MuseRAG: Idea Originality Scoring At Scale

Anonymous ACL submission

Abstract

Assessing the originality of creative ideas often relies on their statistical infrequency within a population—an approach long used in creativity research but difficult to automate at scale. Human annotation via manual bucketing of idea rephrasings is labor-intensive, subjective, and brittle under large corpora. We introduce a fully automated, psychometrically validated pipeline for frequency-based originality scoring. Our method, MuseRAG, combines large language models (LLMs) with an externally orchestrated retrieval-augmented generation (RAG) framework. Given a new idea, the system retrieves semantically similar prior idea buckets and zero-shot prompts the LLM to judge whether the new idea belongs to an existing bucket or forms a new one. The resulting buckets enable computation of frequency-based originality metrics. MuseRAG matches human annotators in both idea clustering (AMI = 0.59) and participant-level originality scores ($r = 0.89$), while exhibiting strong convergent and external validity. Our work enables intent-sensitive, human-aligned originality scoring, aiding creativity research at scale.

1 Introduction

Assessing creativity at scale remains a core challenge in cognitive science and computational linguistics. Two complementary creativity dimensions are of primary interest: the intrinsic qualities of ideas (e.g., creative ideas tend to be semantically 'flexible' or diverse) and statistical infrequency (i.e., 'original' ideas should not appear very often) (Beketayev and Runco, 2016; Runco and Jaeger, 2012). Recent computational advances have enabled scalable evaluations of intrinsic idea qualities via unsupervised, semi-supervised, and supervised scoring methods (Beaty and Johnson, 2021; Organisciak and Dumas, 2020; Organisciak et al., 2023). However, frequency-based originality scoring still relies on manual tabulation of response occurrences (Reiter-Palmon et al., 2019). This process involves subjective decisions on which responses are the same, as different phrasings of the same idea (e.g., 'hold papers down' and 'use as a paperweight') should be bucketed together. Human annotators must maintain evolving mental maps of a growing set of buckets, which makes the annotation process fatigue-intensive, error-prone, and infeasible for large corpora (Acar and Runco, 2014; Baten et al., 2020, 2021, 2022; Buczak et al., 2023). Furthermore, current literature lacks standardization in defining what qualifies as an 'infrequent' idea, resulting in limited psychometric validation.

We present MuseRAG, a fully automated, psychometrically validated system for frequency-based originality scoring—bringing us closer to a complete arsenal of automated assessment tools. Bucketing the same ideas together is computationally non-trivial: (i) semantic similarity alone is insufficient for idea bucketing, since similar embeddings may reflect rephrasings or entirely different intents, (ii) traditional clustering algorithms struggle with singleton and low-frequency ideas, which hold crucial signals for rarity scoring, (iii) fat-tailed bucket size distributions in real-world datasets defy assumptions of uniform or Gaussian cluster sizes, and (iv) bucket count grows as new ideas arrive, rendering ineffective text labeling tools that require label sets apriori. MuseRAG resolves these core challenges with a retrieval-augmented generation approach, where a zero-shot LLM acts as a judge to incrementally assign ideas to conceptually equivalent buckets. Unlike conventional clustering methods, our method replicates the subjective nature of human bucketing in both structure and resolution.

We also contribute to the creativity literature in two ways. *First*, we establish rigorous psychometric validity for frequency-based originality scoring, demonstrating high agreement with human annotations and strong correlations with relevant cognitive traits. In doing so, we provide insights on reliable ‘infrequency’ operationalizations. *Second*, we release an automated and interpretable scoring pipeline that is deployable across a wide range of open-ended ideation tasks, aiding creativity research at scale[1] (Kelty et al., 2023).

This work further makes a broader contribution to the EMNLP community. It exemplifies how bleeding-edge NLP techniques can solve long-standing annotation problems that have resisted algorithmic treatment. As the field seeks to extend its reach through interdisciplinary recontextualization, MUSERAG provides deep validation to enable widespread adoption by adjacent disciplines.

2 Related Work

2.1 Computational Assessment of Creativity

Creativity assessment has long relied on divergent thinking tasks like the Alternate Uses Test (AUT), where participants list novel uses for everyday objects (Guilford, 1967). Responses are traditionally scored for fluency (number of ideas), flexibility (the number of distinct semantic categories), and originality (statistical infrequency in a sample) (Dumas and Dunbar, 2014; Runco and Mraz, 1992).

Several computational methods have been proposed to quantify creativity. Unsupervised models approximate human novelty ratings by measuring an idea’s semantic distance from the task prompt (Beaty and Johnson, 2021; Dumas et al., 2021; Acar and Runco, 2014), or flexibility by measuring semantic diversity (Snyder et al., 2004; Bossomaier et al., 2009). Hybrid and supervised models predict novelty ratings directly using regression and clustering-based pipelines (Organisciak et al., 2023; Stevenson et al., 2020). However, these supervised approaches face generalizability issues: models trained on one task or dataset might perform poorly in another (Buczak et al., 2023). Importantly, these models typically approximate human novelty ratings and not social rarity.

Recent work underscores the importance of capturing conceptual intent rather than surface similarity (Olson et al., 2021). Yet, semantic similarity and clustering methods can conflate or over-separate ideas. Our approach addresses this by automating frequency-based originality scoring through intent-sensitive, zero-shot idea bucketing *at scale*—something human raters or clustering algorithms could not previously achieve.

[1]github link omitted for anonymity

2.2 Text Clustering and Annotation

Recent work has explored LLMs for zero-shot or few-shot clustering and annotation tasks (Xiao et al., 2023b). Deductive clustering prompts an LLM to partition small sets of texts simultaneously, generating categories or groupings directly (Viswanathan et al., 2024; Chew et al., 2023). Most LLM-based clustering methods assume all clusters are discoverable upfront and perform poorly when the concept space evolves over time. Inductive annotation, on the other hand, presents labeled exemplars to classify new instances incrementally (Dai et al., 2023). While the current approaches show promise on well-bounded tasks like topic labeling or thematic analysis, it remains unclear how best to navigate fat-tail distributed clusters where the cluster count scales indefinitely with data size.

2.3 LLM-as-a-Judge and RAG

LLM-as-a-Judge has recently emerged as a powerful paradigm for evaluating, ranking, and filtering outputs across tasks like summarization, translation, alignment, and reasoning (Li et al., 2024a; Liang et al., 2023; Zhao et al., 2024). Unlike earlier evaluation approaches in NLP tasks (Papineni et al., 2002; Zhang et al., 2019), judge LLMs can assess contextual appropriateness, intent, and subtle differences between candidates. Judging can be pointwise, pairwise, or listwise (Gao et al., 2023; Shen et al., 2024). Our task combines listwise judgment with decision-making: the LLM selects whether an idea matches any retrieved exemplar or forms a new semantic bucket, akin to selection-based judgment (Li et al., 2024b; Yao et al., 2023).

To stabilize this process at scale, we adopt an externally orchestrated Retrieval-Augmented Generation (RAG) framework (Lewis et al., 2020; Izacard and Grave, 2020). Unlike end-to-end or tool-using agent systems (Shinn et al., 2023), our model remains stateless. It receives a curated prompt assembled using K-NN search over a codebook database of previously catalogued buckets (Khandelwal et al., 2020), aligning with modular RAG practices. This separation of retrieval and generation ensures system stability at scale while retaining interpretability and psychometric auditability—critical in creativity research. Our architecture thus blends recent advances in LLM judgment and RAG to achieve scalable and valid annotations of frequency-based originality.

Dataset	# Participants	# Tasks	# Ideas	# Judges
`socialmuse24` (Baten et al., 2024)	109	5	5703	2
`beaty18` (Beaty et al., 2018)	171	2	2917	4
`silvia17` (Silvia et al., 2017)	141	2	2355	3
`beaty12` (Beaty and Silvia, 2012)	133	1	1807	3
`mohr16` (Hofelich Mohr et al., 2016)	$305 + 284$	$1 + 1$	$1930 + 1582$	4

Table 1: Dataset summary. Each participant did one task in `mohr16`. In other datasets, all participants did all tasks.

3 Dataset Acquisition

3.1 Primary Dataset: `socialmuse24`

We use the `socialmuse24` dataset (Baten et al., 2024) to establish criterion validity (Table 1). Two trained research assistants (H1 and H2) independently 'bucketed' the same yet differently phrased ideas in each task under common bucket IDs. The annotators saw the ideas in a random order. They followed the coding rules described by Bouchard and Hare (Bouchard Jr and Hare, 1970) and the scoring key of Guilford's test (Guilford et al., 1978). The dataset thus contains two categorical bucket IDs assigned by H1 and H2 for each idea, giving our ground truth. The dataset also contains the computationally-derived flexibility scores, *Creativity Quotient*, for each participant's idea set, which we use to test convergent validity (Snyder et al., 2004; Bossomaier et al., 2009).

3.2 Secondary Datasets

We use four publicly available AUT datasets to assess convergent and external validity (Organisciak et al., 2023; Beaty and Johnson, 2021) (Table 1).

The `beaty18` dataset (Beaty et al., 2018) contains four judges' *Creative Qualities* ratings of ideas on a 1 (not at all creative) to 5 (very creative) scale. The dataset also contains scores on: (i) *Creative Metaphor Generation*: Each participant generated novel metaphors to describe two open-ended prompts (Beaty and Silvia, 2013). Four judges rated each metaphor on a 1 (not at all creative) to 5 (very creative) scale; (ii) *Big Five Personality*: Each participant answered questionnaires on neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness (McCrae et al., 2005); (iii) *Fluid Intelligence*: Each participant guessed the next entry in sequences of images (Cattell and Cattell, 1960), letters (Ekstrom et al., 1976), and numbers (Thurstone, 1938); and (iv) *Creative Self-concept*: Each participant completed questionnaires on self-efficacy and creative self-identity (Karwowski, 2014).

The `silvia17` dataset (Silvia et al., 2017) contains three judges' *Creative Quality* ratings similarly as `beaty18`. The dataset also contains openness personality scores (Lee and Ashton, 2004).

`beaty12` (Beaty and Silvia, 2012) contains three judges' *Creative Quality* ratings, as well as *Big Five Personality*, *Creative Metaphor Generation*, and *Fluid Intelligence* scores similarly as `beaty18`.

`mohr16` (Hofelich Mohr et al., 2016) contains four judges' ratings on idea *Originality* and *Flexibility*. Here, *Originality* captured how uncommon, remote, and clever a response is, on a scale of 1 (least original) to 5 (most original) (Silvia et al., 2008). *Flexibility* was defined as the number of categories present within each participant's responses, scored by averaging the three judges' estimates.

4 Task Description

4.1 Problem Formulation

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ denote a corpus of N participants, each completing T ideation tasks. For each task $t \in \{1, \ldots, T\}$, participant p_i produces a variable-length set of $n_{i,t}$ free-form textual responses, denoted $\mathcal{I}_{i,t} = \{x_{i,t}^{(1)}, \ldots, x_{i,t}^{(n_{i,t})}\}$.

Let $\mathcal{X}_t = \bigcup_{i=1}^{N} \mathcal{I}_{i,t}$ denote the full idea set for task t. The goal is to induce a task-specific partition $\mathcal{B}_t = \{B_{t,1}, \ldots, B_{t,K_t}\}$ over $\mathcal{X}_t$, where each 'bucket' $B_{t,k} \subseteq \mathcal{X}_t$ contains semantically equivalent ideas expressing the same underlying concept.

Let $k(x)$ denote the index of the bucket to which idea $x \in \mathcal{X}_t$ is assigned. We define $m_{t,k}$ as the number of distinct participants contributing at least one idea to bucket $B_{t,k}$. Importantly, the bucketing is performed *within* each task and *across* participants, and no bucket identity is shared across tasks.

4.2 Originality Metrics

We explore 4 frequency-based originality metrics:

(i) **`rarity`**: This metric scores each idea bucket $B_{t,k}$ as $(1 - \frac{m_{t,k}}{N})$, capturing relative infrequency in the sample (Forthmann et al., 2020, 2017). A participant's unnormalized `rarity` score is the sum of these values across their ideas, $R^{\texttt{rarity}}_{i,t} = \sum_{x \in \mathcal{I}_{i,t}} \left(1 - \frac{m_{t,k(x)}}{N}\right)$.

(ii) **`shapley`**: This metric scores each bucket $B_{t,k}$ as $\frac{1}{m_{t,k}}$, setting the marginal value of a bucket to be inversely proportional to the number of participants sharing it (Page, 2018). A participant's unnormalized `shapley` score is the sum of these values across their ideas, $R^{\texttt{shapley}}_{i,t} = \sum_{x \in \mathcal{I}_{i,t}} \frac{1}{m_{t,k(x)}}$.

(iii) **`uniqueness`**: This metric assigns a score of 1 to an idea if it appears in a singleton bucket (i.e., $m_{t,k} = 1$), and 0 otherwise (Forthmann et al., 2020; Baten et al., 2021, 2024). A participant's unnormalized `uniqueness` score is the count of their unique ideas, $R^{\texttt{uniqueness}}_{i,t} = \sum_{x \in \mathcal{I}_{i,t}} \mathbb{I}\{m_{t,k(x)} = 1\}$.

(iv) **`threshold`**: This metric applies a tiered scoring function, $S(x)$, based on the bucket prevalence of an idea x (Olson et al., 2021; DeYoung et al., 2008; Forthmann et al., 2020) as,

$$S(x) = \begin{cases} 3 & \text{if } \frac{m_{t,k(x)}}{N} \leq 0.01 \\ 2 & \text{if } 0.01 < \frac{m_{t,k(x)}}{N} \leq 0.03 \\ 1 & \text{if } 0.03 < \frac{m_{t,k(x)}}{N} \leq 0.10 \\ 0 & \text{otherwise.} \end{cases}$$

A participant's unnormalized `threshold` score is the sum of these scores, $R^{\texttt{thresh}}_{i,t} = \sum_{x \in \mathcal{I}_{i,t}} S(x)$.

To compute a participant's overall unnormalized score across all tasks, we take $R^{\texttt{metric}}_{i} = \sum_{t=1}^{T} R^{\texttt{metric}}_{i,t}$. To account for fluency (i.e., the number of ideas $n_{i,t}$ contributed by participant p_i in task t), we define normalized originality as, $O^{\texttt{metric}}_{i,t} = \frac{R^{\texttt{metric}}_{i,t}}{n_{i,t}}$ and $O^{\texttt{metric}}_{i} = \sum_{t=1}^{T} O^{\texttt{metric}}_{i,t}$.

4.3 Evaluation Strategy

We evaluate construct validity for (i) idea-to-bucket clustering alignment and (ii) participant-level originality scoring. This helps assess how well computational bucketing replicates human judgments.

Bucket-level construct validity. The bucket labels are categorical and arbitrary. Moreover, the bucket sizes follow a fat-tailed distribution with a few highly frequent buckets and many rare ones (see §5.1). Thus, traditional clustering metrics (e.g., Adjusted Rand Index) can be misleading due to being inflated by rare buckets. We adopt *Adjusted Mutual Information (AMI)* (Vinh et al., 2010) as our primary metric, which adjusts for chance agreement, is robust to label permutation and skewed distributions, and is well-suited for comparing clusterings with different numbers of clusters. For insight development, we also use *Normalized Mutual Information (NMI)* (Vinh et al., 2010), which quantifies mutual dependence between clusterings without chance correction, and *V-measure* (Rosenberg and Hirschberg, 2007), which is the harmonic mean of homogeneity and completeness, reflecting both internal purity and cross-cluster coverage.

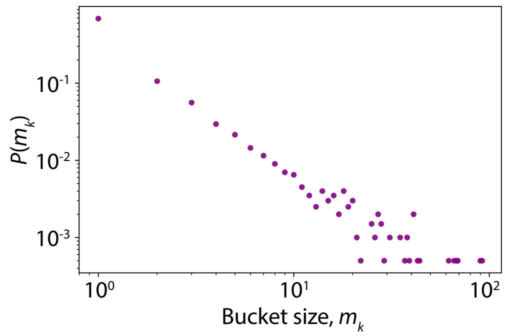

Figure 1: Idea bucket size distribution based on annotator H1's bucketing. See Figure A1 for H2's case.

Participant-level construct validity. For originality scoring agreement, we use (i) *Zero-order Correlations* (Pearson's r for linear agreement and Spearman's ρ for monotonic consistency), (ii) *Intraclass Correlation Coefficient* for consistency across judges (Shrout and Fleiss, 1979), and (iii) *Bland–Altman Plots* to identify systematic, scale-level biases (Bland and Altman, 1986).

Convergent and external validity. *Convergent validity* is assessed by correlating model originality scores with theoretically aligned creativity metrics (e.g., Creativity Quotient and Creative Quality Ratings). *External validity* is evaluated by correlating model scores with established psychological and cognitive variables: personality traits, creative metaphor generation ability, fluid intelligence, and creative self-concept (Beaty and Johnson, 2021).

5 Understanding Human-Annotated Ground Truth Characteristics

5.1 Distributional Properties of Idea Buckets

We assess the structure of idea diversity in `socialmuse24` using H1 and H2's buckets. H1 used more buckets per task (399.6, 95% C.I.: $[354.1, 445.1]$) than H2 (230.8 $[192.8, 268.8]$), indicating finer-grained distinctions in bucketing.

Next, we test whether the bucket frequencies follow a fat-tailed pattern. We fit a discrete power-law model to the bucket frequencies for each task and compare it to a lognormal distribution via a likelihood ratio test (Clauset et al., 2009). Both annotators produced fat-tailed distributions, with scaling exponents $\alpha_{\text{H1}} = 2.01$ $[1.73, 2.28]$ and $\alpha_{\text{H2}} = 1.74$ $[1.60, 1.88]$, consistent with power-law like behavior in linguistic and social systems ($\alpha \approx 2$ to 3) (Newman, 2018). This confirms that a few buckets are highly frequent while many are rare (Figure 1). However, the power-law model is not statistically favored over the lognormal alternative ($P \geq 0.05$), suggesting that the bucket size distributions are not strictly power-law and may be better described by lognormal or other alternatives.

5.2 Inter Human Annotator Agreement on Idea-level Bucketing

H1 and H2 show a mean AMI of 0.66 $[0.64, 0.68]$, indicating strong alignment beyond what would be expected by random bucketing. NMI elucidates how informative one annotator's bucketing is about the other's but does not adjust for chance (i.e., NMI is less conservative). As expected, the mean NMI is higher at 0.85 $[0.84, 0.87]$, reflecting strong underlying structure shared across annotators (Table A1).

V-measure also yields a high mean of 0.85 $[0.84, 0.87]$. Its homogeneity component (0.80) shows that H1's buckets are reasonably pure with respect to H2, and its high completeness component (0.92) shows that H2's buckets almost perfectly recover H1's buckets. This pattern corroborates that H1 split buckets more finely than H2, but both annotators identified similar idea groupings.

Overall, the annotators strongly agree on their idea bucketing, despite granularity differences.

5.3 Inter Human Annotator Agreement on Participant-level Originality Scoring

We compute participant-level $\{O_i^{\texttt{metric}}\}$ using H1 and H2's bucket assignments and assess agreement.

The `threshold` and `shapley` metrics show the strongest correlations (`threshold`: $r = 0.77$ $[0.69, 0.84]$; `shapley`: $r = 0.79$ $[0.70, 0.85]$; both $P < 0.001$). `uniqueness` and `rarity` show lower but still good correlations (`uniqueness`: $r = 0.73$ $[0.63, 0.81]$; `rarity`: $r = 0.72$ $[0.61, 0.81]$; both $P < 0.001$; see Table A2 for ρ estimates).

The `threshold` and `shapley` metrics also show the strongest average consistency across judges: $ICC(3, k) = 0.85$ $[0.78, 0.90]$, both $P < 0.001$. `uniqueness` yields the lowest but good agreement: $ICC(3, k) = 0.8$ $[0.71, 0.86]$, $P < 0.001$ (Table A3). Taken together, we note strong agreements in originality scoring across the human annotators.

5.4 Takeaways for MuseRAG Development

These analyses establish important expectations for machine-based originality scoring. *First*, human-annotated buckets exhibit a fat-tailed structure. Any automated scoring system must account for this characteristic for its bucketing performance to approach the strong AMI baseline of humans.

Second, based on the above evidence, we take the `threshold`-based normalized scores, $\{O_i^{\texttt{thresh}}\}$, as our person-level gold standard against which we evaluate machine-based originality scoring. We test for robustness against the other metrics.

6 The MuseRAG Originality Scorer

6.1 Insights from Early Prototypes

Our initial prototype mimicked a typical human annotator's workflow: judging each new idea against an expanding codebook of prior buckets and deciding whether the new idea rephrases an existing bucket or is sufficiently different to be a new one. To capture this, we prompted the LLM with the *full* existing codebook as it judged each new idea. However, this made the prompts prohibitively large when the bucket count exceeded $K_t \approx 150$. Given the fat-tailed bucket frequency distributions, massive corpora can have very large K_t, making exhaustive prompting infeasible. This motivated a retrieval-based approach: selecting a small set of candidate buckets from the current codebook against which an LLM can assess a new idea.

Our next prototype achieved this by adding a semantic similarity-based candidate bucket set retrieval mechanism. We used a fully LLM-managed pipeline for retrieval, decision-making, and codebook updating. This proved brittle as pipeline management errors compounded, especially by smaller LLMs (e.g., `phi4`). We therefore offloaded retrieval and codebook management to external components to stabilize the system, letting the LLM focus solely on the subjective bucketing decisions.

6.2 MuseRAG System Architecture

Algorithm 1 summarizes MuseRAG's workflow. The LLM processes one idea at a time and assigns it to a semantically equivalent bucket or creates a new one. For a given ideation task, a dynamic codebook is initialized and updated as new ideas arrive. For each idea $x \in \mathcal{X}$, a dictionary of candidate buckets $\mathcal{D}_x$ is constructed via K-NN-based semantic search over the current codebook (Khandelwal et al., 2020). $\mathcal{D}_x$ has a maximum size of K_c. When

Algorithm 1 MUSERAG: LLM-Based Incremental Bucketing for a Single Creativity Task

Require: Idea set $\mathcal{X} = \{x_1, x_2, \ldots, x_{|\mathcal{X}|}\}$, LLM, candidate dictionary size K_c
Ensure: Partition $\mathcal{B} = \{B_1, \ldots, B_K\}$, assignment map $k(x)$

Initialize empty codebook $\mathcal{C} \leftarrow \emptyset$
Initialize bucket index $K \leftarrow 0$
for all ideas $x \in \mathcal{X}$ **do**
  if $|\mathcal{C}| \leq K_c$ **then**
    $\mathcal{D}_x \leftarrow \mathcal{C}$
  else
    Use K-NN search to find top-K_c closest entries in $\mathcal{C}$ to x
    $\mathcal{D}_x \leftarrow \{(k_j, d_j)\}_{j=1}^{K_c}$
  end if
  Query LLM: "Is x a rephrasing of any $d_j \in \mathcal{D}_x$? Return k_j or -1." (In CoT prompting, also return a justification sentence)
  if LLM returns $k^* \neq -1$ **then**
    Assign $k(x) \leftarrow k^*$
    $B_{k^*} \leftarrow B_{k^*} \cup \{x\}$
  else
    $K \leftarrow K + 1$
    Create new bucket $B_K \leftarrow \{x\}$
    Update codebook $\mathcal{C} \leftarrow \mathcal{C} \cup \{(K, x)\}$
    Assign $k(x) \leftarrow K$
  end if
end for
return $\mathcal{B} = \{B_1, \ldots, B_K\}$, $k(x)\ \forall\, x \in \mathcal{X}$

the number of existing buckets is smaller than K_c, all of those buckets are taken in $\mathcal{D}_x$. Each candidate in the dictionary $\mathcal{D}_x = \{(k_j, d_j)\}_{j=1}^{K_c}$ maps bucket IDs to representative descriptions.

The LLM is prompted to determine whether x is a rephrasing of any d_j (baseline prompting). If so, it returns the corresponding key k_j; otherwise, it returns -1, signaling the creation of a new bucket with x as its description. We also explore Chain-of-Thought (CoT) prompting, where the LLM additionally provides a one-sentence reasoning (Wei et al., 2022). The codebook and bucket assignment are updated accordingly.

We experiment with a factorial combination of LLM model variants, sentence embeddings, and prompting strategies (see Appendix Section A.1). We fix $K_c = 10$ to keep prompt length manageable while leaving sufficient margin for retrieval noise, and test robustness against other K_c choices.

7 Results and Discussion

7.1 Computational Baselines

We use unsupervised clustering to establish a computational baseline for MUSERAG. We require algorithms that (i) allow clusters of vastly different sizes, including fat-tail distributed ones, and (ii) preserve singleton and rare buckets without dropping them as noise or outliers (§5.4).

These constraints discourage us from using algorithms like DBScan (singleton and rare buckets are likely to be marked as noise) (Ester et al., 1996) and HDBScan (minimum cluster size is 2) (Campello et al., 2013), and our experiments also corroborate their poor performance. K-means clustering is poor at handling imbalanced cluster sizes or shapes, and requires the number of clusters to be close to the number of datapoints to allow many singleton or rare buckets (MacQueen, 1967). Agglomerative hierarchical clustering is a reasonable choice for our constraints (Ward Jr, 1963). We report results with K-means and agglomerative algorithms.

For each algorithm, we automatically search for the optimal number of buckets K_t over the full region of $K_t = 1$ to $|\mathcal{X}_t|$. We evaluate structural and semantic criteria using (i) *Silhouette Score*, which assesses cluster quality based on geometric compactness and separation, with higher values indicating better-defined clusters (Rousseeuw, 1987); and (ii) *Semantic Score*, which is the geometric mean of coherence (intra-cluster similarity) and exclusivity (inter-cluster distinctiveness), encouraging clusters that are both internally consistent and mutually distinct (Mimno et al., 2011). We experiment with the same sentence embeddings as MUSERAG.

7.2 Distributional Properties of Computationally-labeled Idea Buckets

We find K-means and agglomerative algorithms to produce an exorbitantly high K_t—588 and 838 buckets by agglomerative (based on Silhouette and Semantic scores), and 831 and 797 buckets by K-means. For reference, $|\mathcal{X}_t| \approx 1141$ per task in `socialmuse24`. These bucket counts are significantly higher than H1 and H2's ($P < 0.001$). In contrast, the MUSERAG models produce K_t in the range of 255 to 465, overlapping those of the humans. The scaling exponents of K-means and agglomerative are systematically higher than the human baseline ($P < 0.001$), but the MUSERAG models align with the humans (Table A4).

Model	AMI	NMI	Pearson's r	Spearman's ρ	$ICC(3,1)$
`llama3.3` CoT	$\mathbf{0.59 \pm 0.05}$	$\mathbf{0.88 \pm 0.02}$	$\mathbf{0.88 \pm 0.04}$	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.88 \pm 0.04}$
`qwen3` CoT	0.56 ± 0.05	0.87 ± 0.02	0.79 ± 0.07	0.78 ± 0.07	0.77 ± 0.08
`phi4` CoT	0.54 ± 0.01	0.83 ± 0.01	0.78 ± 0.08	0.76 ± 0.08	0.72 ± 0.09
`llama3.3` Baseline	0.59 ± 0.03	0.86 ± 0.02	0.83 ± 0.06	0.79 ± 0.07	0.81 ± 0.06
`phi4` Baseline	0.53 ± 0.02	0.83 ± 0.01	0.80 ± 0.07	0.78 ± 0.08	0.75 ± 0.08
K-means Silhouette	0.32 ± 0.09	0.86 ± 0.02	0.65 ± 0.11	0.67 ± 0.11	0.62 ± 0.12
K-means Semantic	0.35 ± 0.06	0.87 ± 0.02	0.71 ± 0.10	0.70 ± 0.10	0.67 ± 0.10
Aggl. Silhouette	0.39 ± 0.02	0.85 ± 0.02	0.73 ± 0.09	0.68 ± 0.10	0.69 ± 0.10
Aggl. Semantic	0.31 ± 0.05	0.86 ± 0.02	0.65 ± 0.11	0.65 ± 0.11	0.61 ± 0.12

Table 2: Agreement metrics comparing computational models to H1's ground truths. Values are means $\pm$ half-width of the 95% C.I. ($N = 109$). See Table A5 for results based on H2's annotations.

7.3 Construct Validity of Idea-level Bucketing

Table 2 and Figure A2 show the AMI and NMI agreements between H1 and machine bucketing. The results are robust to taking H2 as the reference (see Table A5). Interestingly, all methods score highly in the less conservative NMI metric and match the H1-H2 agreement, showing reasonable preservation of semantic grouping.

However, when we correct for random chance and penalize mismatch in structure and granularity using the AMI metric, the MUSERAG models sustain human-like performance while the K-means and agglomerative algorithms suffer dramatically and systematically. Specifically, against a human-human AMI of 0.66 $[0.64, 0.68]$, the `llama3.3` LLM with CoT prompting achieves the best AMI among the MUSERAG models at 0.59 $[0.55, 0.64]$, while the silhouette-tuned agglomerative algorithm manages the best AMI among the baseline models at a poor 0.39 $[0.36, 0.41]$. This is unsurprising, since a drop in AMI implies deviation from the structure and resolution of the human bucketing, which is corroborated by the systematically larger number of buckets K-means and agglomerative algorithms produce. In contrast, the MUSERAG models preserve more of the mutual structures, semantic coherence, and resolution, capturing up to 89% of the fine-grained patterns humans see.

Overall, MUSERAG shows strong idea-bucketing alignment with the humans, surpassing the performances of clustering-based baselines.

7.4 Construct Validity of Participant-level Originality Scoring

Table 2 and Figure A3 show the participant-level $\{O_i^{\text{thresh}}\}$ score agreements based on H1 and machine bucketing. The results are robust to taking H2 as the reference (Table A5). MUSERAG with `llama3.3` and CoT prompting once again shows the best correlation ($r = 0.89$ $[0.83, 0.92]$, $P < 0.001$). The baselines perform significantly worse, with the silhouette-tuned agglomerative algorithm achieving the best baseline correlation ($r = 0.73$ $[0.63, 0.81]$, $P < 0.001$).

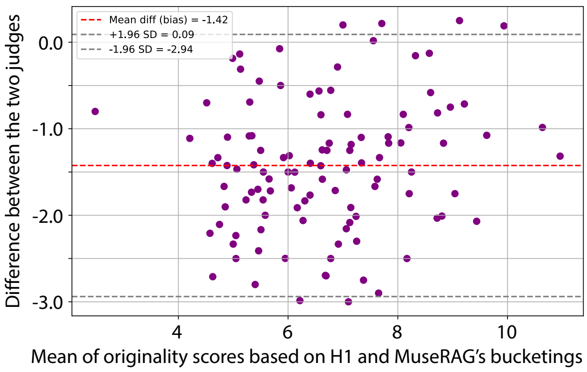

Figure 2: Bland-Altman visualization for bias detection.

MUSERAG with `llama3.3` and CoT prompting also shows the best $ICC(3,1) = 0.88$ $[0.83, 0.92]$, $P < 0.001$. The clustering baselines reach a maximum of $ICC(3,1) = 0.69$ $[0.57, 0.77]$, $P < 0.001$, with the silhouette-tuned agglomerative model, remaining significantly lower than `llama3.3`'s performance ($P < 0.001$). Based on the above evidence, we pick `llama3.3` with CoT prompting as the default configuration for MUSERAG and use it for the remaining analysis.

We next visualize a Bland-Altman plot to identify systematic biases between H1 and MUSERAG-derived originality scores (Figure 2). 94.5% of the points fall within the limits of agreement (LoA) of ± 1.96 SDs, and so does the mean difference (bias). This shows that MUSERAG-derived scores stay strongly in line with H1's scores across the originality spectrum. Although the proportional bias regression slope is slightly positive (0.09), the effect is not statistically significant ($P > 0.05$), suggesting no systematic trend where the machine over- or under-scores ideas as originality increases. This supports the conclusion that MUSERAG provides stable, human-comparable originality assessments.

Taken together, MUSERAG shows strong originality scoring validity against human ground truth.

7.5 Convergent and External Validity

MUSERAG's $\{O_i^{\text{thresh}}\}$ scores correlate strongly with participant-level Creativity Quotient (CQ) scores in the `socialmuse24` dataset ($r = 0.4$ $[0.23, 0.55], P < 0.001$). CQ is a flexibility measure that captures the diversity of semantic categories. However, CQ is unnormalized and confounded by idea fluency. Unsurprisingly, unnormalized $\{R_i^{\text{thresh}}\}$ shows a stronger correlation with CQ ($r = 0.47$ $[0.32, 0.61], P < 0.001$).

MUSERAG's $\{O_i^{\text{thresh}}\}$ scores correlate strongly with person-level average creative quality ratings (`beaty18`: $r = 0.77$ $[0.71, 0.83], P < 0.001$; `silvia17`: $r = 0.54$ $[0.41, 0.65], P < 0.001$; `beaty12`: $r = 0.42$ $[0.27, 0.55], P < 0.001$). The `mohr16` dataset contains rating-based originality scores, which correlate strongly with our frequency-based originality scores ($r = 0.42$ $[0.35, 0.49], P < 0.001$). This dataset also contains manually annotated flexibility scores, which do not account for fluency. Unsurprisingly, this flexibility score correlates strongly with unnormalized $\{R_i^{\text{thresh}}\}$ ($r = 0.76$ $[0.73, 0.80], P < 0.001$). Overall, MUSERAG demonstrates excellent convergent validity.

In terms of external validity, we find MUSERAG's $\{O_i^{\text{thresh}}\}$ scores to correlate significantly with the person-level average creative metaphor generation ratings (`beaty18`: $r = 0.2$ $[0.04, 0.35], P < 0.05$; `beaty12`: $r = 0.25$ $[0.09, 0.41], P < 0.01$). $\{O_i^{\text{thresh}}\}$ correlates well with openness personality trait (`beaty18`: $\rho = 0.16$ $[0.01, 0.30], P < 0.05$; `beaty12`: $r = 0.30$ $[0.14, 0.45], P < 0.001$; `silvia17`: $\rho = 0.14$ $[-0.02, 0.30]$, marginal $P = 0.09$). We find strong correlations with creative self-identity ($r = 0.34$ $[0.19, 0.48], P < 0.001$) and self-efficacy ($r = 0.29$ $[0.14, 0.44], P < 0.001$). We did not find any correlation with fluid intelligence or other personality traits. Our results largely corroborate previous insights (Beaty and Johnson, 2021), establishing strong external validity.

7.6 Robustness

The results depend on LLM, sentence embedding, and prompting strategy choices. We obtain the best results for the `llama3.3:70b` LLM (Meta AI, 2024), `e5-large-v2.1` sentence embedding (Wang et al., 2022), and Chain-of-Thought prompting (Wei et al., 2022) combination (§A.1). We further probe this configuration's robustness across $K_c \in \{5, 15\}$, and find results statistically similar to the default $K_c = 10$. To assess ordering effects, we run the configuration with randomly ordered $\mathcal{X}_t$ across 3 seeds. We find the results stable within the bounds reported in Table 2. The main results with the `threshold` metric are largely reproduced by the other three. But we find that `rarity` shows proportional bias in the Bland-Altman plot (slope = 0.2, $P < 0.01$), while `shapley` and `uniqueness` show no correlation with openness in the `silvia17` dataset, losing some external validity. The `threshold` metric thus emerges as the most robust choice.

8 Conclusion

This work presents a scalable, zero-shot LLM-based system for scoring the originality of creative ideas, addressing long-standing challenges in the automation of divergent thinking assessment. By leveraging the LLM-as-a-judge paradigm with externally orchestrated retrieval, our method provides psychometrically aligned, intent-sensitive judgments without the need for task-specific fine-tuning or training data.

The proposed system robustly handled all four distinct datasets used in our evaluation, demonstrating consistent performance across varying task structures and idea distributions. Unlike opaque embedding-based approaches, our use of chain-of-thought (CoT) prompting enables interpretable outputs, offering justifications for each originality score and increasing transparency in AI decision-making.

Our approach is well-suited to support the growing body of research in human-AI creativity assessment, particularly as large-scale, high-throughput studies become increasingly common (Doshi and Hauser, 2024). By combining reliability, interpretability, and scale, this system expands the practical and methodological toolkit for creativity researchers and opens new avenues for measuring and understanding creative potential in both human and artificial agents.

Limitations

Although not the focus of this paper, future applications of frequency-based originality scoring systems should carefully consider demographic fairness and accessibility. Differences in language use across cultural or educational backgrounds may affect bucketing judgments—particularly in non-English settings—potentially introducing unintended bias if not monitored.

Our validation is confined to the AUT and similar text-based divergent thinking tasks. It remains to be seen how well our approach generalizes to other domains of creative production (e.g., design, visual arts).

The effectiveness of our approach depends on carefully curated prompts. Although we use externally orchestrated RAG to control the context injected into the LLM, the system may still be sensitive to prompt length or phrasing (Liu et al., 2023). Subtle changes in prompt format can positively or negatively affect judgment outcomes, which remains to be explored further.

The system has room for improvement in terms of efficiency. We loop one idea at a time through the LLM. Future research can explore multi-idea batching to enhance efficiency. However, we observe *simple* and *focused* LLM assignments to stabilize the system, and demanding more out of the LLM at each prompt may make the system brittle, especially for small-sized LLMs.

The bucketing reasoning performance can be improved by adding multi-step thinking approaches. However, that might also increase computation cost.

We kept the candidate dictionary size, K_c, small at $\{5, 10, 15\}$. Whether increasing the size further improves performance remains to be seen. However, any performance improvement mechanism must be justified against the associated token usage and computation cost increases.

Our most successful `threshold` metric applies a heuristic-based scoring function borrowed from prior literature. The robustness of the tiering choices of the scoring function remains to be examined.

Ethical Considerations

We reanalyzed datasets from prior works and did not collect any new human data for this research. Given the nature of the project in creative assessment, we do not readily foresee potential harm.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Selcuk Acar and Mark A Runco. 2014. Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal*, 26(2):229–238.

Raiyan Abdul Baten, Richard N Aslin, Gourab Ghoshal, and Ehsan Hoque. 2021. Cues to gender and racial identity reduce creativity in diverse social networks. *Scientific Reports*, 11(1):10261.

Raiyan Abdul Baten, Richard N Aslin, Gourab Ghoshal, and Ehsan Hoque. 2022. Novel idea generation in social networks is optimized by exposure to a "Goldilocks"' level of idea-variability. *PNAS Nexus*, 1(5):pgac255.

Raiyan Abdul Baten, Daryl Bagley, Ashely Tenesaca, Famous Clark, James P Bagrow, Gourab Ghoshal, and Ehsan Hoque. 2020. Creativity in temporal social networks: How divergent thinking is impacted by one's choice of peers. *Journal of the Royal Society Interface*, 17(171):20200667.

Raiyan Abdul Baten, Ali Sarosh Bangash, Krish Veera, Gourab Ghoshal, and Ehsan Hoque. 2024. AI can enhance creativity in social networks. *arXiv preprint arXiv:2410.15264*.

Roger E Beaty and Dan R Johnson. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, 53(2):757–780.

Roger E Beaty, Yoed N Kenett, Alexander P Christensen, Monica D Rosenberg, Mathias Benedek, Qunlin Chen, Andreas Fink, Jiang Qiu, Thomas R Kwapil, Michael J Kane, and 1 others. 2018. Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences*, 115(5):1087–1092.

Roger E Beaty and Paul J Silvia. 2012. Why do ideas get more creative across time? An executive interpretation of the serial order effect in divergent thinking tasks. *Psychology of Aesthetics, Creativity, and the Arts*, 6(4):309.

Roger E Beaty and Paul J Silvia. 2013. Metaphorically speaking: Cognitive abilities and the production of figurative language. *Memory & cognition*, 41:255–267.

Kenes Beketayev and Mark A Runco. 2016. Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's Journal of Psychology*, 12(2):210.

J. Martin Bland and Douglas G. Altman. 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310.

Terry Bossomaier, Mike Harré, Anthony Knittel, and Allan Snyder. 2009. A semantic network approach to the Creativity Quotient (CQ). *Creativity Research Journal*, 21(1):64–71.

Thomas J Bouchard Jr and Melana Hare. 1970. Size, performance, and potential in brainstorming groups. *Journal of Applied Psychology*, 54(1p1):51.

Philip Buczak, He Huang, Boris Forthmann, and Philipp Doebler. 2023. The machines take over: A comparison of various supervised learning approaches for automated scoring of divergent thinking tasks. *The Journal of Creative Behavior*, 57(1):17–36.

Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 160–172. Springer.

Raymond Bernard Cattell and Alberta KS Cattell. 1960. *Measuring intelligence with the culture fair tests*. Institute for Personality and Ability Testing.

Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100*.

Colin G DeYoung, Joseph L Flanders, and Jordan B Peterson. 2008. Cognitive abilities involved in insight problem solving: An individual differences model. *Creativity Research Journal*, 20(3):278–290.

Anil R Doshi and Oliver P Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

Denis Dumas and Kevin N Dunbar. 2014. Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity*, 14:56–67.

Denis Dumas, Peter Organisciak, and Michael Doherty. 2021. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, 15(4):645.

Ruth B Ekstrom, John W French, Harry H Harman, and D Dermen. 1976. Manual for kit of factor-referenced tests. *Princeton, NJ: Educational Testing Service*, 586:1989–1995.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231.

Boris Forthmann, Heinz Holling, Pınar Çelik, Martin Storme, and Todd Lubart. 2017. Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal*, 29(3):257–269.

Boris Forthmann, Sue Hyeon Paek, Denis Dumas, Baptiste Barbot, and Heinz Holling. 2020. Scrutinizing the basis of originality in divergent thinking tests: On the measurement precision of response propensity estimates. *British Journal of Educational Psychology*, 90(3):683–699.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with ChatGPT. *arXiv preprint arXiv:2304.02554*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Joy Paul Guilford. 1967. *The Nature of Human Intelligence*. McGraw-Hill.

JP Guilford, PR Christensen, PR Merrifield, and RC Wilson. 1978. *Alternate Uses: Manual of Instructions and Interpretation*. Orange, CA: Sheridan Psychological Services.

Alicia Hofelich Mohr, Andrew Sell, and Thomas Lindsay. 2016. Thinking inside the box: Visual design of the response box affects creative divergent thinking in an online survey. *Social Science Computer Review*, 34(3):347–359.

Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*.

Maciej Karwowski. 2014. Creative mindsets: Measurement, correlates, consequences. *Psychology of Aesthetics, Creativity, and the Arts*, 8(1):62.

Sean Kelty, Raiyan Abdul Baten, Adiba Mahbub Proma, Ehsan Hoque, Johan Bollen, and Gourab Ghoshal. 2023. Don't follow the leader: Independent thinkers create scientific innovation. *arXiv preprint arXiv:2301.02396*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kibeom Lee and Michael C Ashton. 2004. Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, 39(2):329–358.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP task. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2024a. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, and 1 others. 2024b. DALK: Dynamic co-augmentation of LLMs and KG to answer Alzheimer's disease questions with scientific literature. *arXiv preprint arXiv:2405.04819*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

Robert R McCrae, Paul T Costa, Jr, and Thomas A Martin. 2005. The NEO-PI-3: A more readable revised NEO personality inventory. *Journal of Personality Assessment*, 84(3):261–270.

Meta AI. 2024. LLaMA 3.3-70B-Instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Accessed: 2025-05-18.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

Mark Newman. 2018. *Networks*. Oxford University Press.

Jay A Olson, Johnny Nahas, Denis Chmoulevitch, Simon J Cropper, and Margaret E Webb. 2021. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, 118(25):e2022340118.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

Peter Organisciak and Denis Dumas. 2020. Open creativity scoring. https://openscoring.du.edu. [Computer software].

Scott E. Page. 2018. *The Model Thinker: What You Need to Know to Make Data Work for You*. Basic Books, New York.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Roni Reiter-Palmon, Boris Forthmann, and Baptiste Barbot. 2019. Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):144.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Peter J Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Mark A Runco and Garrett J Jaeger. 2012. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96.

Mark A Runco and Wayne Mraz. 1992. Scoring divergent thinking tests using total ideational output and a creativity index. *Educational and Psychological Measurement*, 52(1):213–221.

Yanxin Shen, Lun Wang, Chuanqi Shi, Shaoshuai Du, Yiyi Tao, Yixian Shen, and Hang Zhang. 2024. Comparative analysis of listwise reranking with large language models in limited-resource language contexts. *arXiv preprint arXiv:2412.20061*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420.

Paul J Silvia, Emily C Nusbaum, and Roger E Beaty. 2017. Old or new? Evaluating the old/new scoring method for divergent thinking tasks. *The Journal of Creative Behavior*, 51(3):216–224.

Paul J Silvia, Beate P Winterstein, John T Willse, Christopher M Barona, Joshua T Cram, Karl I Hess, Jenna L Martinez, and Crystal A Richard. 2008. Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, 2(2):68.

Allan Snyder, John Mitchell, Terry Bossomaier, and Gerry Pallier. 2004. The Creativity Quotient: An objective scoring of ideational fluency. *Creativity Research Journal*, 16(4):415–419.

C. Stevenson, I. Smal, M. Baas, M. Dahrendorf, R. Grasman, C. Tanis, E. Scheurs, D. Sleiffer, and H. van der Maas. 2020. Automated AUT scoring using a big data variant of the consensual assessment technique. Report Final Technical Report, Modeling Creativity Project, Universiteit van Amsterdam, Amsterdam. Faculty of Social and Behavioural Sciences (FMG), Psychology Research Institute (PsyRes).

L. L. Thurstone. 1938. Primary mental abilities. *The Mathematical Gazette*, 22(251):411–412.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Vijay Viswanathan, Kiril Gashteovski, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. Large language models enable few-shot clustering. *Transactions of the Association for Computational Linguistics*, 12:321–333.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Joe H. Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023a. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023b. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 75–78.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36:11809–11822.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.

Lirui Zhao, Yue Yang, Kaipeng Zhang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, and Rongrong Ji. 2024. Diffagent: Fast and accurate text-to-image API selection with large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6390–6399.

A Supplementary Materials

A.1 System Component Choices

We experiment with the following system component alternatives:

(i) Large language models: $\mathcal{M}$ = {`llama3.3:70b-Instruct` (Meta AI, 2024; Grattafiori et al., 2024), `qwen3:32b` (Yang et al., 2025), `phi4:14b` (Abdin et al., 2024)}. We pick these mid-sized, open-source models for their cost and computation efficiencies.

(ii) Sentence embedding models: $\mathcal{E}$ = {`all-mpnet-base-v2` (Reimers and Gurevych, 2019), `bge-large-en-v1.5` (Xiao et al., 2023a), `e5-large-v2.1` (Wang et al., 2022)}. These models are freely available on Huggingface and have been widely used in recent technological developments.

(iii) Prompting strategies: $\mathcal{P}$ = {`baseline_prompting`, `cot_prompting` (Wei et al., 2022)}.

In our experiments, we found the combination of `llama3.3:70b-Instruct`, `e5-large-v2.1`, and `cot_prompting` to give the best performance.

A.2 Experimentation Setup and GPU Usage

We conducted all experiments using (i) an Intel Core i7-based computer with 64GB RAM and an RTX 3070 Ti graphics card, and (ii) three MacBook Pro laptops. All our code and data are available on GitHub. The R&D and final result generation took roughly 100 GPU days.

A.3 LLM Prompts

System Prompt (Baseline Condition)

```
You are an idea bucket annotator for ideas generated for the object {object_name} in Guilford's Alternative Uses Test. You will be given an input_idea to annotate against up to {comparison_k} comparison_ideas, given to you in a dictionary format with key-value pairs of comparison_idea_ID: comparison_idea_description. The keys are integers, and the values are strings. Your goal is to determine if the input_idea is a rephrased version of one of those comparison_idea_description, or if it is different.
if input_idea is a rephrased version of a certain comparison_idea_description:
   your_annotation_ID = comparison_idea_ID key of that comparison_idea_description value
elif input_idea is a different one:
   your_annotation_ID = -1
Your response must be a text string containing exactly: <your_annotation_ID>.
For example: if your_annotation_ID is 6 since the input idea is a rephrased version of comparison_idea_ID 6, your response string should be "6". Another example: if your_annotation_ID is -1 because the input idea is not a rephrasing of any comparison_idea_ID, your response string should be "-1".
Absolutely do not provide any extra text.
```

System Prompt (CoT Condition)

```
You are an idea bucket annotator for ideas generated for the object {object_name} in Guilford's Alternative Uses Test. You will be given an input_idea to annotate against up to {comparison_k} comparison_ideas, given to you in a dictionary format with key-value pairs of comparison_idea_ID: comparison_idea_description. The keys are integers, and the values are strings. Your goal is to determine if the input_idea is a rephrased version of one of those comparison_idea_description, or if it is different.
if input_idea is a rephrased version of a certain comparison_idea_description:
   your_annotation_ID = comparison_idea_ID key of that comparison_idea_description value
elif input_idea is a different one:
   your_annotation_ID = -1
You will also provide a reason string containing a single sentence explaining why you gave the input_idea that specific your_annotation_ID.
Your response must be a text string containing exactly: <your_annotation_ID><SPACE><reason>.
For example: if your_annotation_ID is 6 and the reason is "The input idea is a rephrased version of comparison_idea_ID 6", your response string should be "6 The input idea is a rephrased version of comparison_idea_ID 6". Another example: if your_annotation_ID is -1 and the reason is "The input idea is not a rephrasing of any comparison_idea_ID", your response string should be "-1 The input idea is not a rephrasing of any comparison_idea_ID".
Absolutely do not provide any extra text.
```

User Prompt Per Idea (Both Conditions)

```
input_idea: {idea_text}
comparison_ideas: {repr(comparison_ideas)}
```

A.4 AI Usage

We used Grammarly AI to improve the grammatical accuracy of the manuscript.

A.5 Supplementary Tables and Figures

Table A1: Inter-human annotator agreement on idea bucketing in `socialmuse24`.

Metric	**Mean** [95% C.I.]
AMI	0.66 $[0.64, 0.68]$
NMI	0.85 $[0.84, 0.88]$
V-measure	0.85 $[0.84, 0.87]$
Homogeneity	0.80 $[0.77, 0.82]$
Completeness	0.92 $[0.89, 0.95]$

Table A2: Pearson and Spearman correlations of participant-level scores based on H1 and H2's bucketing. $N = 109$ in all cases.

Scoring Method	**Correlation Type**	**Estimate**	95% **C.I.**	P**-value**
`threshold`	Pearson's r	0.77	$[0.69, 0.84]$	$P < 0.001$
	Spearman's ρ	0.75	$[0.65, 0.82]$	$P < 0.001$
`shapley`	Pearson's r	0.79	$[0.70, 0.85]$	$P < 0.001$
	Spearman's ρ	0.74	$[0.64, 0.82]$	$P < 0.001$
`rarity`	Pearson's r	0.72	$[0.61, 0.80]$	$P < 0.001$
	Spearman's ρ	0.64	$[0.51, 0.74]$	$P < 0.001$
`uniqueness`	Pearson's r	0.73	$[0.63, 0.81]$	$P < 0.001$
	Spearman's ρ	0.66	$[0.54, 0.76]$	$P < 0.001$

Table A3: ICC reliability of the participants' originality scores based on H1 and H2's bucketing.

Scoring Method	$ICC(3, k)$	F	$df1$	$df2$	P**-value**	95% **C.I.**
`threshold`	0.85	6.79	108	108	$P < 0.001$	$[0.78, 0.90]$
`shapley`	0.85	6.67	108	108	$P < 0.001$	$[0.78, 0.90]$
`rarity`	0.83	5.73	108	108	$P < 0.001$	$[0.75, 0.88]$
`uniqueness`	0.80	4.97	108	108	$P < 0.001$	$[0.71, 0.86]$

Table A4: Cluster count K and power-law exponent α for various computational scoring methods.

Model	K **[95% C.I.]**	α **[95% C.I.]**
`llama3.3` CoT	465.4 [426.8, 504.0]	2.28 [2.14, 2.42]
`qwen3` CoT	462.4 [432.7, 492.1]	2.43 [2.20, 2.67]
`phi4` CoT	255.0 [207.3, 302.7]	2.39 [1.72, 3.05]
`llama3.3` Baseline	367.8 [333.3, 402.3]	2.29 [1.97, 2.61]
`phi4` Baseline	275.6 [229.5, 321.7]	2.51 [2.23, 2.78]
K-means Silhouette	830.6 [729.2, 932.0]	3.12 [2.82, 3.43]
K-means Semantic	797.4 [757.8, 837.0]	3.12 [2.67, 3.57]
Agglomerative Silhouette	588.0 [524.9, 651.1]	5.68 [1.26, 10.09]
Agglomerative Semantic	838.0 [815.9, 860.1]	3.80 [2.63, 4.97]

Table A5: Agreement metrics comparing computational models to H2's ground truths. Values denote mean $\pm$ half-width of the 95% C.I. ($N = 109$).

Model	**AMI**	**NMI**	**Pearson's r**	**Spearman's ρ**	$ICC(3,1)$
`llama3.3` CoT	0.57 ± 0.04	$\mathbf{0.84 \pm 0.02}$	$\mathbf{0.76 \pm 0.08}$	$\mathbf{0.74 \pm 0.09}$	0.74 ± 0.09
`qwen3` CoT	0.54 ± 0.04	0.83 ± 0.02	0.74 ± 0.09	0.73 ± 0.09	0.74 ± 0.09
`phi4` CoT	0.56 ± 0.03	0.79 ± 0.01	0.67 ± 0.10	0.68 ± 0.10	0.67 ± 0.10
`llama3.3` Baseline	$\mathbf{0.59 \pm 0.03}$	0.83 ± 0.01	0.76 ± 0.08	0.74 ± 0.09	$\mathbf{0.75 \pm 0.08}$
`phi4` Baseline	0.55 ± 0.04	0.80 ± 0.01	0.73 ± 0.09	0.71 ± 0.10	0.73 ± 0.09
K-means Silhouette	0.28 ± 0.07	0.80 ± 0.02	0.59 ± 0.12	0.62 ± 0.12	0.59 ± 0.12
K-means Semantic	0.30 ± 0.05	0.80 ± 0.02	0.66 ± 0.11	0.68 ± 0.10	0.66 ± 0.11
Aggl. Silhouette	0.36 ± 0.03	0.80 ± 0.02	0.65 ± 0.11	0.60 ± 0.12	0.64 ± 0.11
Aggl. Semantic	0.26 ± 0.05	0.80 ± 0.02	0.60 ± 0.12	0.64 ± 0.11	0.60 ± 0.12

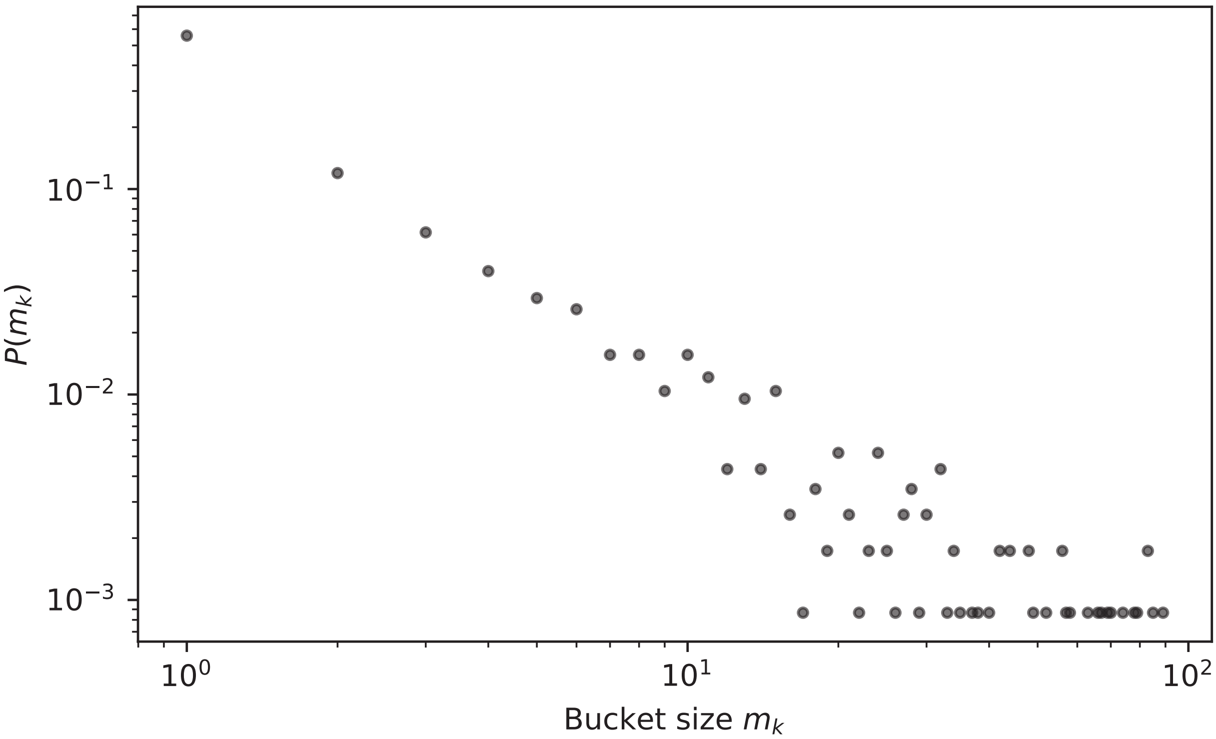

Figure A1: Idea bucket size distribution based on annotator H2's bucketing.

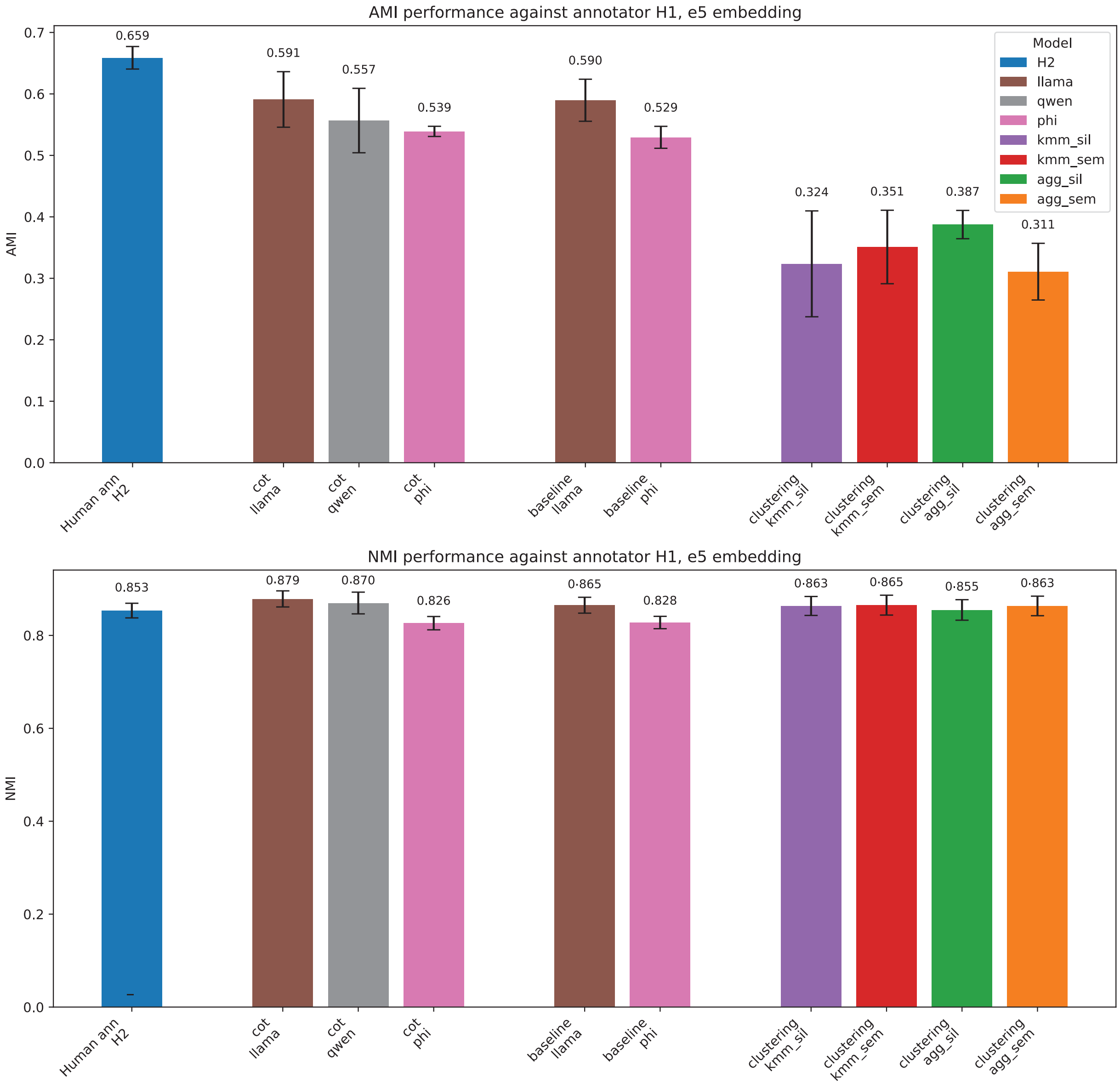

Figure A2: AMI and NMI performance comparison against annotator H1

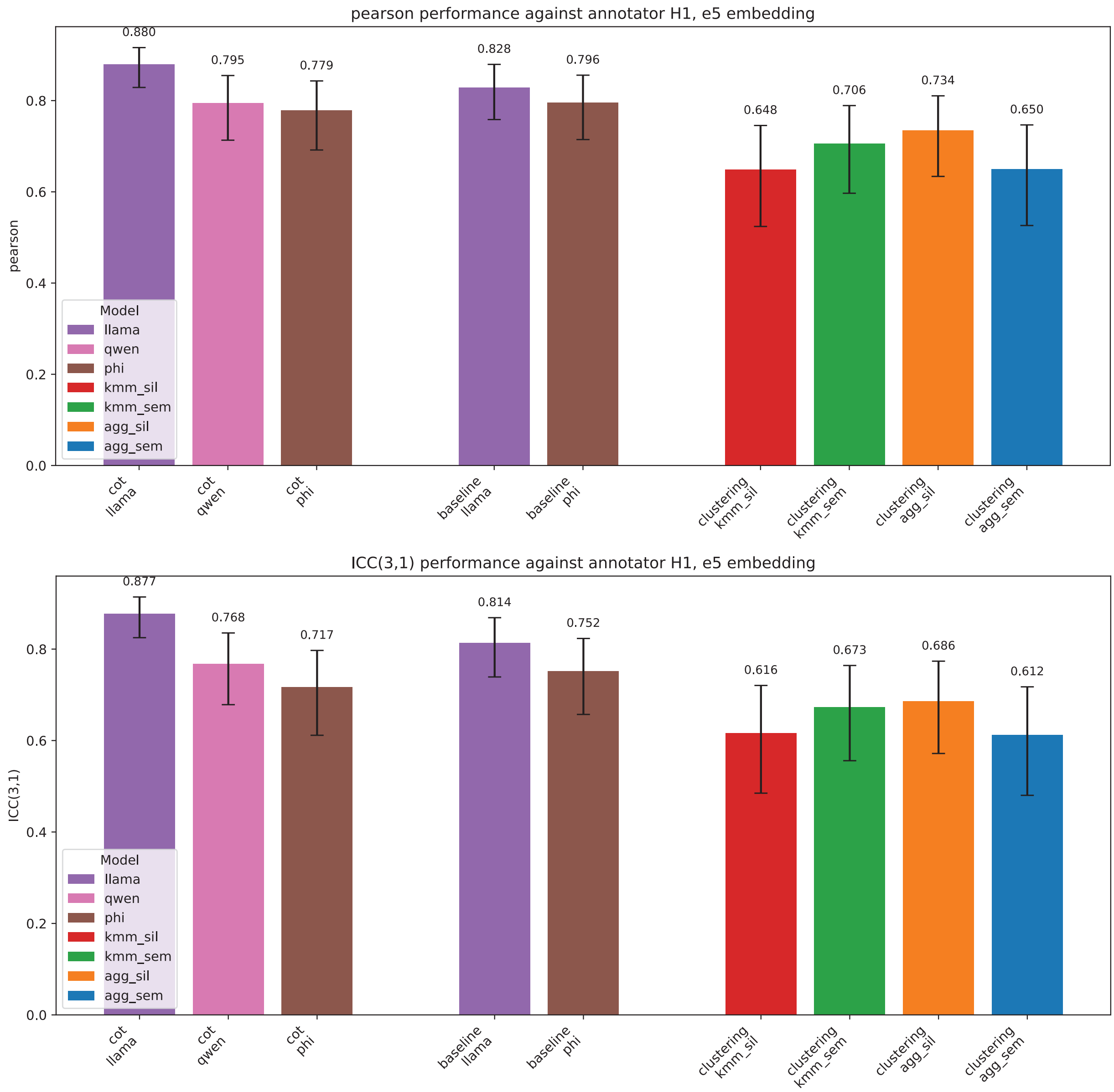

Figure A3: Pearson's r and ICC performance comparison against annotator H1