VISUAL RECODING AS A STRATEGY FOR ACHIEVING DIVERGENT CREATIVITY IN AN LLM POETRY GENERATION TASK

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have revolutionized technical tasks with clear and objective quality signals, such as code generation and mathematical proofs, but yield mixed results on creative, divergent reasoning tasks such as fiction and poetry. Poetry generation may serve as a benchmark for divergent creativity because it requires the ability to construct and stabilize meaning that transcends the well-learned lexical associations of its constituent words within a highly constrained form. Across diverse poetic traditions, visualization has been used to foster novel poetic insights. Here, we used two LLM families (Gemini and Qwen) to generate linked verse poetry in a task inspired by verse capping exercises from the Haikai no renga tradition. From an existing seed verse (maeku), models were asked to generate two additional verses that first extended the seed verse and then pivoted its meaning. We experimentally manipulated the linking modality: in one condition, models linked verses entirely via text generation; in a second condition, models had no direct access to previous verse text and instead generated images to pass meaning to the next verse. Compared to text-based generation, visual linking produced verses with greater lexical density, larger pivots from the seed verse in terms of lexical overlap and embedding distance, and a more dynamic, narrative writing style (e.g., more verbs, fewer articles). The results suggest that visual linking may be a promising direction for enabling creative divergence in LLM generation tasks.

1 INTRODUCTION

Large Language Models (LLMs) continue to improve in many benchmarks that have concrete, verifiable objective success criteria, such as code, math, logic, and textbook science reasoning. However, they continue to lag behind humans in their ability to creatively transcend existing examples or go beyond current theory in any domain (Wang et al., 2026). For example, LLMs can reason about the logical consequences and entailments of current theories and compare and contrast how divergent models may account for an observation, but they often cannot reason outside of current theory to make genuinely novel advances (Ruan et al., 2026). By some measures, such creative potential may even be decreasing as LLMs increasingly align with human preferences (Hernández-Espinosa et al., 2026). This divergent creativity deficit compromises AI's ability to create art or write fiction independently or with humans, where the underlying architecture of AI tends to pull creativity toward dominant attractor states (Chakrabarty et al., 2024; 2025). This can make language more generic in creative reasoning (Wenger & Kenett, 2026), prose (Li et al., 2026), or emphasize surface-layer, stereotypical features of an author's voice, such as using more melancholy and gothic language than Poe when asked to sound like Poe (Anonymous, 2026).

Although poetry has received less attention than product development or idea generation in research on AI creativity (Lin et al., 2025; Su et al., 2025), it has great potential as a target domain for testing short and long-term capability requirements for LLM creativity. Poetry is useful for studying creative reasoning because it concentrates meaning into few tokens and works best when it achieves an unexpected but resonant sense or image—a unique focal sense or meaning—that transcends its lexical constituents in readers' minds as they work through its intra- and inter-textual associations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Poets across Western and Japanese traditions have used visualization techniques to foster poetic creativity (Pound, 1913; Shirane, 1998). Visualization can aid creativity by breaking entrenched lexical associations, refocusing thoughts on other resonances in a poetic image, and then translating this image back into words.

Japenese renga is a form of linked verse that often uses imagery to facilitate the creation of a collaborative poem across several authors (Carter, 1991). In a renga, poets take turns providing verses that link to a specific aspect of a previous verse, thus creating a resonant sense across a combined two-verse sequence or "waka," while while pivoting the poem away from preceding verses to create a sense of movement akin to a montage sequence. Bashō, a prominent master of the *Haikai no renga* form, explicitly advocates for using imagery to create linked verses with layered meanings that convey subtle "scents" between verses like mood, kinetic energy, or social connotations (Shirane, 1998).

1.1 Current work

Here, we used a poetry completion task, inspired by an abbreviated form of renga "verse-capping" exercises (Shirane, 1998), to study how visualization impacts LLM creativity in generating linked verses based on Bashō's scent taxonomy. In the verse-capping task, an LLM was given a hokku seed verse (a *maeku*) and asked to complete or cap it with a second verse using a link strategy from Bashō's taxonomy: content, mood/emotion, kinetic energy, or social connotations. After this verse, the AI then generated a final verse to capture what, in a full renga (often 36 to 1000 verses (Carter, 1991)), would pivot the meaning away from the maeku, again using a specific link type, creating a unique combined sense with its immediate preceding verse.

The key manipulation for our study was that in one condition, the LLM could only connect verses using images: each of the two text verses was generated solely based on an image of the previous verse rather than the text. The impact of imagery on LLM verse capping was assessed by comparing it to a text condition, which handed off text between LLM calls but did not receive any visual input.

We compared the visual and text-based verse-capping strategies' impacts on creativity by examining lexical density (a measure of the share of content words in a text that is associated with creativity; Segundo-Marcos et al. 2025), lexical diversity (McCarthy, 2005), and divergent semantic integration (Johnson et al., 2023). We also examined how generated verses diverged from the maeku seed verse and each other in terms of their pairwise embedding similarities and lexical overlap. Dictionary-based text analysis (Boyd et al., 2022) was used to show how specific types of content and function words differed between visual and text-based verses. Finally, corpus analysis of a published Bashō poetry collection (Arroyave, 2025) licensed for non-commercial use, syllable accuracy, and LLM judges were used to further contextualize the lexical and embedding measures and explore tradeoffs between modalities.

2 Method

2.1 Design and Poem Generation

The design was a 2 (model family: Gemini or Qwen) x 2 (modality: visual or text linking) factorial. The Gemini family used Gemini 2.5 Flash for text generation and Gemini 2.5 Flash-Image for image generation via the Gemini API (Google, 2025). Qwen family (Qwen Team, 2024) used qwen2.5:7b-instruct for text generation and qwen2.5-vl-72b-instruct for vision to text via the OpenRouter (2023) API. Qwen family generations used Flux-schnell (Black Forest Labs, 2024) for image generation, hosted on Replicate API (Replicate, 2023). LLM/VLM temperature $= 1.2$ / top_p $= 0.95$ were consistent for all models. An earlier version of the paper included a Llama (Llama Team et al., 2024) version which defaulted to using qwen2.5-vl-72b-instruct for vlm in the vision condition due to lack of host of Llama vlm on OpenRouter. Thus, Llama text was excluded as a condition, and vision runs (identical to Qwen) were added to the Qwen condition.

Models were presented with one of six selected maeku verses from Arroyave (2025) and asked to first generate a verse with two seven-syllable stanzas and next generate a 5-7-5 verse to follow that (see Table 2 for all six maeku). In the vision condition, all verses were encoded visually before being decoded by VLM into verse. The model was not given the actual preceding verse in text form at

all. Because this naturally splits up the poem into two generation steps, the same strategy was used for text generation, where each verse was generated with a separate function call and the generating model only given the preceding verse.

In addition to the primary vision/text manipulation, the model was instructed to select a linking type from Bashō's Scent Linking taxonomy: Content links were described as carrying over a concrete object from the scene but allowing for shifts in environment or scale; Mood/emotion links were described as carrying over atmospheric or affective tone but not concrete objects. The model was also asked to choose a valence for the mood links based on the circumplex model (positive, negative, or neutral; Russell & Barrett 1999). Kinetic tone links were described as carrying over focal intensity, speed, or motion vectors and to include an arousal tag (high/low) based on circumplex model. Social connotations links were described as carrying over rank or historical era but not objects.

Text models were instructed: "You are a Renga master. You are given this [direct prior 7-5-7 / linking 7-7] verse: 'verse' → taxonomy block → same 3-step task (1. Select a linking category. 2. Write a CoT on how to transition the scene. 3. Generate the [linking 7-7 / closing 5-7-5] verse." Based on quality analysis of an early pilot, Qwen models were additionally instructed with specific JSON format instructions (Gemini enforced format through `pydantic`; Colvin et al. 2017), told to be concise in CoT, and were given up to 3 re-queries for malformed JSON.

Visual generation instructions were: "A visual manifestation of the verse: 'verse"'. Gemini was given additional aesthetic instructions: "underexposed cinematic film grain, stark flat landscape, low-contrast, analog camera noise, raw." Based on quality analysis of an early pilot in which Flux included non-visual elements, instructions were refocused for reducing extraneous text: "Preserve ambiguity and avoid adding text," and included an explicit, enumerated no-text constraint (no letters, words, numbers, captions, signs, logos, watermarks, calligraphy, pseudo-text/glyphs): "do not use text-like marks to communicate with the image decoder."

Visual language model decoding instructions were similar to text but included additional visual instructions: "Analyze the attached image." Based on quality analysis of an early pilot, Qwen was given additional instructions not to attend to any texts or marks that may be in the image and stay concrete in its relation to the image in the chain of thought trace; it was also explicitly told it had no textual context and must not restore unseen objects.

Models generated 1,458 poems total with complete stanzas and caching of all fields. We excluded 246 cases based on condition (Llama text) and 6 for significantly exceeding syllable count, leaving 1206 poems for analysis: Gemini (240 text; 240 visual) and Qwen (263 text; 463 visual). Social connotation links appeared in only 14 poems and were dropped from link analysis due to sparsity. Six additional verses had missing link accuracy scores. Finally, two verses had lexical diversity calculation failures. This led to an additional 19 poems being removed from linear mixed-effects analysis because of inability to calculate an intercept.

2.2 METRICS

Lexical measures. Per verse metrics included lexical density (% content words / total), lexical diversity (calculated using Measure of Textual Lexical Diversity, MTLD; McCarthy & Jarvis 2010), and divergent semantic integration (DSI; Johnson et al. 2023) from automated-intelligence-tests v0.1.11 (dtzx00, 2026).

Embeddings. Embeddings were computed for each verse using Vertex AI's text-embedding-004 on the Vertex API (Google Cloud, 2024). Embedding distances were computed as the cosine distance between the embeddings for each verse and the maeku and between the two verses. An additional seed-sensitivty margin was calculated as the difference between the average embedding distance of the verse to all other maekus minus its embedding distance to its seed maeku.

Jaccard similarity. Jaccard (1901) similarity was calculated as the percentage of overlapping tokens between each generated verse and the maeku, and between the two generated verses.

LLM judges. We used Gemini 2.5 Flash (Google, 2025) with temperature 0.2 to judge each poem. The model was called once per poem via forced function calling (mode = 'ANY') to assess imagery vividness, how naturalistic the focus of poem was, poetic profundity, linking quality, development

arc, waka (each combination of two sequential verses) clarity, and waka pivot quality (how much second combined verse pivoted from first). All were coded on a 1-5 scale. The LLM also provided two blind verse-link classifications for perceived link type, which was scored as a binary accuracy variable at the verse level.

Syllables. Syllable accuracy was estimated using a custom regex-based vowel clustering heuristic, similar to syllable estimation in classic readability formulas (e.g., Kincaid et al. 1975).

Dictionary-based analysis. We used Linguistic Inquiry and Word Count (LIWC-22; Boyd et al. 2022) dictionaries relevant to high-level stylistic differences to assess writing style differences between visual and textual modalities, focusing on articles and prepositions as indicators of a more analytic, expository style, and verbs, adjectives, perceptual words, and affective language as markers of more dynamic, narrative writing (Boyd et al., 2020). Analytical and dynamic language categories tend to be negatively associated in natural language (Markowitz, 2023; Pennebaker et al., 2014). We converted LIWC percentages to binary scores for categories with $<80\%$ zeros and used percentages for adjectives, which had only 18.1% zeros. All measures were accessed via the Receptiviti Inc. (2022a) API using the `receptiviti` R package (Receptiviti Inc., 2022b).

2.3 Golden corpus

In addition to providing the maeku (see above), we used all translated verses of Bashō's Sumidawara ("A Bundle of Charcoal") from Arroyave (2025) as a "golden dataset" reference for expected metric effects for lexical density, lexical diversity, DSI, LLM Judges, and theme tag distribution. This corpus contained 4 rengas (three 36-verse; one 101-verse) that were used for lagged effects to compare verse 2 and verse 3 relationships to maeku in generated poems. An additional 209 verses were 5-7-5 syllable Hokku in non-renga arrangements. Verses were extracted using a combination of automated scraping and manual cleaning. All metrics were calculated the same as for generated verse.

Table 1: Descriptive lexical statistics for the golden corpus

Measure	M (SD)	n
Density	.59 (.12)	364
Diversity (MTLD)	18.8 (14.3)	364
DSI	.78 (.02)	364
Lag-1 Jaccard Similarity	.06 (.08)	205
Lag-2 Jaccard Similarity	.06 (.04)	201
Lag-1 Embedding Distance	.58 (.08)	205
Lag-2 Embedding Distance	.59 (.06)	201

2.4 Statistical Modeling

For verse-level metrics, we used `mixedLM`, implemented in `statsmodels`, with random intercepts for poem (Seabold & Perktold, 2010). Fixed effects included modality (text vs visual), model (Gemini vs. Qwen), maeku, link type, verse position (2nd or 3rd), and the interactions between modality and link type and modality and verse position. To test the significance of each effect, likelihood ratio tests compared a maximum likelihood fit of the full model to the fit of a reduced model leaving the term for the respective effect out. A second, restricted maximum likelihood fit was used to compute coefficient values; follow-up Wald tests were used for the primary visual versus text comparison and for the visual versus text comparison within each link type.

Poem-level variables were tested using ordinary least squares (OLS) regression, with factors for modality (vision vs. text), link type, maeku, model family, and the interaction between link type and modality. Because our primary hypotheses focused on the main effect of modality, type-II sums of squares were used for calculating the main effects and interactions.

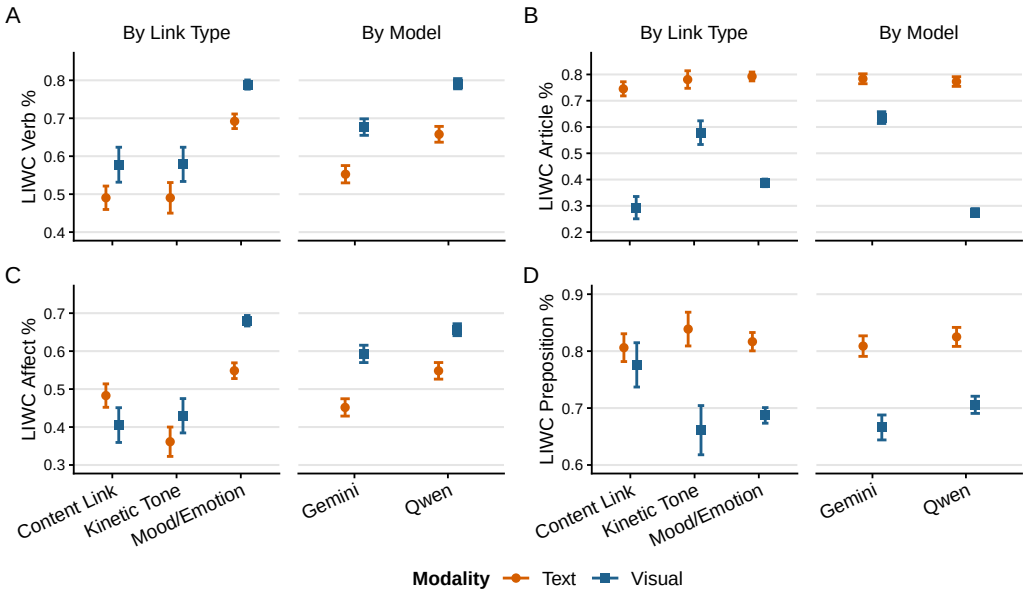

Figure 1: Binary LIWC-22 scores (% of verses containing words in each category) for (A) verbs, (B) articles, (C) affective words, and (D) visual language. Error bars are 95% CIs.

3 Results

3.1 Corpus

Corpus statistics for density, diversity, DSI, and cross-verse overlap statistics are depicted in Table 1. Corpus LLM Judge dimensions are given in Table S3. Within renga sequences in the corpus, semantic similarity between verses tended to decrease from lag-1 to lag-2, $t(395) = 2.05$, $p = .04$, consistent with the expectations to pivot away from the penultimate verse when generating a sequence. The lag difference was not significant for Jaccard, t (315.8) = 1.18, $p = .24$.

Topic tagging revealed that people, plants, and non-biological objects were top three most tagged themes in corpus verses (Table S1). Link classification based on a moving window across renga verses in corpus revealed a predominance of mood (42.2%) and content links (30.5%), followed by social connotations (17.8%), and then kinetic links (9.2%).

3.2 Descriptive Language Differences

Stylistically, text-generated verses tended to have fuller clause structures with articles and prepositions, whereas visual verses were denser, with mostly content words in a montage-like sequence. Canonical Haikai no renga tends to be between these two with fewer full clauses but less pure montage (Table 2).

Content analysis of generated verses revealed stereotypical tendencies to focus on natural elements, sound, and natural scenes. The top 3 themes in generated verses were natural elements (23.1%), time (15.2%) and sound (16.2%). Theme tags differed from corpus considerably across most categories (Table S1). Canonical Haikai no renga also tends to include extensive social content, thus when compared to the corpus, generated verses had much fewer references to people ($\chi^2 = 207.9$, $p < .001$) and non-biological objects ($\chi^2 = 128.2$, $p < .001$), while having substantially more references to natural elements ($\chi^2 = 116.4$, $p < .001$).

These qualitative style and content differences manifested in quantitative differences between modalities. Consistent with its tendency toward montage style, visually generated verses tended to have greater lexical density (0.861 vs. 0.784; Figure 2). However, they also tended to be less

Table 2: Maeku, example verses, and link types

Verse	Model / Corpus	Maeku / Preceding Verse	Judged Link Type
Petals adrift on / The warm water's quiet surface.	Gemini (text)	The high priest leaves / For the hot springs. / Peaking cherry blossoms.	Content
Fields stretch, dawn / Breaks over earth's breath	Qwen (text)	The eastern winds / Carry / The steam of manure.	Mood
Dust trails left behind, Through barren lands, spirits surge, No rest, path unfolds.	Gemini (visual)	I hope they make it in time / For the harvest moon: / A yam field.	Kinetic
By the still pond / A lone crane calls softly	Qwen (text)	In the evening sky / They scatter it: / Clouds on the moon.	Mood
Stone sentinel stands, watching silent gate shut—chest of grass exhales.	Qwen (visual)	Fun is / Shutting the gate / And going quietly to sleep.	Content
Wildfire rages, consuming all.	Gemini (visual)	Amid the scent of plum blossoms, / Boom! / The sun rises.	Kinetic
The little birds from around these parts, They all gather.	Corpus	Perhaps I should rub The gilded Buddha's Delicate feet.	Kinetic
The moon, digging For the yams within the rice.	Corpus	As I unsaddle The errand-running horse, The sun sets.	Content
This year we have had A rainless sixth month.	Corpus	Going back and forth They all look the same Small time peddlers.	Mood
News come from the North And the price of rice climbs.	Corpus	In the lull of spring, House building Begins.	Social

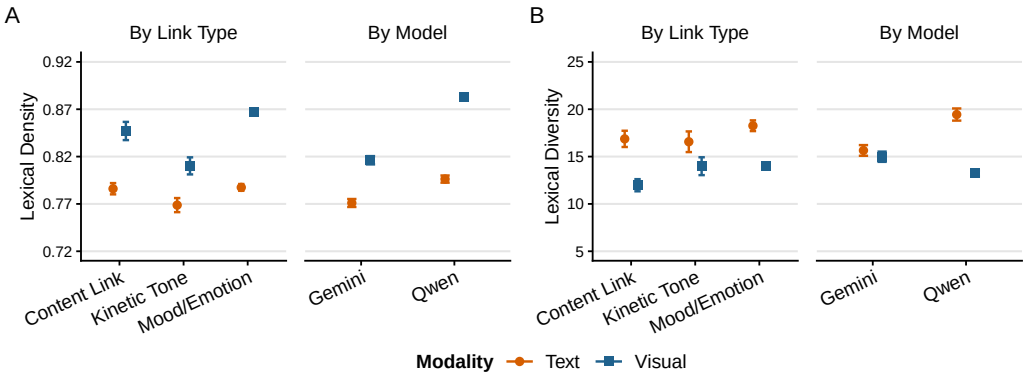

Figure 2: Lexical measures of (A) density, (B) diversity (MTLD) by modality condition, model, and link type. Error bars are 95% CIs.

lexically diverse than text generated verses (13.8 vs. 17.6). DSI did not significantly differ by generation type (0.780 vs. 0.781). Comparatively, the corpus was lower on lexical density than either text or visual (.59), higher in lexical diversity (18.8), and equivalent in DSI (.78; Table 1).

Dictionary-based analyses (Figure 1) showed that text condition poetry included more function words reflecting formal, analytical language such as articles and prepositions, whereas visual condition verses were more likely to include dynamic, narrative words such as verbs and affective language (Pennebaker et al., 2014). Effects of modality controlling for link type, model, and verse position were significant for articles, verbs, and prepositions, all $p < .001$ (Table 3). The condition x link type interaction was significant for prepositions and affective language; for each, condition effects were strongest for mood/emotion links and null for content links (Table S5). See Table S6 for condition x link type model results and descriptive LIWC statistics.

3.3 PIVOT EFFECTS

Visually generated verses tended to have lower Jaccard similarity and greater embedding distance to the maeku (seed) verse than text-generated verses (Table 3, Figure 3). Across both conditions, the the third verse was significantly further from the maeku (.463) in embedding space than the second verse (.426; z = 16.14, $p < .001$); the third verse had significantly lower Jaccard similarity to the maeku (.043) than the second verse (.052; $z = 5.07$, $p < .001$). This makes the generated poems' lag profile consistent with the corpus and the goal to pivot from the maeku in the third verse.

Seed sensitivity margin analysis revealed that poems were more similar to their own maeku than other maekus for both visual and text generated verses, across both verse 2 and 3 (Table 3).

Table 3: Pooled Wald z-tests for modality condition

Variable	Visual M (SD)	Text M (SD)	z	p
Lexical density	.861 (.091)	.784 (.090)	10.8	<.001
Lexical diversity (MTLD)	13.8 (10.1)	17.6 (13.7)	-5.56	<.001
Divergent semantic integration	.780 (.025)	.781 (.018)	0.420	.675
Link classification accuracy	.526 (.499)	.627 (.484)	-1.31	.191
Syllable accuracy (%)	85.6 (14.5)	88.9 (10.9)	-3.54	<.001
Mean embedding distance to other maeku	.512 (.031)	.491 (.031)	12.3	<.001
Min embedding distance to other maeku	.440 (.053)	.415 (.054)	8.02	<.001
Seed sensitivity margin	.058 (.076)	.059 (.080)	-1.72	.086
Embedding distance to maeku	.453 (.076)	.432 (.077)	8.27	<.001
Jaccard sim. to maeku	.041 (.044)	.057 (.043)	-6.49	<.001
LIWC physical binary	.306 (.461)	.348 (.477)	-3.35	.001
LIWC article binary	.396 (.489)	.778 (.416)	-12.3	<.001
LIWC verb binary	.753 (.432)	.607 (.489)	3.20	.001
LIWC affect binary	.635 (.482)	.502 (.500)	1.12	.265
LIWC adj	13.8 (9.5)	11.7 (9.0)	1.97	.049
LIWC visual binary	.567 (.496)	.471 (.499)	0.420	.674
LIWC feeling binary	.327 (.469)	.307 (.462)	-1.30	.193
LIWC auditory binary	.717 (.451)	.528 (.499)	2.60	.009
LIWC lifestyle binary	.157 (.364)	.201 (.401)	-2.43	.015
LIWC prep binary	.692 (.462)	.817 (.387)	-4.70	<.001

Note: Wald z-test contrasts are from each mixed model's coefficients and covariance matrix.

3.4 LLM JUDGES AND SYLLABLE ACCURACY

Evaluation scores tended to show a preference for text-based generation across all dimensions (Table 4). The combined wakas (verse 1 + verse 2; verse 2 + verse 3) that are, in Hakai no renga, intended to work together to create a unit, were rated as clearer in text than vision, and as having a better pivot. Text was rated as having better imagery, development arc, and scent linking quality. Text also tended to score higher on naturalism, which was added to supplement qualitative and content analysis above indicating that generated verse tended to be more stereotypically naturalistic than the corpus. Interestingly, however, generated poems were also rated more highly than the golden data corpus on all dimensions (Table S3) except scent linking quality (3.62 corpus vs 3.44 generated), where the corpus was rated lower than text generated poems (3.79) but not visual (3.18).

Overall, the LLM judges tended to prefer LLM text-based poetry, even over treasured classical poetry by Bashō and colleagues in the golden dataset, and therefore likely illustrates a self-preference biases that LLMs tend to show across diverse texts (e.g., short stories, essays) independent of writing quality (Chen et al., 2025).

Nevertheless, some of the modality-based quality differences are likely partially accurate, as visually generated poems did tend to have numerically lower link classification accuracy (.52 vision vs. .63 text, ns) and significantly lower syllable count accuracy (.86 vs. .89, 3.54, $p < .001$), suggesting that the use of VLM for poetry generation did come at some cost to quality (Table 3).

Table 4: LLM judge ratings for modality conditions

LLM Judge	Visual M	Text M	t	p
Waka clarity (verse 1,2)	3.18	3.91	-12.9	$<.001$
Waka clarity (verse 2,3)	3.80	4.12	-6.54	$<.001$
Imagery vividness	3.73	4.04	-9.87	$<.001$
Naturalism focus	4.06	4.38	-5.58	$<.001$
Poetic profundity	3.63	3.73	-2.90	$<.001$
Scent linking quality	3.19	3.79	-14.8	$<.001$
Development arc	3.71	4.05	-8.96	$<.001$
Waka pivot quality	3.71	4.04	-9.37	$<.001$

Note: Visual n = 703, text n = 503.

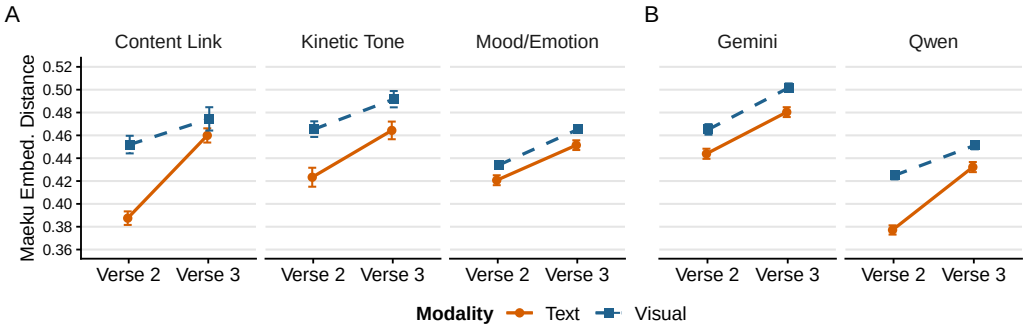

Figure 3: Embedding distance to maeku (seed verses) across modality conditions by (A) link type and (B) model families. Error bars are 95% CIs.

4 Discussion

Visualization is a creativity boosting technique that has been used in poetry writing for centuries across diverse traditions. Here we investigated how poems, generated in a Haikai-no-renga-inspired verse capping (Shirane, 1998) exercise, differed when they were generated by passing visual images across verses instead of text. The results revealed that visually-generated texts tended to be dense and montage-like, often including perceptual, dynamic, narrative, and affect-laden content words. Comparatively, poems generated fully via text tended to have more complete clause structure, with more articles and prepositions. However, visually generated poems may have sacrificed some dimensions of quality related to effective verse-linking strategies and meter (syllable) accuracy.

Although the visually generated poems were clearly differentiated from text-based poems on density and other measures relevant for creativity (Halliday, 1985; Pennebaker et al., 2014; Ure, 1971), both types remained markedly different from historically significant Haikai no renga by Bashō and his colleagues (Arroyave, 2025), which we used as a golden dataset for comparison. Both poem types tended to be much narrower in theme than Bashō's Haikai and more stereotypically focused on natural elements and scenes, whereas Haikai no renga tends to be much more socially focused and contains frequent cross-text allusions (Shirane, 1998). Therefore, none of the results here suggest that LLM-generated renga poetry approaches the level of creativity in the true cultural form or represent substantial progress toward achieving creative parity with humans. Nevertheless, the visual manipulation suggests a promising direction for infusing poetry with different characteristics from pure text-based generation, which may help move toward the more layered, rich textures of real renga poetry. To this end, visualization per se was never intended in the renga tradition, or most poetry traditions, to be a sole technique for generating poetry. Generating poetry from imagery alone may approximate some surrealist techniques (Dalí, 1935), but real renga poets are more likely to merge the unique poetic properties of visualization with other layered textual associations like literary allusions (Shirane, 1998). In that spirit, funneling visualization generated montages into a second text-based verse generation step, instead of generating verses directly from VLM, may

preserve visualizations' impacts on lexical density (as it does in humans Louwerse & Hutchinson 2012), while enabling additional, rich textual associations.

Other studies on poetry and creativity have used images, including one approach with a pipeline specifically designed to generate faithful images over repeated steps (Fang et al., 2025; Liu et al., 2018; Qian et al., 2026) to ensure visual fidelity to the original poem. This differs from our approach of generating single images from single verses, which undoubtedly loses visual qualities that a more iterative pipeline would retain. However, it is not clear whether, for creativity, representing all details of a poem as veridically as possible would correspond to more creative links per se. Real human art, particularly in the visual space, often achieves its effects both through what it foregrounds or makes concrete and what it leaves less articulated. For example, the evocative power of Edvard Munch's (1916) *Man in a Cabbage Field* is enhanced by the contrast between the focal details of the man's face and the broad, ambiguous strokes of the surrounding landscape. Similarly, the mental image of what the poet is trying to convey may benefit from a fast draft with the key linking qualities foregrounded rather than a fully articulated image comprehensively representing an idea.

Notably, although DSI (Johnson et al., 2023) is frequently used as a measure of creative potential in LLMs (including LLM haiku Bellemare-Pepin et al. (2026)), we observed no differences in DSI between our conditions or between LLM generated poems and the golden dataset. DSI is primarily a measure of semantic divergence between tokens within a text. Such divergence does not capture the unique creative aspects of Haikai no renga where, although some linked verse may seem improbable in context, the true creativity arises in creating a unique poetic sense, that often admits interpretation simultaneously across multiple layers of meaning such as on the emotion/mood register while also serving as social commentary. Currently, the LLM creativity field lacks robust ways to operationalize this type of poetic creativity or concretely measure it, as demonstrated by the low scores Bashō et al.'s renga verses received from AI judges relative to LLM verses.

4.1 Limitations

Generation differences across families. There was asymmetry in the instructions to VLM and LLM across model families. These discrepancies were introduced because of challenges in achieving equivalent performance from the models in terms of formatting guideline compliance and avoiding adding or interpreting incidental text on images. However, these differences introduce additional variability in our pipeline. Nonetheless, all statistical models controlled for model family and primary results were robust to family differences, suggesting that these differences likely do not impact our overall conclusions.

Text Size and Metrics. The verses generated in our study had much lower token count compared to many studies of creative writing, including those where measures like DSI and MTTR are often used. Although DSI has previously been used with haiku (Bellemare-Pepin et al. (2026); a component of Haikai renga), it is possible that the small number of words made it less likely to find robust effects with these metrics in the present case.

Broader creativity metrics. The LLM judge's ratings showed self-preference bias and likely do not align with the perspective of human experts. Further, none of our metrics can be interpreted directly as indicating poetic quality or creativity. Future work aiming to not only study the nature of LLM poetry generation but also iteratively write better poetry and improve creativity measurement would require human scoring and likely collaborative generation (Grange et al., 2026; Walser, 2016).

5 Conclusion

Multi-modal techniques such as introducing visualization steps into poetry generation are a promising direction for enhancing LLM creativity. Currently, LLMs not only struggle with creative problem-solving and the divergent reasoning required for novel technical work like scientific discovery, but they do not even approach the layered, multifaceted complexity of poetry and art. Ezra Pound (1934) said, "Great literature is simply language charged with meaning to the utmost possible degree." Given that poetry is literature in its most condensed form, and Haikai no renga is in a sense poetry in its most condensed form (Carter, 1991), the present verse-capping paradigm may prove an effective step toward future benchmarks for creative competence in AI.

AI USE STATEMENT

In this work, we used generative AI tools for code tasks, including co-writing code pipelines and debugging code for Python, R, LaTeX, and APIs; copy-editing and outlining; making qualitative judgments in the LLM judge measure that we report (other qualitative judgments are our own); and aspects of literature review, such as providing entry points to unfamiliar research areas or adjudicating between the best reference for classic works. LLM pipelines included cleaning and formatting of data, particularly for the corpus, and LLMs generated synthetic data (renga verses) in the experimental conditions we describe in the Method. We did not use generative AI tools for writing the final text of the paper, developing theoretical models, translation, or interpreting results (though LLM chat tools sometimes offered opinions unasked while providing help with statistical code or LaTeX formatting). Tasks related to mathematical proofs are not applicable to this work. We have reviewed all AI-assisted work. At least one author manually tested all sections of LLM-generated code for soundness and replicability. We take responsibility for the final content of this work, including potential as-yet-unidentified errors or artifacts that may have been produced through collaboration with generative AI.

ETHICS STATEMENT

Human participants were not involved in this research and thus the study design was not subject to review by an ethics board. We do not see societal or personal ethical risks inherent in the data generated in these experiments, nor are the methods and results produced by our research at special risk of misuse by bad actors. The LLM-generated poems were explicitly influenced by an English translation of an anthology by Bashō. We acknowledge that AI art—especially imitative art derived from individuals' distinct styles—risks harming human creators' reputations and the economic sustainability of their professions (Jiang et al., 2023).

REPRODUCIBILITY STATEMENT

We describe the procedures and models exactly in the Method section, and the shared data and runnable code pipelines (to be shared via a link in the published paper if accepted) enable direct reproduction of our research from data generation to statistical findings. Specific texts of the LLM-generated poetry will differ in replications of this research. The LLM judgments and embedding measures will vary slightly as well even when run on the same texts. Lexical density and diversity measures run on the same texts will produce identical results, as will the deterministic dictionary-based language measures we used (Boyd et al., 2022). We encourage and will be responsive to inquiries regarding replications or extensions of this work.

REFERENCES

Anonymous. I celebrate grass: Parody and social division of labor in multi-agent collaborative poetry. Manuscript under review, authors redacted for blind review, 2026.

Juan Felipe Arroyave. *"Relationships" / "A Bundle of Charcoal, Volume I"*. Ph.d. thesis, University of Washington, 2025. URL `https://washington.edu`. Sumidawara Corpus dataset extracted from ResearchWorks Archive.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A. Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. *Scientific Reports*, 16(1):1279, 2026. doi: 10.1038/s41598-025-25157-3. URL `https://www.nature.com/articles/s41598-025-25157-3`.

Black Forest Labs. Flux.1-schnell, 2024. URL `https://huggingface.co/black-forest-labs/FLUX.1-schnell`.

Ryan L Boyd, Kate G Blackburn, and James W Pennebaker. The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, 6(32):eaba2196, 2020. doi: 10.1126/sciadv.aba2196.

Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. *Linguistic Inquiry and Word Count (LIWC-22)*. Pennebaker Conglomerates, Austin, TX, 2022. URL https://www.liwc.app. Computer software.

Steven D. Carter. *Traditional Japanese Poetry: An Anthology*. Stanford University Press, Stanford, 1991. Translated and edited by Steven D. Carter.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. Can AI writing be salvaged? Mitigating idiosyncrasies and improving human-AI alignment in the writing process through edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–33, 2025.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1653–1672, 2025.

Samuel Colvin, Eric Jolibois, Hasan Ramezani, Adrian Garcia Badaracco, Terrence Dorsey, David Montague, Serge Matveenko, Marcelo Trylesinski, Sydney Runkle, David Hewitt, Alex Hall, and Victorien Plot. Pydantic validation, 2017. URL https://pydantic.dev/docs/validation/latest/get-started/.

Salvador Dalí. *La Conquête de l'irrationnel [The Conquest of the Irrational]*. Éditions Surréalistes, Paris, 1935.

dtzx00. Automated intelligence tests: v0.1.11. https://github.com/dtzx00/Automated-Intelligence-Tests, July 2026. Minimal Python package providing automated tests of associative and creative ability.

Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, et al. Creation-MMBench: Assessing context-aware creative intelligence in MLLMs. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 447–456. IEEE, 2025.

Google. Gemini 2.5 Flash. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash, 2025. Gemini API documentation.

Google Cloud. Vertex AI text embeddings model: text-embedding-004. https://cloud.google.com/vertex-ai/docs/generative-ai/model-reference/text-embeddings, 2024. Accessed: 2026-09-25.

Camille Grange, Théophile Demazure, Mickael Ringeval, Simon Bourdeau, and Cédric Martineau. The human-genai value loop in human-centered innovation: beyond the magical narrative. *Information Systems Journal*, 36(1):29–51, 2026.

Michael A. K. Halliday. *Spoken and Written Language*. Oxford University Press, Oxford, 1985.

Alberto Hernández-Espinosa, Luan Ozelim, Felipe S Abrahão, and Hector Zenil. Superarc: a test for artificial superintelligence based on compressed modelling, recursive prediction and problem complexity. *Nature communications*, 17(1):4885, 2026.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901. doi: 10.5169/SEALS-266450.

Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. AI art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 363–374, 2023.

Dan R Johnson, James C Kaufman, Brendan S Baker, John D Patterson, Baptiste Barbot, Adam E Green, Janet van Hell, Evan Kennedy, Grace F Sullivan, Christa L Taylor, et al. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior research methods*, 55(7):3726–3759, 2023.

Robert P. Kincaid, J. Peter beneath Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. (8-75), February 1975. URL https://apps.dtic.mil/sti/citations/ADA006655. Defense Technical Information Center (DTIC) Accession Number: ADA006655.

Zehan Li, Yutong Zhu, Siyang Wu, Honglin Bao, and James A. Evans. Narrative flattening: How post-training compresses thematic, affective, and stylistic variation in LLM fiction, 2026. URL https://arxiv.org/abs/2605.27878.

Yi-Cheng Lin, Kang-Chieh Chen, Zhe-Yan Li, Tzu-Heng Wu, Tzu-Hsuan Wu, Kuan-Yu Chen, Hung-yi Lee, and Yun-Nung Chen. Creativity in LLM-based multi-agent systems: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27572–27595, 2025.

Bei Liu, Jianlong Fu, Makoto P. Kato, and Masatoshi Yoshikawa. Beyond narrative description: Generating poetry from images by multi-adversarial training. In *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 783–791, 2018.

Llama Team, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Max M. Louwerse and Sterling Hutchinson. Embodied relations in language are structural, visual, and spatial. *Frontiers in Psychology*, 3:247, 2012.

David M Markowitz. Analytic thinking as revealed by function words: What does language really measure? *Applied Cognitive Psychology*, 37(3):643–650, 2023.

Philip M McCarthy. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. PhD thesis, The University of Memphis, 2005.

Philip M McCarthy and Scott Jarvis. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392, 2010.

Edvard Munch. Man in the cabbage field, 1916. Oil on canvas, National Museum of Art, Architecture and Design, Oslo, inventory no. PE.M.00205.

OpenRouter. OpenRouter: A Unified API for Large Language Models, 2023. URL https://openrouter.ai/. Accessed: September 26, 2026.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12): e115844, 2014.

Ezra Pound. A few don'ts by an imagiste. *Poetry*, 1(6):200–206, 1913.

Ezra Pound. *ABC of Reading*. George Routledge 'I&' Sons, London, 1934.

Cheng Qian, Hyeonjeong Ha, Jiayu Liu, Jeonghwan Kim, Emre Can Acikgoz, Bingxuan Li, Kunlun Zhu, Jiateng Liu, Aditi Tiwari, Zhenhailong Wang, et al. Advancing creative physical intelligence in large multimodal models. *arXiv preprint arXiv:2605.26396*, 2026.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org.

Receptiviti Inc. *Receptiviti: Psychological measurement through language analytics*, 2022a. URL https://receptiviti.github.io/receptiviti-r/.

Receptiviti Inc. *receptiviti: Text Analysis Through the Receptiviti API*, 2022b. URL https://receptiviti.github.io/receptiviti-r/.

Replicate. Replicate: Run and push machine learning models with a cloud API, 2023. URL https://replicate.com/. Accessed: September 26, 2026.

Chuang Ruan, Yidong Wang, et al. Evaluating LLMs' divergent thinking capabilities for scientific idea generation. *Nature Communications*, 17(1):70245, 2026. doi: 10.1038/s41467-026-70245-1. URL https://nature.com.

James A. Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76 (5):805–819, 1999. doi: 10.1037/0022-3514.76.5.805. URL https://doi.org.

Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, pp. 92–96, 2010. doi: 10.25080/Majora-92bf1922-011.

Rafael Segundo-Marcos, Verónica López Fernández, and María Teresa Daza González. Creative thinking and lexical richness development and the impact of cooperative learning over time. *The Journal of Creative Behavior*, 59(4):e70054, 2025.

Haruo Shirane. *Traces of Dreams: Landscape, Cultural Memory, and the Poetry of Bashō*. Stanford University Press, 1998.

Fang-Yi Su, Ching Chieh Tsao, and Jung-Hsien Chiang. [TINY] building bridges of thought: Using the power of association to inspire creativity in large language models. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025.

Jean Ure. Lexical density and register differentiation. In G. E. Perren and J. L. M. Trim (eds.), *Applications of Linguistics*. Cambridge University Press, Cambridge, 1971.

Hannah Walser. Humanities without humanists. Stanford Literary Lab (*Techne*), August 2016. URL https://litlab.stanford.edu/humanities-without-humanists/. Accessed: 2026-08-24.

Dawei Wang, Difang Huang, Haipeng Shen, and Brian Uzzi. A large-scale comparison of divergent creativity in humans and large language models. *Nature Human Behaviour*, 10(3):531–540, March 2026. doi: 10.1038/s41562-025-02331-1. URL https://nature.com.

E. Wenger and Y. N. Kenett. Large language models are homogeneously creative. *PNAS Nexus*, 5 (3):pgag042, 2026. doi: 10.1093/pnasnexus/pgag042. URL https://oup.com.

A APPENDIX

Table S1: Tagged themes in golden data corpus and generated poems

Category	Corpus count (%)	Generated count (%)	χ^2	p
Animal	54 (7.4%)	126 (2.6%)	45.0	<.001
Biological_motion	66 (9.1%)	323 (6.7%)	5.10	.024
Elements	41 (5.6%)	1116 (23.1%)	116	<.001
Food_or_beverage	34 (4.7%)	40 (0.8%)	68.1	<.001
Non_biological_motion	28 (3.8%)	522 (10.8%)	33.7	<.001
Non_biological_object	100 (13.7%)	182 (3.8%)	128	<.001
Other	1 (0.1%)	4 (0.1%)	0.00	1.00
Person	146 (20.1%)	251 (5.2%)	208	<.001
Plant	117 (16.1%)	640 (13.3%)	3.99	.046
Sound	38 (5.2%)	781 (16.2%)	59.7	<.001
Temperature	19 (2.6%)	106 (2.2%)	0.320	.572
Time	84 (11.5%)	733 (15.2%)	6.45	.011

Note: P-values are not corrected for false discovery rates.

Table S2: Judged link types in golden data corpus

Link Type	Count (%)
Content link	122 (30.5%)
Kinetic tone	37 (9.2%)
Mood/emotion	169 (42.2%)
Social connotations	71 (17.8%)
Unclassifiable	1 (0.2%)

Table S3: LLM-judged dimensions: Golden corpus compared to mean generated verses

Dimension	Corpus M	Generated M	t	p
Imagery_vividness	3.27	3.86	-11.3	.000***
Naturalism_focus	2.23	4.19	-23.0	.000***
Poetic_profundity	3.32	3.67	-6.01	.000***
Scent_linking_quality	3.62	3.44	3.08	.002**
Development_arc	3.85	3.86	-0.215	.830
Waka_pivot_quality	3.55	3.85	-4.85	.000***
Waka_clarity_1_2	2.54	3.48	-11.4	.000***
Waka_clarity_2_3	2.41	3.93	-18.7	.000***

Note: Corpus n = 200, Generated n = 1,206. * p<.05, ** p<.01, *** p<.001.

Table S4: Mixed-model likelihood-ratio omnibus tests for per-stanza variables repeated within poems: Lexical variables

Variable	Term	χ^2	df	p
density	modality	-0.000	0	–
density	pivot_position	-0.000	0	–
density	link_type_norm	-0.000	0	–
density	maeku	10.415	5	0.064
density	model_family	115.611	1	0.000***
density	modality * pivot_position	41.994	1	0.000***
density	modality * link_type_norm	15.591	2	0.000***
diversity	modality	0.000	0	–
diversity	pivot_position	0.000	0	–
diversity	link_type_norm	0.000	0	–
diversity	maeku	9.203	5	0.101
diversity	model_family	1.123	1	0.289
diversity	modality * pivot_position	0.689	1	0.407
diversity	modality * link_type_norm	3.801	2	0.149
dsi†	modality	-0.000	0	–
dsi†	pivot_position	0.000	0	–
dsi†	link_type_norm	-0.000	0	–
dsi†	maeku	19.312	5	0.002**
dsi†	model_family	30.820	1	0.000***
dsi†	modality * pivot_position	6.911	1	0.009**
dsi†	modality * link_type_norm	0.929	2	0.628
link_match	modality	-0.000	0	–
link_match	pivot_position	-0.000	0	–
link_match	link_type_norm	0.000	0	–
link_match	maeku	2.687	5	0.748
link_match	model_family	69.630	1	0.000***
link_match	modality * pivot_position	0.213	1	0.644
link_match	modality * link_type_norm	11.001	2	0.004**
syllable_accuracy	modality	0.000	0	–
syllable_accuracy	pivot_position	0.000	0	–
syllable_accuracy	link_type_norm	0.000	0	–
syllable_accuracy	maeku	9.338	5	0.096
syllable_accuracy	model_family	91.496	1	0.000***
syllable_accuracy	modality * pivot_position	39.359	1	0.000***
syllable_accuracy	modality * link_type_norm	3.624	2	0.163
other_maeku_dist_avg	modality	0.000	0	–
other_maeku_dist_avg	pivot_position	0.000	0	–
other_maeku_dist_avg	link_type_norm	-0.000	0	–
other_maeku_dist_avg	maeku	151.066	5	0.000***
other_maeku_dist_avg	model_family	99.792	1	0.000***
other_maeku_dist_avg	modality * pivot_position	0.156	1	0.693
other_maeku_dist_avg	modality * link_type_norm	7.179	2	0.028*
other_maeku_dist_min	modality	-0.000	0	–
other_maeku_dist_min	pivot_position	-0.000	0	–
other_maeku_dist_min	link_type_norm	-0.000	0	–
other_maeku_dist_min	maeku	330.780	5	0.000***
other_maeku_dist_min	model_family	74.729	1	0.000***
other_maeku_dist_min	modality * pivot_position	0.102	1	0.749
other_maeku_dist_min	modality * link_type_norm	7.519	2	0.023*
seed_sensitivity_margin	modality	0.000	0	–
seed_sensitivity_margin	pivot_position	0.000	0	–
seed_sensitivity_margin	link_type_norm	0.000	0	–
seed_sensitivity_margin	maeku	1069.615	5	0.000***
seed_sensitivity_margin	model_family	126.416	1	0.000***
seed_sensitivity_margin	modality * pivot_position	2.534	1	0.111
seed_sensitivity_margin	modality * link_type_norm	32.539	2	0.000***
maeku_distance	modality	-0.000	0	–
maeku_distance	pivot_position	0.000	0	–
maeku_distance	link_type_norm	-0.000	0	–
maeku_distance	maeku	910.140	5	0.000***
maeku_distance	model_family	263.684	1	0.000***
maeku_distance	modality * pivot_position	3.119	1	0.077
maeku_distance	modality * link_type_norm	44.334	2	0.000***
maeku_jaccard†	modality	0.000	0	–
maeku_jaccard†	pivot_position	-0.000	0	–
maeku_jaccard†	link_type_norm	0.000	0	–
maeku_jaccard†	maeku	251.390	5	0.000***
maeku_jaccard†	model_family	14.517	1	0.000***
maeku_jaccard†	modality * pivot_position	4.884	1	0.027*
maeku_jaccard†	modality * link_type_norm	0.893	2	0.640

Note: * p<.05, ** p<.01, *** p<.001. † random-intercept variance estimated at or near zero (singular fit). Modality effects are from an omnibus test of whether modality-related terms jointly improve fit, not a direct condition contrast (modality participates in two interactions); see the Wald z-test tables for that comparison.

Table S5: Mixed-model likelihood-ratio omnibus tests for per-stanza variables repeated within poems: LIWC variables

Variable	Term	χ^2	df	p
liwc22_Physical_binary	modality	0.000	0	–
liwc22_Physical_binary	pivot_position	0.000	0	–
liwc22_Physical_binary	link_type_norm	-0.000	0	–
liwc22_Physical_binary	maeku	20.060	5	0.001**
liwc22_Physical_binary	model_family	9.223	1	0.002**
liwc22_Physical_binary	modality * pivot_position	11.385	1	0.001***
liwc22_Physical_binary	modality * link_type_norm	15.872	2	0.000***
liwc22_article_binary	modality	0.000	0	–
liwc22_article_binary	pivot_position	0.000	0	–
liwc22_article_binary	link_type_norm	0.000	0	–
liwc22_article_binary	maeku	10.788	5	0.056
liwc22_article_binary	model_family	91.278	1	0.000***
liwc22_article_binary	modality * pivot_position	20.610	1	0.000***
liwc22_article_binary	modality * link_type_norm	13.252	2	0.001**
liwc22_verb_binary†	modality	0.000	0	–
liwc22_verb_binary†	pivot_position	0.000	0	–
liwc22_verb_binary†	link_type_norm	-0.000	0	–
liwc22_verb_binary†	maeku	28.759	5	0.000***
liwc22_verb_binary†	model_family	23.066	1	0.000***
liwc22_verb_binary†	modality * pivot_position	5.648	1	0.017*
liwc22_verb_binary†	modality * link_type_norm	0.681	2	0.712
liwc22_Affect_binary	modality	-0.000	0	–
liwc22_Affect_binary	pivot_position	0.000	0	–
liwc22_Affect_binary	link_type_norm	-0.000	0	–
liwc22_Affect_binary	maeku	28.380	5	0.000***
liwc22_Affect_binary	model_family	0.788	1	0.375
liwc22_Affect_binary	modality * pivot_position	14.212	1	0.000***
liwc22_Affect_binary	modality * link_type_norm	17.516	2	0.000***
liwc22_adj	modality	0.000	0	–
liwc22_adj	pivot_position	0.000	0	–
liwc22_adj	link_type_norm	-0.000	0	–
liwc22_adj	maeku	5.613	5	0.346
liwc22_adj	model_family	157.630	1	0.000***
liwc22_adj	modality * pivot_position	3.181	1	0.074
liwc22_adj	modality * link_type_norm	16.589	2	0.000***
liwc22_visual_binary†	modality	0.000	0	–
liwc22_visual_binary†	pivot_position	0.000	0	–
liwc22_visual_binary†	link_type_norm	-0.000	0	–
liwc22_visual_binary†	maeku	33.662	5	0.000***
liwc22_visual_binary†	model_family	7.139	1	0.008**
liwc22_visual_binary†	modality * pivot_position	0.156	1	0.693
liwc22_visual_binary†	modality * link_type_norm	17.525	2	0.000***
liwc22_feeling_binary	modality	0.000	0	–
liwc22_feeling_binary	pivot_position	0.000	0	–
liwc22_feeling_binary	link_type_norm	0.000	0	–
liwc22_feeling_binary	maeku	29.371	5	0.000***
liwc22_feeling_binary	model_family	20.002	1	0.000***
liwc22_feeling_binary	modality * pivot_position	0.079	1	0.778
liwc22_feeling_binary	modality * link_type_norm	6.001	2	0.050*
liwc22_auditory_binary†	modality	0.000	0	–
liwc22_auditory_binary†	pivot_position	0.000	0	–
liwc22_auditory_binary†	link_type_norm	0.000	0	–
liwc22_auditory_binary†	maeku	93.335	5	0.000***
liwc22_auditory_binary†	model_family	50.978	1	0.000***
liwc22_auditory_binary†	modality * pivot_position	0.593	1	0.441
liwc22_auditory_binary†	modality * link_type_norm	3.431	2	0.180
liwc22_Lifestyle_binary	modality	-0.000	0	–
liwc22_Lifestyle_binary	pivot_position	-0.000	0	–
liwc22_Lifestyle_binary	link_type_norm	-0.000	0	–
liwc22_Lifestyle_binary	maeku	10.789	5	0.056
liwc22_Lifestyle_binary	model_family	1.826	1	0.177
liwc22_Lifestyle_binary	modality * pivot_position	3.423	1	0.064
liwc22_Lifestyle_binary	modality * link_type_norm	1.639	2	0.441
liwc22_prep_binary	modality	0.000	0	–
liwc22_prep_binary	pivot_position	-0.000	0	–
liwc22_prep_binary	link_type_norm	-0.000	0	–
liwc22_prep_binary	maeku	6.065	5	0.300
liwc22_prep_binary	model_family	3.796	1	0.051
liwc22_prep_binary	modality * pivot_position	8.028	1	0.005**
liwc22_prep_binary	modality * link_type_norm	7.632	2	0.022*

Note: * p<.05, ** p<.01, *** p<.001. † random-intercept variance estimated at or near zero (singular fit). Modality effects are from an omnibus test of whether modality-related terms jointly improve fit, not a direct condition contrast (modality participates in two interactions); see the Wald z-test tables for that comparison.

Table S6: Pooled Wald z-tests for modality condition x link type

Variable	Content			Kinetic			Mood		
	Visual M (SD)	Text M (SD)	z	Visual M (SD)	Text M (SD)	z	Visual M (SD)	Text M (SD)	z
Lexical Density	.847 (.104)	.786 (.0960)	4.66***	.810 (.0990)	.769 (.0930)	3.94***	.867 (.0870)	.787 (.0860)	17.0***
Lexical Diversity (MTLD)	12.0 (6.91)	16.9 (13.9)	-4.19***	14.0 (10.4)	16.6 (13.6)	-1.21	14.0 (10.3)	18.3 (13.6)	-6.74***
Divergent Semantic Integration	.789 (.029)	.786 (.021)	0.768	.780 (.018)	.778 (.015)	0.134	.779 (.025)	.779 (.018)	-0.572
Link Classification Accuracy	.707 (.457)	.897 (.304)	-2.92**	.388 (.489)	.284 (.452)	1.79	.522 (.50)	.595 (.491)	-2.33*
Syllable Accuracy (%)	82.8 (18.7)	86.2 (11.7)	-2.83**	89.6 (9.24)	91.3 (9.82)	-0.255	85.4 (14.4)	89.5 (10.5)	-5.12***
Mean Other-Maeku Embedding Dist.	.516 (.036)	.488 (.032)	8.23***	.519 (.036)	.509 (.029)	3.88***	.511 (.030)	.488 (.029)	14.9***
Min Other-Maeku Embedding Dist.	.449 (.056)	.416 (.050)	5.16***	.439 (.055)	.437 (.047)	2.02*	.439 (.052)	.408 (.057)	11.5***
Seed Sensitivity Margin	.0530 (.062)	.078 (.081)	-5.31***	.0410 (.056)	.0610 (.074)	1.79	.061 (.079)	.0500 (.080)	0.956
Embedding Distance to Maeku	.462 (.068)	.410 (.080)	9.63***	.478 (.056)	.448 (.074)	0.216	.450 (.078)	.438 (.075)	7.13***
Jaccard Sim to Maeku	.0390 (.049)	.064 (.051)	-4.03***	.0340 (.035)	.0510 (.041)	-2.39*	.0410 (.045)	.0550 (.039)	-7.41***
LIWC Physical Binary	.224 (.419)	.433 (.496)	-3.82***	.231 (.423)	.316 (.466)	-1.60	.322 (.467)	.318 (.466)	0.760
LIWC Article Binary	.293 (.457)	.745 (.437)	-7.79***	.579 (.496)	.781 (.415)	-3.19**	.387 (.487)	.792 (.406)	-16.7***
LIWC Verb Binary	.578 (.496)	.490 (.501)	1.18	.579 (.496)	.490 (.502)	1.54	.789 (.408)	.692 (.462)	4.51***
LIWC Affect Binary	.405 (.493)	.483 (.501)	-2.11*	.430 (.497)	.361 (.482)	1.25	.680 (.467)	.548 (.498)	5.30***
LIWC Adj	10.6 (9.10)	12.1 (9.10)	-0.764	11.9 (7.80)	11.6 (8.40)	0.496	14.3 (9.60)	11.6 (9.10)	7.09***
LIWC Visual Binary	.517 (.502)	.445 (.498)	1.18	.289 (.455)	.465 (.500)	-2.50*	.601 (.490)	.484 (.500)	4.77***
LIWC Feeling Binary	.233 (.424)	.304 (.461)	-1.53	.107 (.311)	.206 (.406)	-1.10	.360 (.480)	.336 (.473)	1.58
LIWC Auditory Binary	.379 (.487)	.266 (.443)	1.58	.355 (.481)	.381 (.487)	0.133	.790 (.408)	.687 (.464)	4.97***
LIWC Lifestyle Binary	.129 (.337)	.224 (.418)	-2.25*	.124 (.331)	.148 (.357)	-0.432	.163 (.370)	.204 (.403)	-2.26*
LIWC Prep Binary	.776 (.419)	.806 (.396)	-0.273	.661 (.475)	.839 (.369)	-3.80***	.687 (.464)	.817 (.387)	-6.17***

Wald z-tests contrast modality within each link type (marginal over verse position) built from each fitted mixed model. Dist. = distance. Content visual n = 116, text n = 263; Kinetic visual n = 121, text n = 155; Mood/emotion visual n = 1141, text $n = 578$. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table S7: Poem-level F tests

Variable	Term	SS	df	F	p
pivot1_pivot2_distance	modality)	0.590	1.0	92.981	0.000***
pivot1_pivot2_distance	link_type_2_norm)	0.103	2.0	8.120	0.000***
pivot1_pivot2_distance	maeku)	0.270	5.0	8.505	0.000***
pivot1_pivot2_distance	model_family)	0.755	1.0	119.063	0.000***
pivot1_pivot2_distance	modality * link_type_2_norm)	0.152	2.0	11.958	0.000***
pivot1_pivot2_distance	Residual	7.556	1191.0	–	–
pivot1_pivot2_jaccard	modality)	0.196	1.0	52.301	0.000***
pivot1_pivot2_jaccard	link_type_2_norm)	0.004	2.0	0.487	0.615
pivot1_pivot2_jaccard	maeku)	0.058	5.0	3.072	0.009**
pivot1_pivot2_jaccard	model_family)	0.027	1.0	7.160	0.008**
pivot1_pivot2_jaccard	modality * link_type_2_norm)	0.041	2.0	5.411	0.005**
pivot1_pivot2_jaccard	Residual	4.465	1191.0	–	–
waka_clarity_1_2	modality)	85.706	1.0	105.415	0.000***
waka_clarity_1_2	link_type_1_norm)	42.967	2.0	26.424	0.000***
waka_clarity_1_2	maeku)	104.607	5.0	25.733	0.000***
waka_clarity_1_2	model_family)	0.866	1.0	1.065	0.302
waka_clarity_1_2	modality * link_type_1_norm)	3.097	2.0	1.904	0.149
waka_clarity_1_2	Residual	961.814	1183.0	–	–
waka_clarity_2_3	modality)	19.599	1.0	26.601	0.000***
waka_clarity_2_3	link_type_1_norm)	1.635	2.0	1.110	0.330
waka_clarity_2_3	maeku)	9.327	5.0	2.532	0.027*
waka_clarity_2_3	model_family)	2.127	1.0	2.887	0.090
waka_clarity_2_3	modality * link_type_1_norm)	2.832	2.0	1.922	0.147
waka_clarity_2_3	Residual	871.595	1183.0	–	–
imagery_vividness	modality)	14.993	1.0	59.077	0.000***
imagery_vividness	link_type_1_norm)	8.066	2.0	15.891	0.000***
imagery_vividness	maeku)	34.613	5.0	27.277	0.000***
imagery_vividness	model_family)	2.037	1.0	8.028	0.005**
imagery_vividness	modality * link_type_1_norm)	0.600	2.0	1.182	0.307
imagery_vividness	Residual	300.231	1183.0	–	–
naturalism_focus	modality)	33.508	1.0	53.261	0.000***
naturalism_focus	link_type_1_norm)	3.350	2.0	2.662	0.070
naturalism_focus	maeku)	333.250	5.0	105.940	0.000***
naturalism_focus	model_family)	82.267	1.0	130.763	0.000***
naturalism_focus	modality * link_type_1_norm)	5.019	2.0	3.989	0.019*
naturalism_focus	Residual	744.262	1183.0	–	–
poetic_profundity	modality)	0.274	1.0	0.865	0.353
poetic_profundity	link_type_1_norm)	1.253	2.0	1.976	0.139
poetic_profundity	maeku)	22.205	5.0	14.000	0.000***
poetic_profundity	model_family)	25.525	1.0	80.466	0.000***
poetic_profundity	modality * link_type_1_norm)	1.359	2.0	2.142	0.118
poetic_profundity	Residual	375.258	1183.0	–	–
scent_linking_quality	modality)	69.794	1.0	133.293	0.000***
scent_linking_quality	link_type_1_norm)	1.691	2.0	1.614	0.199
scent_linking_quality	maeku)	9.906	5.0	3.784	0.002**
scent_linking_quality	model_family)	34.677	1.0	66.226	0.000***
scent_linking_quality	modality * link_type_1_norm)	4.624	2.0	4.416	0.012*
scent_linking_quality	Residual	619.434	1183.0	–	–
development_arc	modality)	14.289	1.0	31.002	0.000***
development_arc	link_type_1_norm)	5.542	2.0	6.013	0.003**
development_arc	maeku)	13.468	5.0	5.844	0.000***
development_arc	model_family)	42.347	1.0	91.880	0.000***
development_arc	modality * link_type_1_norm)	5.999	2.0	6.508	0.002**
development_arc	Residual	545.245	1183.0	–	–
waka_pivot_quality	modality	15.533	1.0	36.750	0.000***
waka_pivot_quality	link_type_1_norm	2.810	2.0	3.324	0.036*
waka_pivot_quality	maeku	12.510	5.0	5.919	0.000***
waka_pivot_quality	model_family	32.541	1.0	76.991	0.000***
waka_pivot_quality	modality * link_type_1_norm	5.620	2.0	6.648	0.001**
waka_pivot_quality	Residual	500.010	1183.0	–	–

Type II ANOVA, one-per-poem outcomes. * $p < .05$, ** $p < .01$, *** $p < .001$.