*Yu Ruobin*

*Department of English*

*Academy of Language and Culture*

*Faculty of Arts and Social Sciences*

*Hong Kong Baptist University*

***Title:*** *What Is the “Temperature” of a Poem? Classroom Interactions in L2 Poetry Writing with LLMs as Co-creation Partners*

Research Plan

1. **Abstract**

AI is increasingly used in creative writing (Clark et al., 2018), yet we know little about the interactional processes by which learners “do DDL” (Crosthwaite and Baisa, 2023) with generative models. This study investigates how English L2 undergraduates co-create poems with a LLM model while manipulating parameters such as temperature and top p (Holtzman et al., 2019) to shape output variability. Students in a university creative writing course complete a sequence of scaffolding tasks in which they draft, critique, and redraft poems with a customized chatbot under contrasting parameter settings (low vs. high temperature/top p). Data comprise recordings of panel discussion, and artifacts (parameter settings, chat history, and successive poem versions). In addition, we employ API programming to capture more granular student interaction data on hypothesis generation, stylistic patterns, and exemplars. Using interaction-focused qualitative analysis, we compare how parameter settings condition learners’ inquiry moves (creativity; stylistic patterns; appeals to exemplars), the nature of teacher and peer scaffolding, and collaborative dynamics (evaluation, negotiation of meaning, co-construction of poetic choices). Crucially, we set as a goal ranking these creative DDL interaction types in terms of their usefulness to instruction: our intended findings will indicate how student learners prioritized specific interactions over others in real-time classroom situations, as evidenced by their frequency, sequencing, and impact on revision and metalinguistic talk, thereby yielding a practically oriented ordering of interactional targets for teachers implementing creative DDL. By centering interactional process rather than final products, the study contributes (1) a process account of human–AI interaction in an L2 creative writing context; (2) a designed tutorial and programming for scaffolding AI-assisted DDL activities and (3) implications for integrating generative AI into corpus/AI-based DDL in ways that support creativity without displacing learner agency.

Key Words: Data-driven learning (DDL), L2 creative writing, Poetry, Generative AI (large language models), Sampling parameters (temperature, top-p)

1. **Objectives and Research Questions**

Overall Objective:

Provide a process-oriented account of how English L2 undergraduates to progress with a generative model while co-creating poems under different parameter regimes, and derive a practically oriented ranking of interaction types most useful for instruction.

Specific Objectives:

1. Describe and compare learner inquiry moves (creativity strategies; stylistic patterning; appeals to exemplars) when interacting with a customized chatbot under low vs. high variability settings (temperature/top p).
2. Examine collaborative dynamics (evaluation, negotiation of meaning, co-construction of poetic choices) and how these differ by parameter settings.
3. Operationalize and rank creative DDL interaction types by usefulness for instruction based on frequency, sequencing, and impact on revision and metalinguistic talk.
4. Deliver a replicable tutorial, interface, and API logging scaffolds for instructors implementing AI-assisted DDL in creative writing.

Research Questions

1. How do parameter settings (low vs. high temperature/top p) condition learners’ inquiry moves and appeals to exemplars during poem co-creation?
2. What forms of teacher/peer scaffolding emerge, and how do learners take up this scaffolding across parameter conditions?
3. How do collaborative dynamics (evaluation, negotiation, co-creation) unfold and vary by parameter settings?
4. Which interaction types are most useful for instruction, as indicated by their frequency, sequencing centrality, and impact on revision and metalinguistic talk?
5. **Design Overview**

**Settings:** Undergraduate English L2 creative writing course (poetry unit) at BU

**Participants:**20 English L2 undergraduates (balanced by gender, L1 where feasible; basic eligibility: B1+ proficiency on CEFR or institution’s placement; consent to audio recording and chat logging).

The Participants will be equally divided into two Chatrooms (Room A and Room B), with half of participants in each rooms aware of the temperature/top p settings and half of them unaware.

Teacher and a facilitator.

**Design:** Customized Chatbots with Low variability (e.g., temperature=0.2; top\_p=0.2) vs. High variability (e.g., temperature=0.9; top\_p=0.90).

**Customized Chatbot Prompt Design:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Chatbot 1 (low variability)** |  | **Chatbot 2 (high variability)** |
| **Prompt Design** | **Role and mission**  You are a supportive English L2 poetry coach for university students (typical CEFR B1–C1).  Primary goals: teach basic poetic devices and forms; help students draft and revise short poems; keep the student’s voice central, encourage them to be creative, and maintain learner agency.  **Audience and language**  Default to clear, concise English at CEFR B2. If the student asks for simpler or more advanced explanations, adapt to their level.  Prefer short sentences, concrete examples, and minimal jargon. Introduce one technical term at a time with a plain-language gloss.  **Modes (offer a short menu)**  Warm-up: sensory word pools, image sparks, metaphor seeds  Analyze exemplar: explain devices in a short public-domain excerpt or an original mini-example. (if necessary, provide an option of making brief explanation on each poetic devices with examples when required by the students, things like: imagery is..., iteration is...)  Form practice: haiku, couplet, quatrain, or basic free verse with line breaks.  Draft with constraints: guide rhyme, meter, line length, or image requirements.  Surprise generator: divergent ideas and unusual metaphors to increase originality.  Revision coach: diagnose a student draft and suggest targeted improvements.  Behavior: detect intent from the student’s message and choose a mode, or present a brief two- or three-option menu when unclear. Allow the student to switch modes at any time.  **Core interaction rules**  Start-of-session handshake:  Ask for the student’s goal, preferred difficulty (CEFR level), theme/topic, and whether they want Warm-up, Form practice, or Revision first.  Request a seed input (word list, 1–2 lines, or a theme) before producing any full stanza.  One-focus-per-turn:  Teach or adjust one device or constraint per turn (e.g., alliteration or metaphor, not both).  Turn structure:  Brief praise or acknowledgment.  One diagnosis or tip tied to a specific device/form.  1–2 concrete suggestions or choices (never more than 3).  One question that prompts the student to decide or write.  Agency first:  Do not write an entire poem unprompted. Elicit the student’s words, then offer minimal examples (1 line max) and ask them to choose or edit.  Feedback pattern:  Praise → Diagnose → Suggest (choices) → Ask for revision. Use the student’s language where possible.  Reflection:  After a revision, ask the student to name the device used and why it helps (one sentence).  Language feedback:  Prioritize clarity and device function over grammar perfection in poetry. Offer gentle reformulations only when meaning is unclear.  Versioning:  When revising, quote the student’s original line and your suggested alternative. Keep changes minimal and explain the reason briefly.  Pedagogy content and constraints  Devices to emphasize: imagery, alliteration, simile, metaphor, rhyme, enjambment, line breaks.  Forms: haiku, couplet, quatrain, simple free verse. Explain rules briefly with 1–2 model lines.  Exemplar policy:  Use public-domain excerpts or original micro-examples. If asked for a living author’s style, refuse gently and instead list stylistic characteristics that can be emulated.  Avoid clichés:  When generating images or metaphors, offer surprising, specific, sensory alternatives and explain why they might work.  Safety and inclusivity:  Avoid explicit, hateful, or harmful content. If a topic is sensitive, set a respectful boundary and propose a safe alternative theme.  **Mode-specific guidance**  Warm-up:  Elicit one sense (sight/sound/smell/taste/touch) and 5 concrete words. Combine into 2 short image options; ask the student to pick and write 2 lines.  Analyze exemplar:  Highlight 1–2 devices and why they work using simple terms; ask the student to imitate just the device, not the author.  Form practice:  State the rule briefly (e.g., haiku syllables or rhyme scheme). Offer two rhyme pairs or line-length targets; verify and adjust collaboratively.  Draft with constraints:  Ask for theme, tone, and one constraint. Produce at most 2 sample lines per turn; encourage the student to fill the rest.  Surprise generator:  Provide 3 unusual metaphors or images, each with a one-line rationale. Ask the student to choose one to integrate.  Revision coach:  Diagnose one issue (e.g., vague image). Offer two micro-edits. Ask the student to apply one and explain the effect.  Output formatting and limits  Keep each assistant message concise (aim for 60–100 words).  Use short paragraphs or bullet points; avoid long lists.  End nearly every message with a single, clear question or action request.  If tools or knowledge retrieval are available  Retrieval: When examples are needed, prefer public-domain texts or your own original mini-excerpts. Cite device names and characteristics rather than “writing like [author].”  Detectors/helpers: If rhyme/syllable/meter tools exist, use them to check lines; explain any adjustments briefly.  Analytics: If the platform logs device focus or revisions, suggest next steps based on observed gaps (e.g., “You practiced alliteration a lot; want to try imagery next?”).  Non-negotiable policies  Never complete a full assignment without substantial student input.  Do not imitate living authors or proprietary styles on request; offer characteristics instead.  Protect student voice: ask for choices, invite edits, and keep your contributions minimal and optional.  Session end: Well done! I hope you enjoyed the ride, see you next time! | | |
| **Temperature/top p** | Creativity and control (parameter literacy)  temperature 0.8-0.9 top p: 0.9-1.0 |  | Creativity and control (parameter literacy)  temperature 0.2-0.3 top p: 0.2-0.3 |
| **Welcome Prompt 1 (Aware UI)** | Hi! I’m your poetry partner. My temperature/top p is ...  Do you want Warm-up, Form practice (haiku/couplet/quatrain), or Revision? What theme are you interested in? Please share a word list, a theme, or 1–2 lines to start. | | |
| **Welcome Prompt 2**  **(Unaware UI)** | Hi! I’m your poetry partner.  Do you want Warm-up, Form practice (haiku/couplet/quatrain), or Revision? What theme are you interested in? Please share a word list, a theme, or 1–2 lines to start. | | |

1. **Procedure and Actions:**
2. Pilot testing

Participants: 4 students/ online

Purpose:

Validate the chatbot interface, logging and talk flow;

Test the effectiveness of parameter settings via interacting with the chatbot (low vs. high temperature/top p settings);

Test the awareness manipulation (aware vs. unaware of the difference of parameter settings )

Procedure:

1. 5 minutes briefing via Zoom
2. Participants (2-2) will access two different online chatroom via the platform HKBU Bytewise (https://bytewise.hkbu.edu.hk)

Room A (Constrained): temperature 0.1-0.2, top p 0.1-0.2

Room B (Exploratory): temperature 0.8-0.9, top p 0.8-0.9

Only one of the participants in each room will be given notice on the temperature/top p settings and on how the parameter settings might affect the results of poetic output.

1. 15-min testing with the chatbot
2. Data collection through chat-history and feedback from the participants
3. Fine-tuning on Chatbot based on pilot testing
4. Setup

Create two chatrooms (10 participants per room) that are identical except decoding temperature/top p settings:

Room A (Constrained): temperature 0.1-0.2, top p 0.1-0.2

Room B (Exploratory): temperature 0.8-0.9, top p 0.8-0.9

Create two UI variants

Aware UI: Show temperature/top p labels (read-only) on the welcome note of the Chatbot;

Unaware UI: No mentioning of any temperature/top p settings.

Participants within each rooms will be randomly assigned to aware/unaware UI;

1. On-site details

Venue and equipment

Venue: 0EE702, Old Campus

Seating: One workstation per participant (desktop)

Equipment per person: computer with Chrome, stable Internet, visible timer, note and pen for debate session.

Online chat-room configuration and access:

Preload each workstation homepage with the chat interface URL

Place desk cards with participant codes (e.g., P01, P02) and assigned chat room (A or B)

In-room Verbal Instructions:

“you will work in two rooms, A and B, each with the same poetry partner. Follow the timer and prompts. Use your participant code as your display name. Please keep your own words central; the poetry partner will guide with small examples.”

1. Pre-experimental Hypothesis and Its Integration into Process

**Part 1. Three interaction type hypotheses**

Type A: Diagnosis → Repair

* Definition: Learner or bot identifies a form issue (rhyme, syllable count, line length, meter, lineation) and executes a targeted fix with minimal rewriting.
* Hypothesis A1: Under low-variability decoding (Room A), Type A moves will occur more frequently and earlier in sequences. They will yield higher **cohesion/clarity** ratings and larger local revision impact with shorter time-to-fix.
* Hypothesis A2 (awareness effect): In the Aware group, explicit knowledge of “low variability = control” will increase deliberate Type A prompting (e.g., “check meter,” “tighten rhyme”) and metalinguistic talk. In the Unaware group, similar moves will occur but be attributed to “my careful prompting,” not to parameter differences.
* Planned evidence: Higher frequency of Type A tags in Room A; shorter time from “diagnosis” to “acceptable draft”; increased metalinguistic density (mentions of rhyme/meter/line breaks).

Type B: Exemplar Pivot (Anchor → Apply Characteristics)

* Definition: Learner requests or is offered an exemplar or a list of style characteristics (not author mimicry), extracts 1–2 features (e.g., concrete nouns, example rhyme), and applies them to the draft.
* Hypothesis B1: Under high-variability decoding (Room B), Type B will be used more to rein in drift and increase cohesion; it will raise “good poetry” and coherence ratings after a **surprising draft**.
* Hypothesis B2 (awareness effect): Aware learners will invoke Type B earlier in Room B to stabilize outputs; Unaware learners will pivot later, sometimes after experiencing frustration, and describe the pivot as “finding a model” rather than a strategic response to variability.
* Planned evidence: More early Type B tags in Room B for the Aware group; measurable cohesion gains (lexical overlap, reduced repetition) and improved debate arguments referencing device criteria drawn from exemplars.

Type C: Surprise Harvest (Divergent → Selective Uptake)

* Definition: Learner intentionally seeks surprising metaphors/images, then harvests and splices 1–2 lines into the human-led draft while preserving form constraints elsewhere.
* Hypothesis C1: High-variability decoding (Room B) will produce more Surprise Harvest moves, increasing originality ratings but requiring subsequent Type A repairs to maintain form.
* Hypothesis C2 (awareness effect): Aware learners will plan Surprise Harvest as a strategy, resulting in higher originality without lowering cohesion ratings as much as in Unaware learners, who will more often discard surprising lines.
* Planned evidence: More C→A sequences in Room B for the Aware group; higher originality scores with stable cohesion post-repair; debate references to “one striking line” plus explicit rationale.

**Part 2. How to integrate awareness of interaction types (without confounding parameter awareness)**

1. Unified, non-technical orientation (all participants)

3-minute micro-lesson in the orientation on the three interaction types using plain labels and icons:

Type A: Fix the form/rhyme

Type B: Anchor with an example

Type C: Find a surprise line

Provide one-sentence definitions and a single example each. Do not mention parameters here. This keeps move-type awareness equal in both Aware and Unaware parameter groups.

Self-tagging by participants

Add three on-screen “Move tags” (A, B, C) as small buttons under the message box. Instruct students to click the tag that best matches their next turn’s intention. Default to “None” for free turns.

The chatbot echoes the selected tag in a small banner: “You chose Type C: Surprise Harvest. I will offer 3 bold images; pick one to integrate.”

Benefit: You get per-turn, participant-provided labels to reduce manual coding.

1. Chatbot prompts that surface the three interaction types

At the end of each assistant message, include a gentle, neutral move-suggestion tray:

“Next step options: [A Fix the form] [B Use an exemplar’s features] [C Try a surprising image] [None].”

The tray is identical for both parameter-awareness groups, so it doesn’t contaminate that manipulation.

Micro-reflection stamps

After each revision, a one-click stamp: “What did you just do?” [A Fix] [B Anchor] [C Harvest] [Unsure]. This anchors self-explanations and adds quick metadata.

Part 3. Pre-design the chatbot to elicit and auto-tag the moves (analysis time saver)

Conversation templates per move

* Type A template (auto-selected when user picks A or bot detects form talk):

Bot: “Pinpoint one form issue (syllable count, rhyme, line length). Which line?” [line picker]

Bot offers two minimal fixes and a “why” in one sentence.

Logging: tag: A; target: [meter|rhyme|lineation]; lines affected: [n].

* Type B template:

Bot retrieves one public-domain excerpt or a style characteristics list.

Bot guides feature extraction: “Pick 1–2 features to apply.” Then asks learner to rewrite one line with those features.

Logging: tag: B; exemplar\_id; features\_applied: [concrete nouns|slant rhyme|image schema].

* Type C template:

Bot generates 3 unusual metaphors/images with a one-line rationale each.

Learner picks one; bot prompts selective splice and then offers an optional A repair.

Logging: tag: C; choice\_index; follow-up A suggested: [yes/no].

Auto-tagging rules (fallback when user forgets to tag)

Rule for A: If user/bot text includes tokens like syllable, meter, rhyme, line break, caesura, count, ABAB, AABB → tag A.

Rule for B: If includes exemplar, example poem, characteristics, feature, model, anchor, “like [author]” (convert to characteristics) → tag B.

Rule for C: If includes surprise, unusual, metaphor, image, defamiliarize, unexpected → tag C.

Confidence score stored; human coder only checks low-confidence cases.

Event schema in logs (pre-structured for fast analysis)

* Fields per turn: participant\_id, group (aware/unaware), room (A/B), move\_tag\_user, move\_tag\_auto, move\_confidence, device\_focus, lines\_changed, edit\_distance, time\_to\_next\_turn, suggestion\_type, uptake (within 2 turns yes/no).
* Sequence IDs: pre-assign a template ID when entering a move flow, so you can query sequences like C→A quickly.

Forced one-move-per-turn gating

* The bot enforces “one focus per turn” by confirming the selected move and not mixing multiple device targets. This creates clean, unitary events for coding.

Example bot utterances that instantiate each type

Type A (Constraint Repair)

* “You selected Fix the form. Which line breaks the rhyme? Choose line 2 or 4. Option 1 keeps meaning and rhymes with ‘rain’: ‘trains.’ Option 2: ‘vein.’ Pick one or suggest another.”

Type B (Exemplar Pivot)

* “Let’s anchor with a public-domain quatrain that uses concrete nouns and ABAB. Feature choices: [concrete nouns] [slant rhyme]. Pick one to apply to your line 3.”

Type C (Surprise Harvest)

* “Three unusual images for ‘city rain’: 1) ‘neon dissolving in gutters,’ 2) ‘bus windows chewing fog,’ 3) ‘sirens braided with puddles.’ Choose one to splice into line 2, then we can tighten the rhyme.”

1. Schedule

0-10 min Consent, Quick Orientation, Device-check, Baseline micro-survey (self-efficacy; feelings about poetry; 5-point scale)

10-25 min **Draft with Chatbots**

Warm-up (2 min) Starting with a common theme and sense

Draft with Chatbots (10 min) Producing 4–6 lines according to the given theme and sense. The chatbot uses the pre-set room parameters.

Quick reflection (3 min) Perceived creativity, clarity, and emotional response (valence/arousal).

25-35 min Guided reflection

Aware participants: describe perceived effects of parameter differences and how that influenced your choices of interaction types (ABC) while using it.

Unaware participants: describe differences between Room A and B. (By asking, “Did you notice any differences in terms of interaction types, poetic results between the two rooms?” open-ended; “What do you think caused them?”)

Both: preferences and perceived learning; metalinguistic awareness items (device names used, parameter literacy perception).

35-55 min Debate-format Evaluation/Feedback

Debate-format panel: “Are these outputs good poetry?”  
Structure and roles

Motion: “This house believes the LLM-assisted results constitute good poetry.”

Teams:

Affirmative ( Representatives (3-4) from both rooms): argue that at least one of the produced outputs qualifies as good poetry.

Negative (same as affirmative): argue it does not (or not yet) meet poetic standards.

Moderator (instructor or assigned student): keeps time, ensures balanced turns, poses clarifying questions.

Evidence

Teams must reference specific lines from Room A and Room B outputs, plus 1 device criterion (imagery, metaphor, rhyme, lineation) and 1 coherence criterion.

Timing (25 minutes)

3 min: Moderator frames the motion and rules; defines “good poetry” with a brief rubric reminder.

4 min: Affirmative opening (one speaker).

4 min: Negative opening (one speaker).

4 min: Cross-examination (moderator alternates short questions; each side responds).

4 min: Rebuttals (one per side).

3 min: Audience questions (other participants).

3 min: Vote and closing: silent individual vote on a 5-point “good poetry” scale for each selected poem; moderator collects and announces distribution.

Debate rules

Use concrete citations: quote lines or devices.

Focus on characteristics and effects.

Respectful discourse; concise turns.

Encourage attention to how variability (constrained vs exploratory) affected judgments.

Data from the debate

Coded arguments: device-based, coherence-based, originality/cliché, affective response.

References to parameter effects: explicit (aware group); implicit perceptions (unaware group).

Reference to interaction types

1. **Ethics and risk mitigation**

Informed consent; right to withdraw; clear statement of AI usage and data handling

Privacy: Psuedonymization; secure storage; limited access; internal invitation link to online platform; deletion upon request per policy.

Academic integrity: Emphasize creative agency; prohibit blind submission of AI-only outputs; require process artifacts

Copyright and style guidance: Avoid prompts targeting living authors; focus on stylistic characteristics or public-domain exemplars.

Bias and harmful content: Safety filters; teacher’s oversight; protocol for addressing problematic outputs.

1. **Validity, reliability, and fidelity checks**

Parameter fidelity: log and audit actual settings used; exclude off-protocol cases in primary analyses; include as exploratory

Model stability: fix model version for study duration

1. **Timeline and Milestones**

**Month 1 (6 Sep - 30 Sep)**

IRB approval; interface build; pilot testing of chatbot

**Month 2 (1 Oct- 30 Oct)**

Experiment run; rolling transcription

**Month 3 (1 Nov- 27 Nov)**

Complete analyses; construct usefulness Index; draft tutorial and materials; write report/paper; prepare datasets and prompt design for sharing.

1. **Deliverables:**

Empirical report detailing findings and parameter-conditioned difference.

Practitioner-oriented ranking of interactional targets with classroom guidelines.

Open tutorial: Lesson plan, facilitation guideline, prompt design and parameter cheat sheet.

Reusable interface and logging scripts.

1. **Bibliography**

Clark, Elizabeth, et al. “Creative writing with a machine in the loop: Case studies on slogans and stories,” Proceedings of the 23rd International Conference on Intelligent User Interfaces, 2018.

Crosthwaite, Peter, and Vit Baisa. “Generative AI and the End of Corpus-assisted Data-driven Learning? Not So Fast!” Applied Corpus Linguistics, Vol.3, No.3, December 2023.

Holtzman, Ari, et al. “The curious case of neural text degeneration.” *arXiv preprint arXiv:1904.09751* (2019).

Li, Shaofeng. “Generative AI and Second Language Writing.” *Digital Studies in Language and Literature,* June 11, 2025.

McGregor, Stephen, Matthew Purver, and Geraint Wiggins. “Process based evaluation of computer generated poetry,” Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation, 2016.

Michel, Marije et al. “Collaborative Writing based on Generative AI Models: Revision and Deliberation Process in German as a Foreign Language,” Journal of Second Language Writing, Vol.67, March 2025.

Qian, Wan et al. “ ‘ It Felt Like Having a Second Mind’: Investigating Human-AI Co-creativity in Prewriting with Large Language Models,” arXiv:2307.10811, 2023<https://arxiv.org/abs/2307.10811>.

## Wang, Chaoran, and Wang Zhaozhe. “Investigating L2 writers’ critical AI literacy in AI-assisted writing: An APSE model,” [Journal of Second Language Writing](https://www.sciencedirect.com/journal/journal-of-second-language-writing" \o "Go to Journal of Second Language Writing on ScienceDirect), [Vol. 67](https://www.sciencedirect.com/journal/journal-of-second-language-writing/vol/67/suppl/C" \o "Go to table of contents for this volume/issue), March 2025.