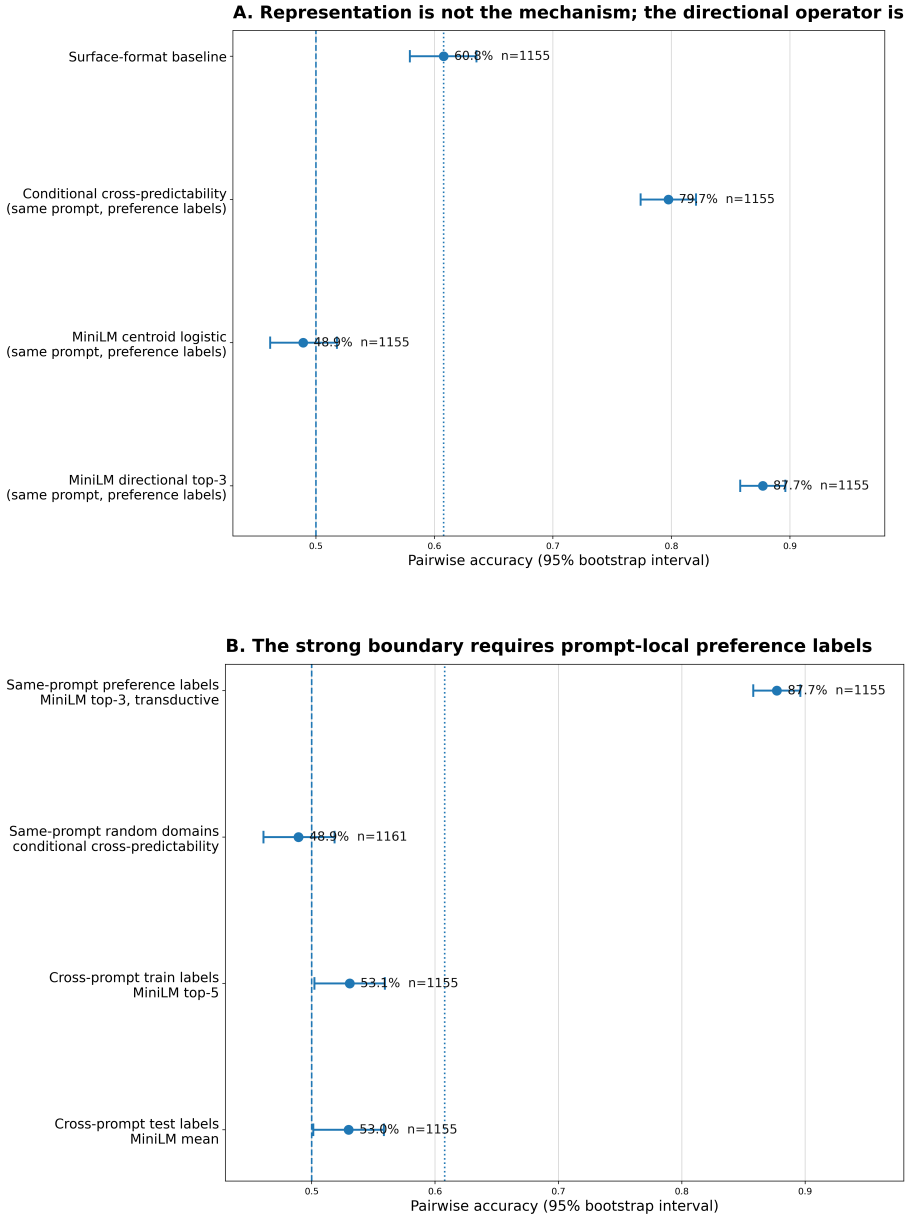

Figure 1. (A) On the exact repeated-prompt overlap, conditional cross-predictability is strong, centroid MiniLM is at chance, and directional top-3 MiniLM is stronger. (B) The strong boundary disappears when labels are randomized within prompt or when preference-labeled domains are drawn from different prompts. The vertical dotted reference is the surface-format baseline; the dashed reference is chance.

Table 2. Canonical LitBench mechanism table

Condition	Domain source	Prompt relation	Label source	Operator	n	Accuracy (95% CI)	Status
Surface format	None	N/A	None	5-fold logistic	1,155	60.78 [57.92, 63.55]	Feature baseline
Conditional cross-predictability	Other test responses	Same prompt	Test preferences	LM sign rule	1,155	79.74 [77.40, 82.08]	Transductive
MiniLM centroid	Other test responses	Same prompt	Test preferences	Centroid logistic	1,155	48.92 [46.15, 51.77]	Transductive control
MiniLM directional top-3	Other test responses	Same prompt	Test preferences	Top-3 sign rule	1,155	87.71 [85.80, 89.61]	Transductive probe
Random same-prompt domains	Other test responses	Same prompt	Random split	LM sign rule	1,161	48.92 [46.08, 51.85]	Prompt-only control
Train-domain retrieval	Training responses	Different/similar prompts	Train preferences	MiniLM top-5	1,155	53.07 [50.22, 55.93]	Inductive cross-prompt
Cross-prompt test domains	Other test responses	Different/similar prompts	Test preferences	MiniLM mean	1,155	52.99 [50.13, 55.85]	Transductive cross-prompt

Rows use the strongest or canonical domain-only operator for each pre-registered control. The random-domain row has a slightly different n and is not row-identical. The 87.71% row is not comparable to an inductive reward-model score because its domains are constructed from other labeled test items.

7. Discussion

7.1 The label-free hypothesis failed. The poetry study blocks the strongest version of the project. A generic literary domain does not provide a reliable mechanical aesthetic selector, and can reward conventionality in the opposite direction from contemporary human appraisal. Raw language-model surprise also fails as a substitute for human surprise.

7.2 Compression was interchangeable. The LitBench operator control is a direct self-correction. Once the embedding baseline uses the same directional per-instance structure, MiniLM exceeds conditional cross-predictability. The earlier centroid null was an aggregation artifact. Compression may remain a useful kernel in some settings, but no compression-specific creative-value mechanism is established.

7.3 The strong result was transductive class probing. The 87.7% number should not be interpreted as an inductive creative-writing evaluator. It classifies test pairs against domains built from other test labels in the same prompt ecology. This is legitimate evidence that the labels induce a strong local class boundary, but it is not comparable to the 78% inductive reward-model result reported by LitBench [2].

7.4 The boundary was prompt-local and label-defined. The control ladder identifies both necessary ingredients. Random labels within the same prompt eliminate the effect. Preference labels from different prompts also eliminate it, even when those labels come from the test set. The result is not simply leakage in the colloquial sense, nor simply topic matching. It is a prompt-local label ecology.

7.5 Negative with relocation. The investigation does leave a constructive engineering result. Human judgment can be represented through contrastive domains, and directional readouts can cheaply extrapolate within the local boundary those domains define. But the source of value has not disappeared. It sits in the human construction and continual revision of D. Software can preserve, retrieve, audit, and amortize that judgment; it does not manufacture a transferable aesthetic value function from generic text statistics.

7.6 Implications for creative systems. A defensible system should treat preferred/rejected exemplars as an explicit human interface rather than hiding them behind a universal score. It should maintain context-specific pools, expose surface confounds, and route candidates that are dissimilar to both pools back to a human rather than automatically rejecting them. These are design implications, not claims that the specific MiniLM operator is a production evaluator.

8. Limitations

The strongest LitBench result is restricted to a favorable high-density repeated-prompt subset. It demonstrates local class structure, not performance over the full distribution.

The same-prompt result is transductive. Other test labels enter the domain construction, so the accuracy cannot be compared directly with inductive reward models or zero-shot judges.

Cross-prompt retrieval uses MiniLM prompt similarity and a fixed neighborhood/domain-size design. The collapse bounds this tested transfer mechanism but does not exhaust every possible alignment, hierarchical domain construction, or learned retrieval method.

The poetry study has only 36 item-level observations. Its resolved generic-domain anti-alignment is informative, but fine-grained comparisons among compression, TF-IDF, embeddings, and surface-matched contrastive pools remain underpowered.

The MiniLM encoder may truncate long stories and is only one embedding family. The crossed-operator result establishes that compression is not uniquely required; it does not identify a universally optimal representation.

Creative appraisal is context-dependent and heterogeneous across readers. Aggregated ratings and pairwise choices compress that heterogeneity into a single target. Personalized or multi-dimensional preference structures may behave differently.

We do not claim that no mechanical aesthetic signal can ever exist. We reject the tested generic compression, raw-surprise, and cross-prompt domain-transfer mechanisms as sufficient label-free selectors in these datasets.

9. Conclusion

This project began with a strong hypothesis: compression progress over a held-out literary domain might mechanize creative appraisal. The evidence does not support that claim. Generic domains fail, raw surprise is not human appraisal, and the apparent LitBench rescue depends on a transductive same-prompt preference-label ecology.

The decisive controls also show why. Centroid pooling erases the same-prompt boundary, while a directional MiniLM operator recovers and exceeds the language-model result. Yet the boundary disappears when the labels come from different prompts, whether from train or test. The metric is interchangeable; the domain construction is load-bearing.

The final finding is therefore negative with relocation. No transferable label-free selector was found. The selector remains human, and the engineering leverage lies in helping that human construct, maintain, and apply D.

Data, Code, and Ethics

All analyses use previously released datasets; no new human-subject data were collected. The poetry ratings originate from Chaudhuri et al. [7], and the story

preferences originate from LitBench [2]. Analysis code, generated tables, and result artifacts are maintained in the PoemForge paper repository. The final public release should preserve dataset licenses and citation requirements from the original sources.

References

[1] J. Schmidhuber. Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes. Dagstuhl Seminar Proceedings 9291, 1–35, 2009. doi:10.4230/DagSemProc.09291.14.

[2] D. Fein, S. Russo, V. Xiang, K. Jolly, R. Rafailov, and N. Haber. LitBench: A Benchmark and Dataset for Reliable Evaluation of Creative Writing. Proceedings of EACL 2026, 7740–7755. doi:10.18653/v1/2026.eacl-long.362.

[3] N. Reimers and I. Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP-IJCNLP 2019, 3982–3992. doi:10.18653/v1/D19-1410.

[4] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. ICLR 2020. arXiv:1904.09675.

[5] R. A. Bradley and M. E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika 39(3/4), 324–345, 1952. doi:10.1093/biomet/39.3-4.324.

[6] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep Reinforcement Learning from Human Preferences. NeurIPS 2017. arXiv:1706.03741.

[7] S. Chaudhuri, A. Pickering, M. Dooley, and J. Bhattacharya. Beyond the Words: Exploring Individual Differences in the Evaluation of Poetic Creativity. PLOS ONE 19(10):e0307298, 2024. doi:10.1371/journal.pone.0307298.

[8] B. Porter and E. Machery. AI-Generated Poetry Is Indistinguishable from Human-Written Poetry and Is Rated More Favorably. Scientific Reports 14:26133, 2024. doi:10.1038/s41598-024-76900-1.

[9] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. NeurIPS 2020.

[10] M. Walsh, A. Preus, and M. Antoniak. Sonnet or Not, Bot? Poetry Evaluation for Large Models and Datasets. arXiv:2406.18906, 2024.