Restored Lines: A Model Finds the Poet's Hardest Line Breaks and Declines to Write Them

Hyunjun Yi
Seoul International School
Seoul, Republic of Korea
hjyi2027@gmail.com

Abstract

A line break is the smallest unit of poetic decision. Nothing in the sentence *so much depends upon a red wheel barrow glazed with rain water beside the white chickens* requires anyone to break it after *wheel*. Williams did that. We strip the line breaks from 2,059 public-domain poems, grade every break by how much dependency structure it severs, and ask Claude Opus 5 to put the breaks back. We expected the model to recover breaks that follow the syntax and to miss the ones that fight it. On free verse, with rhyme, metre and line-initial capitalisation removed, and on poems the model cannot name, it recovers 70% of the breaks that strand a determiner or a conjunction from its host, where a grammar-and-line-length solver recovers 0%. Asked to write free verse of its own, matched to each human poem on subject and length, the same model produces those breaks at 6.1% against the human 12.3%. It locates the poet's override of the sentence and does not perform one. We read that gap as a dissociation between recognising creative agency and exercising it, and argue it makes the model a better instrument for reading poems than a collaborator in writing them.

1 The line break as a unit of decision

Everything else in a poem answers to something. Diction answers to sense, metre to convention, syntax to grammar. The line break answers to nobody. Free verse has no rule fixing where a line ends, which is why Hartman (1980) has to define its prosody by counting what the line does rather than by any constraint it satisfies.

Enjambment is the case where the break contradicts the sentence. The line ends but the clause does not, and the reader is held for the length of a line-turn on a fragment that cannot stand: *a red wheel / barrow* (Williams, 1923). The device is old and the term is old, and automatic detection of it exists for Spanish sonnets (Ruiz Fabo et al., 2017). We are not aware of prior work that uses enjambment as a probe of a model. If enjambment is the poet resisting the sentence, and a language model is a machine for continuing sentences, then a model asked to reconstruct lineation should recover the compliant breaks and fail on the resistant ones. That is a falsifiable prediction about where a model's competence stops.

It is false. This paper reports the failure of that prediction, a second experiment the failure made necessary, and the gap between them.

Our contributions:

- A five-grade typology of the line break, defined by which dependency arcs a break severs, validated against four literary facts the code has no access to (Section 2).
- Two cue controls that had to be built before the probe meant anything. Rhyme and metre give away lineation in 88% of the corpus. Line-initial capitalisation survives flattening and

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

gives away 68% of the hardest breaks on its own. Both would have been reported as findings (Section 3).

- A restoration probe on 245 free-verse poems, case-stripped and filtered for memorisation, with five baselines scored on identical poems (Section 4).
- A generation experiment in which the same model produces hard enjambment at half the human rate (Section 5).

2 Grading the break by what it costs the syntax

Every break sits at a boundary between two words. Parse the poem in its flattened prose form, where the breaks are invisible, and ask which dependency arcs must be cut to put a break there. That cut is the syntactic price of the break, and the poet paid it deliberately. Grades run from a break that costs nothing to a break inside a word. Parses come from spaCy (Honnibal et al., 2020).

Table 1: The grade scale. Grades 0 and 1 go with the syntax, 2 through 4 against it, ordered by how strongly the material before the break demands continuation. Shares are over 36,086 breaks in 2,059 poems. Mean cut is the number of arcs severed, a quantity independent of the relation tier that defines the grades.

Grade	Name	Share	Mean cut	Definition
G0	end-stopped	23.2%	0.0	no arcs cut, the sentence ended anyway
G1	clause-bounded	50.1%	3.1	arcs cut, punctuation already offered the pause
G2	phrase-bounded	19.2%	3.3	no punctuation, a clause-level arc severed
G3	bound-phrase	7.5%	4.4	determiner or modifier cut from its host
G4	lexical	5 breaks	7.0	the break falls inside a word

One design decision carries most of the weight. Tightness is a property of the juncture, not of any arc that happens to span it. An early version graded Milton's soft breaks before *And* as maximally hard, because a coordinator forty words back threw a long `cc` arc over the gap without binding anything across it. Only arcs with an endpoint on the line's last word or the next line's first word count, which is also how the enjambment literature defines the bond.

Four validations. The classifier has no access to literary history. On *The Red Wheelbarrow* it flags exactly four G3 breaks, after *upon*, *wheel*, *rain* and *white*, which are the four the poem is read for, including the stranded preposition Williams isolates on its own line. It flags zero end-stopped breaks, correct, since the poem is one sentence. Ranking poets by enjambment rate reproduces the canon unprompted: Whitman at 4%, whose long lines are end-stopped parallel clauses, against Milton at 46%, Hopkins at 42%, Williams at 75% (Figure 1). By half-century the trough is 1750–1799 at 10.5%, the closed heroic couplet of the Augustans, and the peak is 1900–1949 at 31.5%. Mean cut size rises monotonically across the scale, and cut size is not what defines the grades, so the agreement is not circular. Grades are stable under a larger parser, with 94.3% exact agreement between `en_core_web_sm` and `en_core_web_md` and 98.9% agreement on enjambed-or-not.

3 Two cues that had to be removed first

The probe assumes a line break is a free choice. Twice that turned out to be false, and both times a pilot caught it before the result was believed.

Rhyme and metre. The first pilot returned F1 1.000. There was no bug. The model had perfectly restored a Sophie Jewett sonnet it said it did not recognise, and it did not need to know what enjambment is to do so. The poem is iambic pentameter rhyming ABAB, so the breaks sit at ten syllables on a rhyme word. Classifying the corpus by syllable regularity and rhyme density found that 88% of it has lineation partly given away by sound: 1,181 poems rhymed and metrical, 488 rhymed, 145 metrical, and only 245 free verse. Everything below runs on those 245. A prosody baseline, which does nothing but count syllables and land on rhyme words, is retained to show what sound alone buys. On free verse it buys nothing, at F1 0.126.

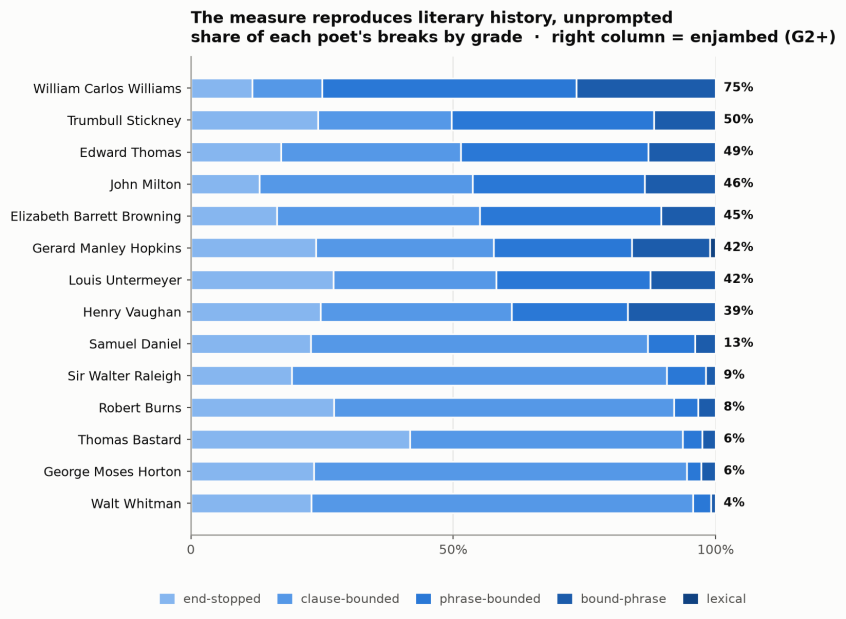

Figure 1: Share of each poet's breaks by grade, over the full 2,059-poem corpus. Nothing in the classifier knows who these people are. The right column is the enjambment rate (G2 and above).

Capitalisation. Much verse before 1900 capitalises the first word of every line, and flattening to prose keeps the casing. In the free-verse set, 67.1% of break boundaries are followed by a capitalised word against 5.3% of non-break boundaries, a 12.6-fold cue with nothing to do with poetics. A baseline that does nothing but break before mid-sentence capitals reaches F1 0.634 and finds 68% of both phrase-bounded and bound-phrase breaks. The headline condition therefore lowercases the prose before the model sees it. That costs the model 8.2 points of bound-phrase recall across all 245 poems and 15.4 points on the memorisation-filtered subset, which is the size of the cue.

Neither leak is a fact about poetics. Both would have been reported as insight.

4 Restoration

Task. The model receives the flattened, lowercased prose and one instruction: put the line breaks back, reproduce every word and mark exactly, output nothing else. The prompt never mentions enjambment, punctuation or line length. It is told the true line count, which separates knowing how many breaks there are from knowing where they go. Predictions are aligned back to the gold word sequence by longest common subsequence, so a dropped word does not shift every downstream break, and poems where the model reproduced under 90% of the original are dropped rather than scored against a text the poet did not write. No poem fell below that threshold.

Memorisation control. A model that has the poem stored can restore it without reading anything. In a separate call we show the first 90 words and ask for title and author, or `UNKNOWN`. Of 245 poems, 195 were named. The headline rests on the 50 that were not. Recognition over-flags, since naming a poem is easier than recalling where its lines end, and for a control that decides which poems the claim rests on, over-flagging is the safe direction. An earlier and sharper control asked for verbatim continuation, but the content filter refused it on the first poem, so it was replaced.

Baselines. Every baseline is handed the gold number of breaks, which makes it stronger than it would otherwise be and removes line-count error as an explanation for any gap. **Syntax + line length** is the strong null: exact dynamic-programming segmentation minimising summed grade cost plus a squared penalty on deviation from the mean line length. This is the best a reader can do knowing only grammar and typical line length. **Punctuation + length** runs the same solver but costs boundaries on visible punctuation alone, so phrase-bounded and bound-phrase boundaries are indistinguishable

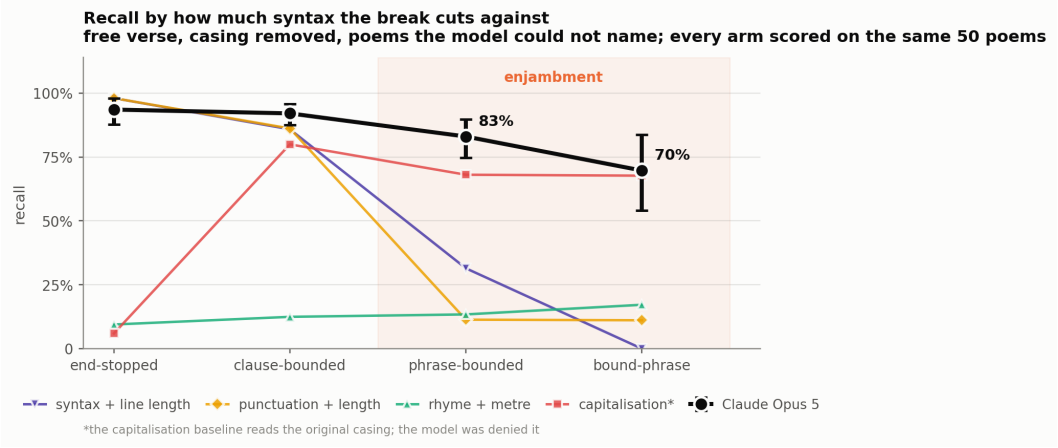

Figure 2: Recall by grade. Intervals are 95% cluster bootstrap over poems (Efron and Tibshirani, 1993), since breaks inside one poem share a poet, a period and a syntax. Every baseline that knows one thing about surfaces falls off a cliff at the enjambment boundary. The grammar-and-length solver reaches zero. The model does not fall.

to it, and any gap between those grades is a fact about the breaks rather than an artefact of the cost function. **Rhyme + metre** and **capitalisation** are the two cue controls above. **Random** places breaks uniformly at the correct rate.

Table 2: Free verse, casing removed from the model's input, restricted to the 50 poems the model could not name. Every arm is scored on those same 50 poems and 784 breaks. The capitalisation baseline reads the original casing, which the model was denied.

	F1	G0	G1	G2	G3
random	0.196	15.9%	20.4%	20.1%	24.2%
rhyme + metre	0.126	9.5%	12.5%	13.4%	17.2%
punctuation + length	0.611	98.0%	86.2%	11.3%	11.1%
syntax + line length	0.645	98.0%	85.8%	31.4%	0.0%
capitalisation	0.634	6.0%	79.9%	68.0%	67.7%
Claude Opus 5	**0.874**	93.5%	92.0%	**83.0%**	**69.7%**

Result. Against the punctuation solver at bound-phrase, the paired difference is $+58.6$ points with a 95% interval of $[+43.1, +72.7]$. Against the grammar-and-length solver it is $+69.7$ points, $[+54.1, +83.6]$, because that solver never breaks there at all. The model is finding breaks that no account of grammar, sound or orthography predicts, on poems it cannot identify, in a form where nothing about the line is fixed in advance. A sentence's lawyer does not do that.

One arm does not fall, and the honest version of this result requires reporting it. The capitalisation baseline reaches 68% at phrase-bounded and 68% at bound-phrase, statistically indistinguishable from the model at both ($+2.0$ points at G3, interval $[-19.7, +23.0]$). The model still beats it on F1 by 24 points, because capitalisation is blind at end-stopped boundaries, where the poet's line-initial capital coincides with a sentence-initial one. The right reading is not that capitalisation rivals the model. Line-initial capitals are a typographic trace of the same authorial decision the task asks about, which makes that baseline a label leak rather than a competing explanation, and it is exactly why we removed the casing from the model's input.

5 Generation

If the model can locate the poet's break, does it make one? We asked the same model for free verse matched to each of the 245 human poems on subject and line count, and ran the identical classifier over its output.

Table 3: Break profiles, 245 poems each, matched on subject and line count.

	breaks	end-stopped	enjambed	hard (G3)	mean cut
human free verse	3,867	23.6%	36.3%	12.3%	2.51
Claude Opus 5	3,850	38.7%	25.6%	6.1%	1.76

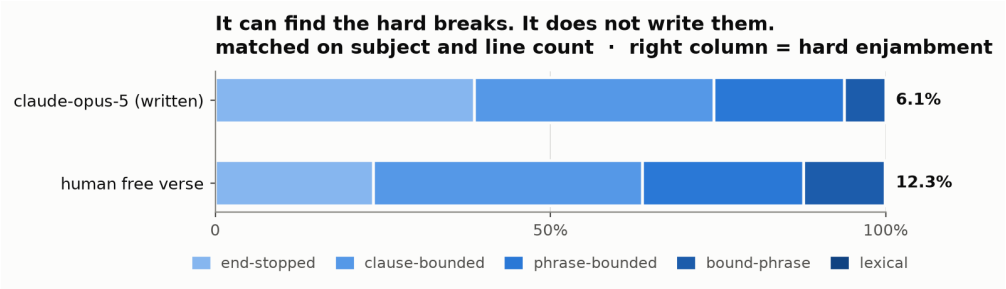

Figure 3: Where the model puts its own breaks. Half the human rate of hard enjambment, and 38.7% of its breaks are end-stopped against the human 23.6%.

The model writes 3,850 breaks against the human 3,867, so it is not compensating with line count. It severs 1.76 arcs per break against 2.51. Recognition sits at 70% of the hardest human breaks. Production sits at half the human rate of making them. Whatever lets it locate the break after *wheel* does not make it write one.

6 Agency

The dissociation is the finding, and it is a claim about agency rather than about capability. A line break in free verse is the one place in a poem where the writer's choice cannot be attributed to a rule. When Williams breaks after *wheel* he is overriding the sentence, and the reader feels the override as a held beat. That is agency in a narrow and checkable sense: a decision with no external warrant, visible in the artefact.

Our model reads those decisions well. At the grade where the fragment before the break cannot stand alone, where a determiner or a conjunction is left hanging, it recovers seven in ten, having been shown nothing but lowercased prose from a poem it cannot name. It is not reconstructing grammar. Grammar, given a line-length budget and asked to place the same number of breaks, recovers none of them.

Then it declines. Given the same subjects and the same line counts and told to write free verse, 38.7% of its breaks are end-stopped and 6.1% are hard enjambments. The capacity to identify the override does not produce the disposition to make one. We do not know the mechanism. One plausible account is that recognition is a discriminative judgment over a text that already exists, where the poem's own difficulty is evidence, while production is a sampling process with no such evidence and a strong prior toward completing the phrase. On that account the model has the reader's competence without the writer's willingness to spend it.

For creative practice the consequence is asymmetric. As a writing collaborator, a model with this profile will smooth a poem toward the sentence whenever it is asked to draft or continue, and it will do so at the exact point where free verse does its work. As a reading instrument, the same model beats every surface heuristic we tested: it can annotate an unfamiliar poem's hardest junctures with no rhyme, metre, punctuation or typography to lean on. The agency the model handles is the poet's, which it can find. Its own, in this domain, it does not exercise. A tool that locates where a human made an unmotivated choice, and declines to make one itself, is worth having, so long as nobody mistakes the first ability for the second.

7 Limitations

The headline rests on 50 poems, and the interval on bound-phrase recall runs from 54% to 84%. The recognition control over-flags, since naming is easier than recalling lineation, and under-flags, since a model can have trained on a text it cannot name. The 245 free-verse poems skew toward translation, Sappho and Trakl among them, and toward early modernism, because free verse in public-domain English before 1929 is largely those two things. That is a sampling constraint of the source corpus, not a filtering artefact. We ran one model at one effort setting, `claude-opus-5` at medium effort with adaptive thinking, so nothing here says whether the effect grows or shrinks with scale. Parses come from `en_core_web_md`, trained on prose, and Milton is both the canonical enjamber and the hardest parse in the corpus, a pairing that deserves a hand-checked subset. Mid-word breaking is nearly absent from public-domain verse, 5 instances corpus-wide, so G4 stays out of the recall plots. Total API spend for every run reported here was about 14 US dollars, and each response is cached under a hash of its request, so re-analysis never re-generates.

Data, code and prior work

The poem corpus and the reversible poem-to-prose flattening come from Bhyravajjula et al. (2025), whose `WISPunspacer` is vendored unmodified here and round-trips every poem used. They released 2,857 public-domain Poetry Foundation poems with whitespace preserved and compared whitespace distributions in generated against published poems. The reconstruction probe, the grade typology, the prosody split and the cue controls are new here. Filtering to English verse of 4 to 40 lines with at least 3 breaks leaves the 2,059 poems used above. Code, cached responses and all run artefacts: https://github.com/hjyi2027/Restored-Lines.

References

Sriharsh Bhyravajjula, Melanie Walsh, Anna Preus, and Maria Antoniak. so much depends / upon / a whitespace: Why whitespace matters for poets and LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China, 2025. Association for Computational Linguistics.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

Charles O. Hartman. *Free Verse: An Essay on Prosody*. Princeton University Press, Princeton, 1980.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python, 2020. https://spacy.io.

Pablo Ruiz Fabo, Clara Martínez Cantón, Thierry Poibeau, and Elena González-Blanco. Enjambment detection in a large diachronic corpus of Spanish sonnets. In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 27–32, Vancouver, 2017.

William Carlos Williams. *Spring and All*. Contact Publishing, Dijon, 1923.