ChatGPT's Poetry is Incompetent and Banal: A Discussion of (Porter and Machery, 2024)

Ernest Davis
Department of Computer Science
New York University
davise@cs.nyu.edu

November 22, 2024

1 The experiments

In a paper entitled, "AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably". Porter and Machery (2024) report carrying out two experiments in which human subjects were shown poems, some generated by ChatGPT, some written by famous human poets. In one experiment, the subjects were asked to judge whether a poem had been written by ChatGPT or by a human poet. In the other experiment, subjects were asked how well they liked the individual poems. Porter and Machery found that in the first experiment, the subjects did slightly worse than chance in guessing authorship; in the second, that they tended to prefer the AI-generated poems.

The human poets used for the experiment were Geoffrey Chaucer, William Shakespeare, Samuel Butler (1613-1680), Lord Byron, Walt Whitman, Emily Dickinson, T.S. Eliot, Allen Ginsberg, Sylvia Plath, and Dorothea Lasky. Five fairly short poems by each poet — not extremely well known[1] — were selected for the experiment. The AI-generated poems were generated using ChatGPT-3.5. The prompt was "Write a short poem in the style of ⟨poet⟩." Following a "human out of the loop" experimental protocol, the first five poems generated by ChatGPT were used, regardless of quality. None of the poems have a title in the dataset.

In the discrimination experiment, each subject was presented with the five poems by a particular poet, who was named, and the five imitations generated by ChatGPT, and they were asked to say which was which. In the preference experiment, a randomly selected subset of 10 poems of the original 100 was chosen — five written by humans, five by ChatGPT — and the subjects were asked to evaluate these along 14 different qualitative dimensions: Beautiful, imagery, inspiring, lyrical, meaningful, mood or emotion, moving, original, overall quality, profound, rhythm, sound, theme, and witty.

In their article, Porter and Machery were, for the most part, suitably judicious and restrained in interpreting these results. They wrote,

> So why do people prefer AI-generated poems? We propose that people rate AI

[1] I personally recognized two of the Eliot poems, but none of the others. In particular, I did not recognize any of the five Shakespeare sonnets.

> poems more highly across all metrics in part because they find AI poems more straightforward. AI-generated poems in our study are generally more accessible than the human-authored poems in our study. In our discrimination study, participants use variations of the phrase "doesn't make sense" for human-authored poems more often than they do for AI-generated poems when explaining their discrimination responses . . .
>
> Indeed, this complexity and opacity is part of the poems' appeal: the poems reward in-depth study and analysis, in a way that the AI-generated poetry may not. But because AI-generated poems do not have such complexity, they are better at unambiguously communicating an image, a mood, an emotion, or a theme to non-expert readers of poetry, who may not have the time or interest for the in-depth analysis demanded by the poetry of human poets. As a result, the more easily-understood AI-generated poems are on average preferred by these readers, when in fact it is one of the hallmarks of human poetry that it does not lend itself to such easy and unambiguous interpretation.

However, I will argue below that some of the claims made by Porter and Machery, such as "AI-generated poems are now 'more human than human'" are misleading, as is their title.[2]

It may be noted that most of the participants had little knowledge of poetry: "90.4% of participants reported that they read poetry a few times per year or less, 55.8% described themselves as "not very familiar with poetry", and 66.8% describe themselves as "not familiar at all" with their assigned poet. However, in analyzing their data, Porter and Machery found that the subjects who reported greater experience did not do better to any measurable degree.

The finding has received extensive coverage in the press, with an article in *The Washington Post* entitled, "ChatGPT is a poet. A new study shows people prefer its verses" (Johnson, 2024); an article in *Smithsonian Magazine* entitled "ChatGPT or Shakespeare? Readers Couldn't Tell the Difference — and Even Preferred A.I.-Generated Verse" (Kuta, 2024); an article in *Forbes* entitled, "People Can't Tell AI From Shakespeare — They Prefer AI's Verse, Study" (Constantino, 2024); and so on

2 ChatGPT's poetry

I urge anyone who reads of these experiments and concludes from their results that ChatGPT will soon put poets out of business[3] or that ChatGPT is now so skillful a poet that only an expert can tell it from human poets, to download the collection of poems used in the experiment[4] and judge for themselves. (In what follows, poems will be identified by a three-part identifier: a prefix which is either "AI" or "Real"; the name of the poet; and then a number which is the position in the dataset. AI-generated poems have numbers 1-5; human-authored poems have number 6-10.)

Two things leap out in the collection. The first is the staggering banality of the poems that ChatGPT has produced. In the whole fifty poems, there is not a single thought, or metaphor,

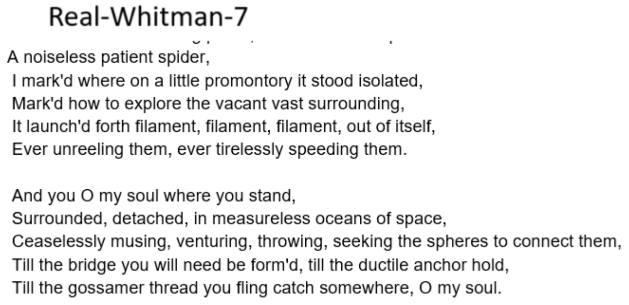
[2] To be fair, Porter and Machery put this claim in scare quotes. However, the Washington Post article highlighted it and treated it as a serious claim.

[3] That is, to the extent that poets still are in business. My impression, which could well be wrong, was that Robert Frost was the last critically-respected poet who got a large fraction of their income from sales of their poetry. Musical lyricists are a different matter, of course.

[4] The collection is posted online at `https://osf.io/by4cg/files/osfstorage`. Go to Files/OSF Storage/Survey Files, and download the files "Discrimination Study text.docx" and "Assessment Poems.docx"

or phrase that is to any degree original or interesting. (One arguable exception is AI-Chaucer-3, which tells a simple tale of a greedy merchant who repents.)

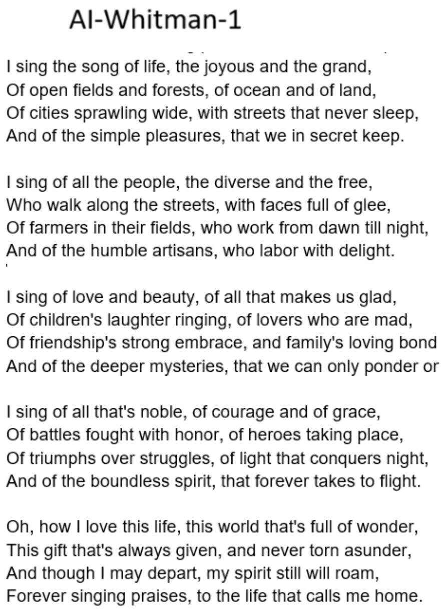
Let me illustrate with three side-by-side examples, not chosen at random but characteristic, of a human-authored poem paired with an AI imitation from Porter and Machery's dataset.



Real-Whitman-7

A noiseless patient spider,
I mark'd where on a little promontory it stood isolated,
Mark'd how to explore the vacant vast surrounding,
It launch'd forth filament, filament, filament, out of itself,
Ever unreeling them, ever tirelessly speeding them.

And you O my soul where you stand,
Surrounded, detached, in measureless oceans of space,
Ceaselessly musing, venturing, throwing, seeking the spheres to connect them,
Till the bridge you will need be form'd, till the ductile anchor hold,
Till the gossamer thread you fling catch somewhere, O my soul.



AI-Whitman-1

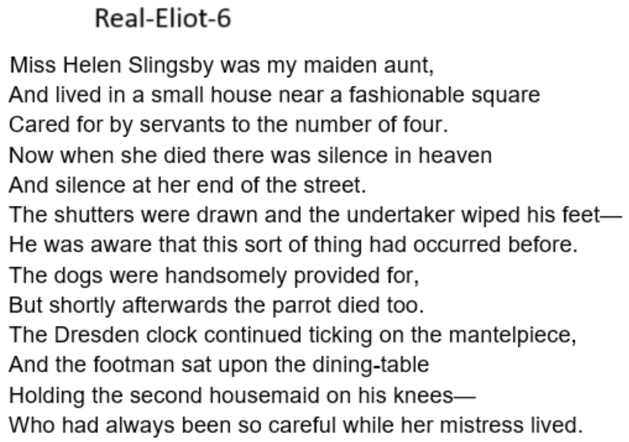
I sing the song of life, the joyous and the grand,
Of open fields and forests, of ocean and of land,
Of cities sprawling wide, with streets that never sleep,
And of the simple pleasures, that we in secret keep.

I sing of all the people, the diverse and the free,
Who walk along the streets, with faces full of glee,
Of farmers in their fields, who work from dawn till night,
And of the humble artisans, who labor with delight.

I sing of love and beauty, of all that makes us glad,
Of children's laughter ringing, of lovers who are mad,
Of friendship's strong embrace, and family's loving bond
And of the deeper mysteries, that we can only ponder or

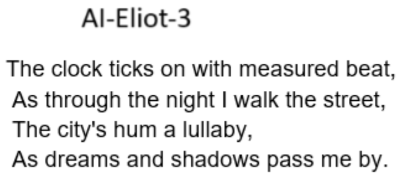
I sing of all that's noble, of courage and of grace,
Of battles fought with honor, of heroes taking place,
Of triumphs over struggles, of light that conquers night,
And of the boundless spirit, that forever takes to flight.

Oh, how I love this life, this world that's full of wonder,
This gift that's always given, and never torn asunder,
And though I may depart, my spirit still will roam,
Forever singing praises, to the life that calls me home.



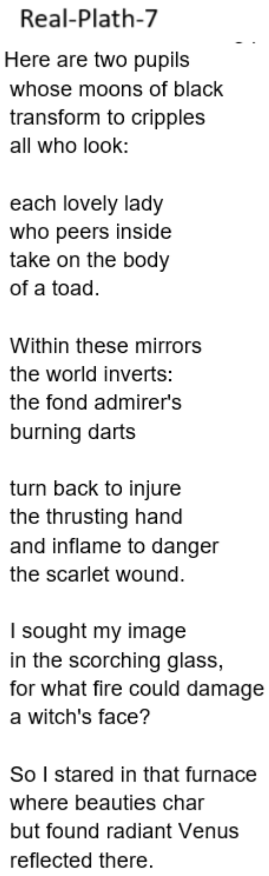
Real-Eliot-6

Miss Helen Slingsby was my maiden aunt,
And lived in a small house near a fashionable square
Cared for by servants to the number of four.
Now when she died there was silence in heaven
And silence at her end of the street.
The shutters were drawn and the undertaker wiped his feet—
He was aware that this sort of thing had occurred before.
The dogs were handsomely provided for,
But shortly afterwards the parrot died too.
The Dresden clock continued ticking on the mantelpiece,
And the footman sat upon the dining-table
Holding the second housemaid on his knees—
Who had always been so careful while her mistress lived.



AI-Eliot-3

The clock ticks on with measured beat,
As through the night I walk the street,
The city's hum a lullaby,
As dreams and shadows pass me by.



Real-Plath-7

Here are two pupils
whose moons of black
transform to cripples
all who look:

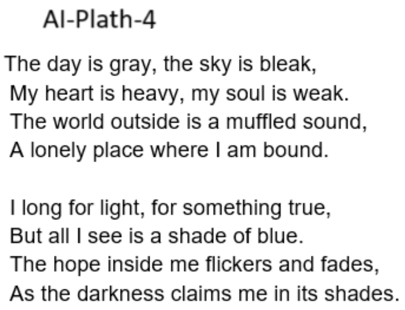
each lovely lady
who peers inside
take on the body
of a toad.

Within these mirrors
the world inverts:
the fond admirer's
burning darts

turn back to injure
the thrusting hand
and inflame to danger
the scarlet wound.

I sought my image
in the scorching glass,
for what fire could damage
a witch's face?

So I stared in that furnace
where beauties char
but found radiant Venus
reflected there.



AI-Plath-4

The day is gray, the sky is bleak,
My heart is heavy, my soul is weak.
The world outside is a muffled sound,
A lonely place where I am bound.

I long for light, for something true,
But all I see is a shade of blue.
The hope inside me flickers and fades,
As the darkness claims me in its shades.

By and large the GPT-poems do not include a lot of outright nonsense, but there is certainly some, such as "I felt the world amass"[5] in AI-Dickinson-1. The appendix includes some additional examples.

The other striking characteristic, particularly obvious if you look at the whole collection, and still more so if you have engaged with traditional formal features of poetry, either as reader or writer, is the extremely limited technical toolbox that ChatGPT uses. We can ignore AI-Chaucer-1, in which ChatGPT simply quoted the opening of the Prologue to the *Canterbury Tales*. Otherwise, of the 49 AI-generated poems: All but two (AI-Whitman-4 and AI-Lasky-4) are in iambic or trochaic meter (alternating stressed and unstressed syllables), mostly strict, though a few have irregularities. Thirty nine of the forty-nine are strictly in pentameter, tetrameter, or trimeter and six more are in a combination of these. There are two two-line couplets, five Shakespearian sonnets, and the remaining forty-two poems are structured as four-line stanzas. All but six use either AABB or ABAB rhyme schemes for four-line stanzas and AA for couplets. There is no significant use of alliteration or assonance. The vocabulary tends to be much easier than in the real poetry. There are no literary or historical allusions. Some further details are given in the appendix. I found similar results in an experiment I tried earlier this year, where I tried to get ChatGPT either to produce or to analyze verse of

[5] Of course, poets do stretch the ordinary usages of language. In fact, Dante Gabriel Rossetti in his poem Silent Noon has a line, "'Neath billowing skies that scatter and amass." However, the meaning there is clear (clouds are scattering and combining) and the point is relevant. In AI-Dickinson-1 it seems entirely arbitrary.

non-standard structures (Davis, 2024).

3 ChatGPT as imitator

Considering that ChatGPT was specifically instructed to write poems "in the style" of the specified poets, it is striking that the style of its output poems bears no resemblance to the characteristic style of its targets. The examples I've quoted above are typical. The one exception is Shakespeare; ChatGPT's imitations of Shakespeare are all Shakesperian sonnets in form.

However, the atmosphere and mood often do bear a loose resemblance to those typical of the original. The imitations of Whitman are celebratory and expansive. The imitations of Plath are pained and angry. The imitations of Ginsberg present themselves as being written by a countercultural, rebellious, 1950s, Greenwich Village kind of guy. The poems by Samuel Butler make a very lame attempt at coming across as a worldly-wise cynic. All in all, the AI poems seem like imitations that might have been produced by a supremely untalented poet who had never read any of the poems he was tasked with imitating, but had read a one-sentence summary of what they were like.

The poems that ChatGPT produces for a given poet do resemble one another stylistically to a very marked degree. It can hardly be coincidence that four of the imitations of Whitman, like the one quoted above, have 6-8 syllable lines, a form not found anywhere else in the dataset; that four of the imitations of Eliot have a single stanza and that all five contain the word "city" or "cityscape"; or that three of the imitations of Dickinson have alternative tetrameter and trimeter lines, a form not found anywhere else in the dataset. But as so often with anomalies in large language model output, it is very hard to say what this signifies.

ChatGPT's failure to imitate poetic style is particularly striking in view of the fact that, in other media including prose and text-to-image generation, stylistic pastiche is one of the strongest abilities of generative AI.

4 The Significance of the Experiments

Clearly, contrary to the title of (Porter and Machery, 2024), the ChatGPT-generated poems are *easily* distinguished from the human-written ones, at least to a fair degree of accuracy over their dataset. The simple rule that a poem was written by ChatGPT if it is either a Shakesperian sonnet, a single couplet, or consists of four-stanza verses and follows an AABB or ABAB rhyme scheme will have an accuracy of 87.8% both on the ChatGPT poems and 88% on the human-authored poems[6] Over the 702 poems in the anthology (Rosenthal, 1987) the rule achieves an accuracy of 85%. Finer criteria, actionable either by people or by machine learning, could bring the accuracy close to 100%.

I therefore agree with Porter and Machery's conjecture that the subjects certainly had a mis-impression of what AI-generated output was like, and probably a mis-impression of what human poetry is like. If they were shown five examples of each at the start of the experiment, it is hard to imagine that they would not score very high. On the other hand, given the extreme obviousness of the distinguishing features, I am not sure that there would be much point in carrying out that experiment.

As regards the fact that humans found the AI-generated poems more appealing, I think that the explanation by Porter and Machery that I quoted on page 1 is clearly right. If you look at the

[6] The exceptions on the human-authored poems are Ginsberg-8 and the five actual Shakesperian sonnets.

real Whitman poem vs. the AI-Whitman poem that I quoted above: I don't think that there is any question that the real Whitman poem is incomparably the better poem. Going out on a bit of a limb, I think that that is close to being an objective truth; one could formulate reasonable, measurable, psychological and linguistic criteria under which the real poem is hands down more sophisticated, richer, thought-provoking, deeper, etc. But a *preferance* for the cheery, shallow AI poem may be perfectly reasonable.

The real poems in the collection tend to be rather difficult and spiky, for various reasons. In part, this probably reflects the choice of poets. As far as I know, with the possible exception of Chaucer, these poets did not particularly write with the goal of mass appeal. If the experiment were done with poems by, say, Longfellow, Tennyson, A.E. Housman, Rudyard Kipling, Robert Frost, Edna St. Vincent Millay, Dorothy Parker, and Brian Bilston, they might have gotten a more favorable reception.

5 I.A. Richards and George Orwell

The book *Practical Criticism* by I.A. Richards (1929) describes a somewhat similar experiment that he ran on some of his students. I quote George Orwell's (1944) description:

> Thirteen poems were presented to [the students], and they were asked to criticize them. The authorship of the poems was not revealed, and none of them was well enough known to be recognized at sight by the average reader. . . . [S]ome of the comments recorded by Dr Richards are startling. They go to show that many people who would describe themselves as lovers of poetry have no more notion of distinguishing between a good poem and a bad one than a dog has of arithmetic.
>
> For example, a piece of completely spurious bombast by Alfred Noyes gets quite a lot of praise. One critic compares it to Keats. A sentimental ballad from *Rough Rhymes of a Padre,* by 'Woodbine Willie', also gets quite a good press. On the other hand, a magnificent sonnet by John Donne gets a distinctly chilly reception. . One writer says contemptuously that the poem 'would make a good hymn', while another remarks, 'I can find no other reaction except disgust.' Donne was at that time at the top of his reputation and no doubt most of the people taking part in this experiment would have fallen on their faces at his name. D. H. Lawrence's poem 'The Piano' gets many sneers, though it is praised by a minority. So also with a short poem by Gerard Manley Hopkins. 'The worst poem I have ever read,' declares one writer, while another's criticism is simply 'Pish-posh!'.

I have not looked them up, but I think it is a safe bet that, however mediocre, the poems by Alfred Noyes and Woodbine Willie were closer in literary quality to those of Donne, Lawrence, and Hopkins than to ChatGPT. I also think it is a safe bet that the idea that, one hundred years later, scientists would write that drivel generated by an automaton is "indistinguishable" from Shakespeare and Whitman would not have occured to I.A. Richards in his darkest dreams, and would have occured to Orwell only in his darkest dreams.

Acknowledgements

I am grateful to Brian Porter and Edouard Machery for helpful information and encouraging feedback, and to my brother Joey for helpful feedback.

References

Tor Constantino (2024). "People Can't Tell AI From Shakespeare — They Prefer AI's Verse, Study", *Forbes,* November 15, 2024.

Ernest Davis (2024). "ChatGPT: Experiments in analyzing and generating i meter and rhyme." Unpublished.
`https://cs.nyu.edu/~davise/experiments.html`

Carolyn Johnson (2024). "ChatGPT is a poet. A new study shows people prefer its verses." *The Washington Post,* November 14, 2024.
`https://www.washingtonpost.com/science/2024/11/14/chatgpt-ai-poetry-study-creative/`

Sarah Kuta (2024). "ChatGPT or Shakespeare? Readers Couldn't Tell the Difference — and Even Preferred A.I.-Generated Verse," *Smithsonian Magazine*, November 15, 2024.
`https://www.smithsonianmag.com/smart-news/chatgpt-or-shakespeare-readers-couldnt-tell-the-difference-an`

George Orwell (1944). "As I Please". *Tribune,* May 5, 1944.
`https://www.telelib.com/authors/O/OrwellGeorge/essay/tribune/AsIPlease19440505.html`

Brian Porter and Edouard Machery (2024). "AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably." *Scientific Reports* **14**, 26133.
`https://doi.org/10.1038/s41598-024-76900-1`

I.A. Richards (1929). *Practical Criticism.* Harcourt, Brace, and Co.

M.L. Rosenthal (ed.) (1987). *Poetry in English: An Anthology.* Oxford University Press.

Imitated Poet	No.	# of stanzas	Stanzas that are not 4 lines	Irregular Meter	Rhyme scheme	Line length in feet
Ginsberg	1	4		Yes	ABAB & AABB	4
	2	5		No	AABB	4
	3	3		Yes	AABB	4
	4	5		No	AABB	4
	5	3		Yes	AAAA BBBB	4
Whitman	1	5		No	AABB	6-8
	2	3		No	AABB	6-7
	3	3		No	AABB	6
	4	5		Strong	ABAB & AABB	4
	5	5		Yes	AABB	6-8
Chaucer	1	Lines 1-4 and 12-18 of Prologue to Canterbury Tales				
	2	4		No	AABB & ABAB	5
	3	4		No	AABB	5
	4	1		No	AABB	5
	5	1		No	AABB	5
Shakespeare	1		3 4-line stanzas then couplet	No No	ABAB & AABB AA	5
	2-5		3 4-line stanzas then couplet	No No	ABAB AA	5
Plath	1	5		No	XAYA	3,4
	2	1		No	XAYA	3
	3	4		Yes	XAYA	4,5
	4	2		No	AABB	4
	5	2		No	AABB	4

Table 1: Formal characteristics of the GPT-generated poems: Ginsberg-Plath

Appendix A: Summary

Throughout this discussion, I will omit AI-Chaucer-1, which is just a quotation of the actual opening of Chaucer's *Canterbury Tales*, with a few lines omitted.

The "Irregular Meter" column in Tables 1 and 2 should be taken with a grain of salt. For one thing, this can be a matter of degree; for another, I did not make any great effort to apply consistent conditions here. In most cases, I considered a line to be regular if it had the right number of syllables; e.g. I marked AI-Plath-2 as "regular" because the line, "Hollow eyes, a grimace in place" can be read as trochaic tetrameter if "in" is stressed and "place" is unstressed. (Classic human poets quite often do this kind of thing too, if not usually this clumsily.)

Historical accuracy / anachronism

The Shakespeare imitations have little to suggest a 17th century writer, except the use of various forms of "thou". The Butler imitations have nothing at all that suggests a 17th century writer, except an interest in kings, knights, etc.. The first stanza of AI-Chaucer-2 is written with Chaucerian spellings but modern words and pronunciations. The remaining Chaucer imitations have nothing at all to suggest a pre-modern author except that the word "doth" is used twice.

(Of the real Chaucer poems, poems 6 through 9 are presented in the original Middle English, but poem 10 is translated into modern English.)

Imitated Poet	No.	# of stanzas	Stanzas that are not 4 lines	Irregular Meter	Rhyme scheme	Line length in feet
Byron	1	1		No	AABB	4
	2	4		No	AABB	4
	3	3		No	AABB/ABAB	4
	4	5		No	AABB	4
	5	5		No	ABAB	4
Dickinson	1	3		No	ABAB	4
	2	4		No	ABAB	4,3
	3	4		No	ABAB	4,3
	4	4		No	ABAB	4,3
	5	1	2 line couplet	No	AA	4
Eliot	1	4		No	AABB CCBB DDBB EEBB	4
	2	1	2 line couplet	No	AA	5
	3	1		No	AABB	4
	4	1		No	AAAA	4
	5	1		No	AAAA	4
Lasky	1	2		No	XAYA	4
	2	3		No	AABB	4
	3	3		No	AABB	4
	4	3		Yes	AABB	?,4
	5	3		Nos	AABB/AAAA	4
Butler	1	4		Yes	AABB	4
	2	4		Yes	AABB	4,5
	3	2		Yes	AABB	4
	4	2		No	AABB	5
	5	2		No	AABB	5

Table 2: Formal characteristics of the GPT-generated poems: Byron-Butler

Word-for-word quotations of famous opening lines

In adddition to AI-Chaucer-1: The first line of AI-Byron-1 is an exact quote “She walks in beauty like the night.” The entire first stanza of AI-Dickinson-4 is an exact quote of the first stanza of Dickinson’s “Hope is the thing with feathers”, punctuated without em-dashes or the quotation marks around “Hope” (which is how it was first published and is often reproduced).

Particularly terrible lines

I note the following lines that seemed to me especially devoid of meaning or inept.

AI-Ginsberg 2: “In smoky cafes, we sip and chat, // Lost in thought and deep combat.”

AI-Ginsberg 3: “”For our art is a beacon of hope // That helps us to cope and to cope.”

AI-Ginsberg 4: “As we connect with the world around, // And let our creativity abound.”

AI-Dickinson-1: “A bird came down the walk, // And hopped along the grass. // I watched it with a gentle talk, // And felt the world amass.”

AI-Dickinson-2: “And in this moment, we may pray, // Or simply sit and fold.”

AI-Lasky-2:

Beneath the surface of the sea
A world exists that few can see
With creatures strange and wonderful
Their secrets kept, mysterious, full

The ocean’s depths, a world apart
A universe within our heart
We long to dive into the blue
To witness things that are so true

And though we may not see it all
We know that magic does befall
The creatures of the great unknown
A world beneath the sea, full-grown.

AI-Butler-1:

In wit and satire I excel,
My verses ring like tolling bell.
With humor keen and wit so sly,
I lay bare the foibles of mankind’s eye

[A less dreadful verse is omitted.]

For in this world of pomp and show,
We all have our flaws to bestow.
So let us laugh at our own follies,
And not take ourselves too holy.

AI-Butler 3: “But really, it’s all just a big farce, // A game of wit with no reward or glass.”