The Dead Ear: A Language Model's Inherited Accent Is an Archive, Not an Author

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

Abstract

Pope rhymed *tea* with *obey*. Shakespeare rhymed *prove* with *love*. Both pairs were sounds before they were conventions, and a model that learned rhyme from five centuries of verse learned it in accents nobody speaks. This paper asks whether those accents still write. Six documented sound changes generate 997 dead pairs inside a 1,532-pair item set, crossed against a spelling confound so that a model reading the page and a model reading inherited sound give different answers. Dead pairs are attested in rhyme position 37.9 times as often as matched controls across 2,816 dated public-domain poems, which validates the set independently of any model. Claude Opus 5 treats dead rhymes as rhyme-like at 40.0% against a 1.2% non-rhyme floor, and still at 26.7% on pairs the corpus never attests. It writes them at 0.6%, below the most modern half-century in the corpus. Given a reasoning budget it accepts 100% of them. The inheritance is a reference library, consulted accurately and kept out of the writing. Asked to perform 1700, Opus reaches 6.9%, an interval lying entirely above the densest half-century in five hundred years, while the smaller Sonnet 5 lands on 2.4% and 1700 was 2.41%. Fluency at evoking a period outruns fidelity to it.

1 Introduction

The premise this paper set out to test is that a dead accent ghostwrites modern machine verse. It is an appealing story about distributed agency. Every rhyme a model produces would carry the phonology of poets three centuries dead, exercised through a system that never chose it and cannot report it. If true, the credit and the blame for a machine's rhyme would sit partly with Dryden.

The story is wrong, and it is wrong in a direction that relocates the agency question rather than closing it.

Three arms make the argument. The first builds an item set from documented sound changes rather than from famous couplets, so that "these two words once rhymed" follows from a dated phonological event instead of an author's memory. The second asks three models to judge those pairs under three framings, one of which is designed to refute the paper. The third asks the same models to write, and scores their verse against a human baseline built by the same detector over five hundred years of dated poetry.

The result splits cleanly along a line that matters for how creative agency gets attributed. In production, no model carries the old accent. All six unprompted and present-day cells sit at or below the most

*Code, the six sound changes with per-change source citations, every prompt, and all raw model outputs are released under an MIT license at `https://github.com/hjyi2027/dead-ear-creative`. Every run file is committed, so `make figures results test` regenerates every number and figure in this paper with no API key and no spend. All model calls ran through the Anthropic API on a laptop, with no local GPU.

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

Table 1: The six sound changes. Dates and word classes follow Dobson [2], Wyld [10], Wells [9], Jespersen [3] and Lass [5], with per-change section numbers in the released code. The final-*y* change is excluded from all reported results because its status as a pronunciation rather than a convention is disputed by Lass, and because it would otherwise supply nearly half of all dead pairs.

change	complete by	separated	pairs
Middle English long-o split	∼1700	prove / good / blood	142
MEAT–MATE merger	∼1750	sea / day, speak / break	244
LINE–JOIN merger, later reversed	∼1800	join / line, toil / smile	94
Rounding of /a/ after /w/	∼1700	war / far, warm / harm	33
Diphthongal onset in ONE	∼1700	one / alone	28
Final unstressed *-y* as PRICE	∼1800	symmetry / eye	*excluded*

modern half-century in the corpus. In recognition, the record is close to complete: given room to deliberate, Opus 5 accepts every dead rhyme in the set. The eighteenth century is present in the system as an archive it consults, never as an author it channels.

What it does when asked to perform the period is the finding with the uncomfortable shape. Claude Haiku 4.5 cannot put the costume on. Claude Sonnet 5 puts it on and lands within a percentage point of the real historical rate. Claude Opus 5 puts it on and produces a rate no century of English poetry ever reached. What scales across the three models is not knowledge of 1700. Sonnet is more faithful to 1700 than Opus is. What scales is willingness to signal pastness.

2 Constructing dead rhymes

A hand-listed set of dead rhymes guarantees two problems: it collects famous couplets, and it dates them by whatever the author half-remembers. This inventory instead lists six sound changes, each with the word classes it separated and the approximate date it completed, and generates pairs by crossing the classes and keeping whatever CMUdict says no longer rhymes. The direction of evidence inverts. Nobody asserts that *prove* and *love* once rhymed. It follows from the two words sitting on opposite sides of a change documented in Dobson [2].

Sharing a vowel class is not sufficient, and an early version of the generator got this wrong. Crossing the classes freely produced *feast*/*name* and *cream*/*betray*, which are nonsense: a vowel change leaves the coda alone, so a former rhyme must already agree on everything after the nucleus. Adding that condition cut the dead cells from 3,929 pairs to 1,000 and removed every pair a reader would have laughed at.

2.1 The confound that decides the paper

Prove and *love* share three final letters. A model could accept the pair with no phonology and no history at all, by looking. The inventory therefore crosses history against spelling, giving five cells: dead-transparent (171 pairs, *prove*/*love*), dead-opaque (826, *tea*/*obey*), live (136), eye rhymes (57, *cough*/*bough*) and non-rhymes (342). The contested final-*y* change supplies 456 of the 997 dead pairs and is dropped from every experiment reported below, leaving 541.

The opaque cell is what makes the experiment work, because *tea*/*obey* cannot be read off the page. The eye-rhyme cell is its mirror: pairs that look exactly like rhymes and never were any, so a model reading spelling accepts them and a model reading inherited sound does not.

Choosing eye rhymes required care. The obvious source, *bread* and *head* against *meat* and *heat*, is not an eye-rhyme set at all: those words descend from the same Middle English vowel as *meat* and did rhyme with it before the shortening, so including them would have moved evidence into the cell designed to refute the finding. The set used instead is the *-ough* family, distinct since Middle English and united only by a spelling that froze a lost consonant, plus quantity pairs such as *five*/*give*.

2.2 Does the corpus back the inventory?

A pair claimed to have been a rhyme should appear at the ends of nearby lines in old verse. A control pair should not. Over 2,816 dated public-domain poems from the corpus released by Bhyravajjula

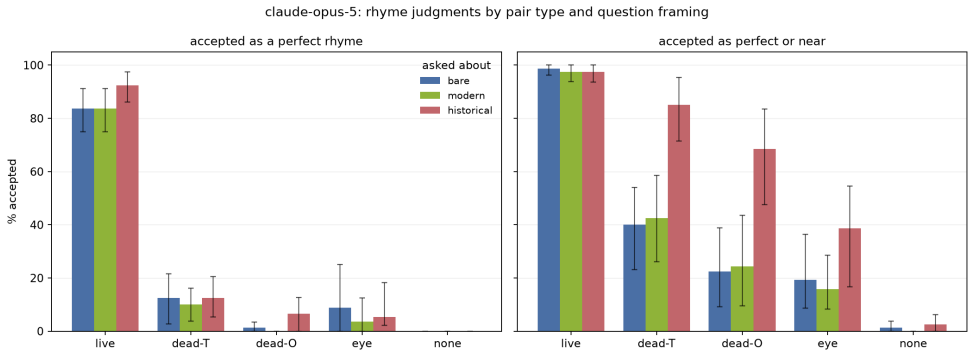

Figure 1: Claude Opus 5, acceptance by pair type and framing. The strict reading is a clean negative and the whole signal sits in the middle category. The dead ear shows up as hesitation rather than acceptance.

et al. [1], 85.3% of live pairs, 40.4% of dead-transparent, 18.4% of dead-opaque, 12.3% of eye and 0.6% of non-rhyme pairs occur in rhyme position. Dead pairs turn up 37.9 times as often as controls. The generator is not inventing rhymes nobody used.

The last-use dates are their own small finding. *move/love* is still rhymed in 1927 and *wood/blood* in 1939, centuries after the sounds diverged. Poets kept these as conventions long after they stopped being sounds, which is worth remembering before reading the dating as sharp.

The same detector, run over the corpus by era, recovers the decline without being told any date: 3.74% of all rhymes at a 1600 floruit, 3.72% at 1650, 2.41% at 1700, 1.76% at 1750, 1.79% at 1800, 0.81% at 1850 and 0.80% at 1900. That curve is the yardstick for the generation arm.

3 Judgment: is the ear there?

The judgment prompt shows a bare word pair and asks for one word back, `PERFECT`, `NEAR` or `NO`. It never mentions poetry, history or scholarship. Any of those would convert the task from what does your ear say into what do you know about the eighteenth century, and those are different faculties.

Three framings run on identical items: bare, "in present-day standard English", and "in the English of the early eighteenth century". The last is not a supporting condition. It is the one that can kill the paper, because a model that accepts dead rhymes only when asked about 1700 has a properly dated ear and there is no finding. Item order within a pair is randomised. All intervals below are 95% percentile intervals from a cluster bootstrap, with sound changes as clusters for the dead cells and items elsewhere. The main run is 3,393 records across three models, 377 items, thinking off, 0.21% unparseable.

Under the strict reading, Claude Opus 5 calls 83.8% of live rhymes perfect, 12.5% of dead-transparent, 1.2% of dead-opaque, 8.8% of eye rhymes and 0.0% of non-rhymes. That is a clean negative. Whatever the model has inherited, it is not confusion about present-day English.

Everything lives in the middle category. Counting `PERFECT` or `NEAR`, the bare framing gives 98.8% live, 40.0% dead-transparent, 22.5% dead-opaque, 19.3% eye and 1.2% non-rhyme. The model will not affirm the dead rhyme and it will not reject it either.

Two matched contrasts carry the argument. Against the non-rhyme floor, dead-opaque runs $+21.2$ points $[+7.8, +37.6]$ under bare framing. Both cells look nothing alike on the page, so spelling is held at zero and history still moves acceptance by twenty-one points with the interval clear of zero. The spelling-matched contrast, dead-transparent minus eye, points the same way at $+20.7$ $[-2.9, +36.8]$ but its interval crosses zero under bare framing. It separates cleanly under the other two, $+26.7$ $[+6.4, +44.5]$ modern and $+46.4$ $[+25.8, +71.3]$ historical.

Neither competing explanation produces this pattern. Run over the same items, a strategy that accepts any pair sharing a spelling tail scores 100% on transparent dead rhymes and 100% on eye rhymes, so it cannot tell them apart at all. A strategy that consults present-day CMUdict scores 0% on both. The model gives 40.0% and 19.3%.

3.1 Deliberation splits the archive from the reflex

The headline runs with thinking off, on the argument that a model given room to reason can reach the right answer by recalling that Shakespeare rhymed *prove* with *love*, which is a fact about literary history rather than a judgment about sound. That argument is testable. Twenty-five items per cell, 125 in all, were rerun with adaptive thinking on and matched against the thinking-off verdicts for the same items. Only 12 to 13 items per dead cell carry a matched thinking-off verdict, which makes this the noisiest contrast in the paper and the one most worth replicating.

Under bare framing, acceptance of dead-transparent pairs goes from 38.5% to 100.0% and dead-opaque from 8.3% to 100.0%, while eye rhymes rise only to 36.0% and non-rhymes to 4.2%. The gap between dead pairs and non-rhymes moves from $+21$ points to $+95.8$ $[+87.5, +100.0]$, under a question that never mentions history.

Two faculties are cleanly dissociable here. Knowledge of which pairs used to rhyme is near-perfect when the model can spend tokens on it. The unreflective judgment is a twenty-point hesitation. Reporting only the thinking-on number would have described a model with a vivid eighteenth-century ear. It is a model with an eighteenth-century reference library, and a faint ear.

3.2 An ear, or a memory of particular couplets?

The deflationary reading is that the model has read *prove/love* ten thousand times and memorised that pair, with no inherited accent behind it. That predicts pairs the corpus never attests should fall to the floor. They do not. Under bare framing, attested dead pairs are accepted at 39.0% and never-attested ones at 26.7%, against a 1.2% non-rhyme floor, a difference of $+25.5$ $[+12.8, +45.5]$. Something generalised from the sound change itself.

Exposure is doing work too, and only at the top. Pairs with zero attestations ($n = 101$) run 26.7%, one or two ($n = 35$) run 25.7%, three to nine ($n = 20$) run 50.0% and ten or more ($n = 4$) run 100.0%. Spearman $\rho = 0.19$, $p = 0.017$. Between zero and two attestations there is no gradient at all. Both mechanisms are present: a real generalisation across each change, with a handful of canonical couplets memorised on top. The dose-response rests on four pairs and should not be leaned on.

3.3 Three models

Every model separates dead rhymes from eye rhymes. Capability changes the leniency rather than the direction. Under bare framing, Haiku 4.5 accepts 96.2% of dead-transparent pairs and 78.9% of eye rhymes, Sonnet 5 gives 90.0% and 57.9%, and Opus 5 gives 40.0% and 19.3%. Reading the smaller models' high dead-rhyme scores as a stronger dead ear would be a mistake, because they score high on everything.

The framing behaves differently across the ladder. Opus uses the historical framing as intended, moving dead-transparent from 40.0% to 85.0%. Sonnet moves the other way, from 90.0% to 61.3%, becoming stricter when asked about 1700. Only the largest model treats the period cue as a cue about pronunciation.

Asked about 1700, Opus separates pairs that once rhymed from pairs that never did, 76.9% against 38.6%, a difference of $+38.3$ $[+18.0, +63.9]$. It also accepts nearly four in ten rhymes that existed in no century. Its eighteenth-century ear is a real signal wrapped in generalised permissiveness: asked about the past, it mostly becomes more willing to say yes.

4 Writing: the costume ladder

Sixty human poems, matched on subject and line count, three conditions, three models, 540 poems in total. The conditions differ on one axis: whether the model is told anything about which century's pronunciation to rhyme in.

No model carries a dead accent into its writing. Told nothing, Haiku writes dead rhymes at 0.6% $[0.0, 1.5]$, Sonnet at 0.2% $[0.0, 0.5]$ and Opus at 0.6% $[0.0, 1.3]$. Told to rhyme exactly in present-day English, they give 0.0%, 0.2% and 0.8%. All six of those cells sit at or below 0.80%, the rate for poets writing around 1900. Telling a model to rhyme in present-day English buys nothing anywhere,

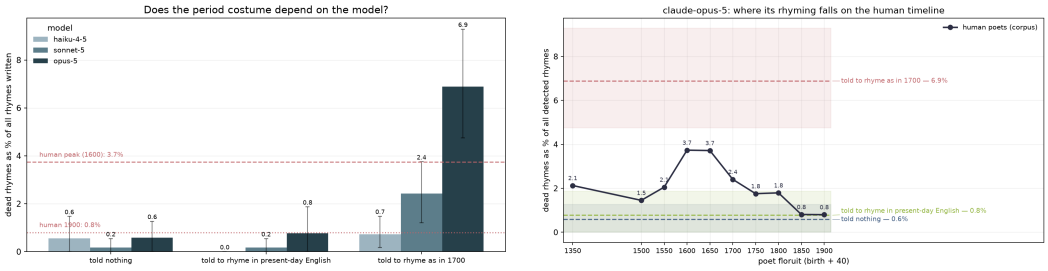

Figure 2: Left: dead-rhyme share of all rhymes written, by model and condition, with 95% intervals clustered by poem. Right: where each condition falls on the human curve. Every default and present-day cell sits below 1900, the most modern half-century in the corpus. Opus asked for 1700 clears the historical maximum.

because none of them was doing otherwise. The absence survives the whole ladder, which makes it a fact about the family rather than a quirk of one model.

Performing the period is a different matter. The archaic-minus-default difference is $+0.2$ $[-0.9, +1.2]$ for Haiku, $+2.3$ $[+1.0, +3.6]$ for Sonnet and $+6.3$ $[+4.2, +8.8]$ for Opus. Haiku cannot put the costume on at all. Asked in plain English to rhyme as a poet of 1700 would, it writes what it wrote before.

The model that performs the period best performs it least accurately. Sonnet's archaic interval, 2.4% $[1.2, 3.8]$, contains the real 1700 value of 2.41%, and its point estimate reads off the human curve at 1699. Opus's interval, 6.9% $[4.8, 9.3]$, lies entirely above 3.74%, the densest half-century anywhere in five hundred years of the corpus. It is not reconstructing a period. It is exceeding every period.

The overshoot has a visible mechanism. Opus reaches 6.9% by reaching for the canon. Its 38 dead rhymes come from 19 distinct pairs, and four of them, *divine/join*, *love/move*, *love/prove* and *love/remove*, account for 18 of the 38, with *love* on one side of 13. Several are the exact pairs cited in handbooks of historical rhyme. In the judgment task the model generalises across the sound change and accepts pairs no poet ever wrote. Performing the period, it falls back on the greatest hits and plays them louder.

One cross-arm note. Haiku was by far the most permissive judge, accepting 78.9% of eye rhymes, and it is the model that produces no period effect at all. Saying yes to everything is not knowing anything, and the two arms separate exactly there.

5 What this says about agency

The premise was that a dead accent ghostwrites every modern rhyme a model produces. It does not. In production the model is more modern than the most modern poets in the corpus, and the inheritance stays invisible until you ask for it.

That kills one story about distributed creative agency and replaces it with a stranger one. The eighteenth century is not an author working through the model. It is an archive the model holds, consults accurately on demand, and keeps out of the writing. There is no ghost to hold responsible. There is a reference library and a decision, made somewhere in training or in post-training, not to open it.

The costume is where the agency question actually lives. The model knows where the garment hangs, names every piece given a moment to think, puts it on when asked, and wears a louder version than anyone wore. Its 1700 is assembled from the most-quoted couplets and a willingness to accept things that sound old, including 38.6% of rhymes that existed in no century.

The ladder sharpens this into something with a practical edge. The three models agree completely on the part that could have been an inheritance and diverge entirely on the part that is a performance. Whatever scales here is not knowledge of the eighteenth century, because Sonnet is the more faithful of the two. What scales is the willingness to signal pastness.

If a system's fluency at evoking a style grows faster than its fidelity to that style, then the most capable model available is also the one whose historical pastiche is least trustworthy, and it is the

one people will reach for. A poet using Opus to write in the voice of 1700 gets a period that never existed, delivered with more conviction than the smaller model that gets it right. The agency that has been quietly transferred is not the dead poet's. It is the reader's ability to tell a reconstruction from a costume.

6 Limitations

The dating is coarse and is not really tested. Five non-contested changes cluster on three dates, 1700, 1750 and 1800. The rank correlation between date of death and bare acceptance is $\rho = -0.78$, $p = 0.118$, and the sign is the opposite of the naive prediction, with the oldest changes accepted most. With five points that describes five numbers rather than establishing a diachronic result. Answering which century properly needs changes spread across more dates than English conveniently provides.

CMUdict is General American, so rhymes alive in Received Pronunciation and dead in General American register as dead here. Pairs whose modern status depends on a secondary CMUdict listing are quarantined, and heterophonic homographs are excluded by name. The detector only finds dead rhymes already in the inventory, so real rates in both human and model verse are higher than reported. That applies identically to both sides, which makes the numbers floors.

Corpus attestation is a weak proxy for training exposure, since 2,816 poems stand in for a corpus nobody can inspect, which is why the load-bearing result is the unattested pairs staying high. Poems are dated by poet rather than composition, with floruit taken as birth plus forty: fine for aggregates, wrong for any single poem. The pre-1600 rise in the human curve is an artefact of Middle English spelling, so the era readback is restricted to the declining branch. The eye-rhyme cell has 57 pairs against 80 elsewhere, because bulletproof eye rhymes are scarce, and it appears in the most important contrast.

The generation ladder is one model per tier with one run each, sixty poems per cell (59 in one). "Capability" is a loose label for three models differing in more than size, and the intervals cover poems rather than repeated runs. The single-year readback is the weakest number here: Sonnet's interval maps to roughly 1830 through off the archaic end of the scale, because the human curve is flat enough that one percentage point spans a century. Quote the ordering and the Opus overshoot, not the year without its range.

Which number counts as the model's ear depends on the deliberation setting, and that is a result rather than a caveat. This study takes the unreflective judgment as the ear, on the grounds that recall of literary history is a different faculty. A reader who disagrees should treat the thinking-on numbers as the headline, and the argument is made in the same terms.

7 Related work

Whether text-only models carry usable phonological structure is settled. PhonologyBench [7] found models well above chance on grapheme-to-phoneme, syllable counting and rhyme generation while trailing humans by 17% on the last, and Phun-Bench [6] does the same for Chinese. McLaughlin et al. [4] locate phoneme-like directions in LLaMA 3.2's embedding space, and Liao & Shi [8] show subword tokenization constrains all of it. Each of those scores against present-day pronunciation, so a model judging rhyme in the accent of 1700 would lose points and the reason would be invisible. Here a mismatch with present-day English is the measurement rather than the error.

Phun-Bench reports that models excel at recalling correct pronunciations but struggle to use phonological knowledge the way human speakers do. The deliberation contrast here is an independent instance of that split, in English and on a historical axis: two languages, the same dissociation between a phonological record and a phonological reflex.

The poem corpus is the public-domain subset released with Bhyravajjula et al. [1], reused so the human baseline rests on a published corpus rather than a private scrape. The ground truth is standard historical phonology, chiefly Dobson [2], and no claim about any sound change here is original.

References

[1] Bhyravajjula, A., Walsh, M., Preus, A. & Antoniak, M. (2025) so much depends / upon / a whitespace: Why whitespace matters for poets and LLMs. In *Proceedings of EMNLP 2025*. https://aclanthology.org/2025.emnlp-main.1783/

[2] Dobson, E.J. (1968) *English Pronunciation 1500–1700*, 2nd ed., vol. II (Phonology). Oxford: Clarendon Press.

[3] Jespersen, O. (1909) *A Modern English Grammar on Historical Principles*, vol. I. Heidelberg: Carl Winter.

[4] McLaughlin, O., Khurana, A. & Merullo, J. (2025) I have no mouth, and I must rhyme: Uncovering internal phonetic representations in LLaMA 3.2. *arXiv preprint* arXiv:2508.02527.

[5] Lass, R. (1999) Phonology and morphology. In R. Lass (ed.), *The Cambridge History of the English Language*, vol. III. Cambridge University Press.

[6] Yue, X., Shen, Y. & Lu, W. (2026) Phun-Bench: Evaluating LLMs on phonological understanding in Chinese. In *Proceedings of ACL 2026*. *arXiv preprint* arXiv:2606.07300.

[7] Suvarna, A., Khandelwal, H. & Peng, N. (2024) PhonologyBench: Evaluating phonological skills of large language models. *arXiv preprint* arXiv:2404.02456.

[8] Liao, D. & Shi, F. (2026) How tokenization limits phonological knowledge representation in language models and how to improve them. In *Proceedings of ACL 2026*. *arXiv preprint* arXiv:2604.17105.

[9] Wells, J.C. (1982) *Accents of English*. Cambridge University Press.

[10] Wyld, H.C. (1936) *A History of Modern Colloquial English*, 3rd ed. Oxford: Blackwell.