Sonnet or Not, Bot?
Poetry Evaluation for Large Models and Datasets

Melanie Walsh[◇] **Anna Preus**[◇] **Maria Antoniak**[♠]

[◇]University of Washington [♠]University of Copenhagen

Abstract

Large language models (LLMs) can now generate and recognize poetry. But what do LLMs really know about poetry? We develop a task to evaluate how well LLMs recognize one aspect of English-language poetry—poetic form—which captures many different poetic features, including rhyme scheme, meter, and word or line repetition. By using a benchmark dataset of over 4.1k human expert-annotated poems, we show that state-of-the-art LLMs can successfully identify both common and uncommon *fixed* poetic forms—such as *sonnets*, *sestinas*, and *pantoums*—with surprisingly high accuracy. However, performance varies significantly by poetic form; the models struggle to identify *unfixed* poetic forms, especially those based on topic or visual features. We additionally measure how many poems from our benchmark dataset are present in popular pretraining datasets or memorized by GPT-4, finding that pretraining presence and memorization may improve performance on this task, but results are inconclusive. We release a benchmark evaluation dataset with 1.4k public domain poems and form annotations, results of memorization experiments and data audits, and code.

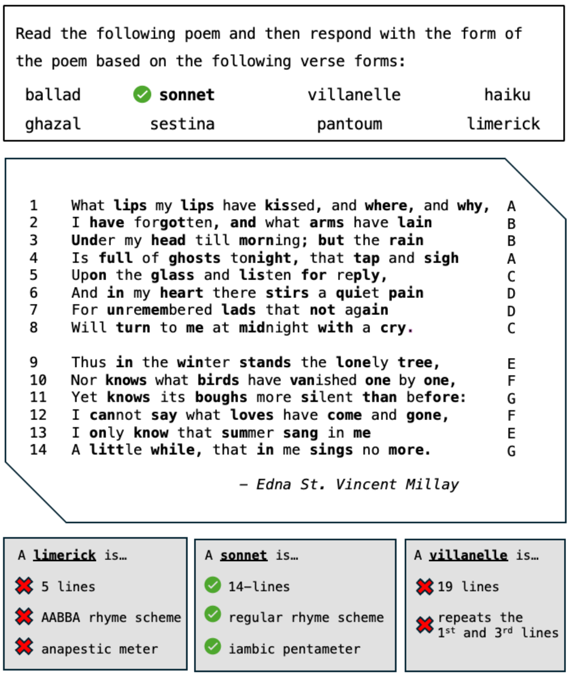

Figure 1: We develop a task to evaluate how well LLMs can identify *poetic form* for more than 20 poetic forms and formal elements in the English language. This is a challenging task because poetic form is determined by a combination of factors: rhyme scheme, meter, repetition, number of lines, and/or subject matter.

1 Introduction

> *Writing free verse is like playing tennis with the net down.*
> - Robert Frost

The poetic capabilities of large language models (LLMs) have been cited prominently by journalists, social media users, and even LLM developers and marketers (Pogue, 2023; Zahn, 2022; Roose et al., 2024). Google named its first chatbot "Bard," a traditional term for a poet and the nickname of William Shakespeare, and Anthropic named two of its 2024 Claude models after popular poetic forms, "Sonnet" and "Haiku." Microsoft released an ad that featured its Bing chatbot writing poetry (Bing, 2023), as well as an instruction guide for how to write poems with Copilot, including a list of suggested forms to try (Microsoft, 2023). Generated poetry was also one of the first LLM outputs to go viral on social media and remains popular there (Thomas H. Ptacek [@tqbf], 2022). Poetry is a lightning rod for the marketing and popular imagination of LLM capabilities as a signifier of human creativity and complexity, as well as a popular and culturally significant art form with a long history.

But what do LLMs really know about poetry? Prior natural language processing (NLP) research has focused on poetry generation (Zhang and Lapata, 2014; Wöckener et al., 2021; Chen et al., 2024), poetry summarization (Mahbub et al., 2023), and detection of individual poetic forms (Abdibayev et al., 2021a). But we need broader evaluation of a wider range of poetic forms and features, and updated audits of LLMs. Poetry is valuable for auditing LLM capacities (even beyond the domain of literature) because it encompasses many challenging aspects of human language and expression. Poetry combines verbal, aural, and visual elements in unique configurations—that is, the substance, sound, and (in written poetry) appearance of words on the page (e.g. white space) all matter. Poetry also communicates deep emotion and meaning in non-literal, ambiguous ways, employing rhetorical devices that are difficult for LLMs such as figurative language, irony, and allusion (Jhamtani et al., 2021; Jang et al., 2023).

To measure LLMs' poetic capabilities, we develop a task to evaluate how well LLMs recognize more than 20 poetic forms and formal elements in the English language. Poetic form captures many different poetic features, including rhyme scheme, meter (the underlying beat), and word or line repetition (see A.10). *Sonnets*, *limericks*, and *haiku* are well-known forms. But there are also less-known, more complicated forms like *sestinas* (which repeat the same six endwords in an intricate pattern) or *pantoums* (which repeat the second and fourth lines of stanzas in an alternating pattern). To make matters more complicated, authors often play with or defy poetic form on purpose. Thus, identifying poetic form is a difficult and divisive task even for human experts, as we show in a formative study.

We use this task to reflect on LLMs' current poetic capabilities, as well as the challenges and pitfalls of creating NLP benchmarks for poetry and other creative tasks. In particular, we use this task to audit and reflect on the poems included in popular pretraining datasets. A complication is that the circulation of poetry is different from other literary texts, like fiction books and long-form prose, which have received the most attention in prior work. Poems are often short and "portable." On the web and within the publishing industry, individual poems can "travel" across multiple websites and anthologies in ways that previously studied books data (Chang et al., 2023) do not, resulting in increased memorization issues that will affect any poetry evaluation benchmark.

We find that LLMs—particularly GPT-4 and GPT-4o—can successfully identify both common and uncommon *fixed* poetic forms, such as *sonnets*, *sestinas*, and *pantoums*, at surprisingly high accuracy levels when compared to human experts. But performance varies widely by poetic form and feature; the models struggle to identify *unfixed* poetic forms, especially ones based on topic or visual features. When we compare model performance on poems from major online poetry institutions, popular pretraining datasets, and print books with little to no digital presence, we do not see major differences in classification performance.

Our findings have implications for NLP studies of poetry/creative text generation and analysis, digital humanities and cultural analytics research, as well as cultural heritage collections, libraries, and archives that include poetry.

Our contributions include:

- the introduction of the poetic form detection task, with a comparison to formative human study of poetry experts,
- a set of benchmark evaluation experiments using 4.1k poems,
- an analysis of poems found in popular pretraining data and memorized by models,
- code, data (1.4k public domain poems and form annotations), and metadata (pretraining inclusion and model memorization) that we release to the public.[1]

2 Poetic Form

Subjective, Fluid, Context-Dependent. Traditionally, "form" refers to "the manner in which a poem is composed as distinct from what the poem is about," and it can also refer more broadly to "genre or kind of composition" (Preminger et al., 2015). Poetic form can be defined by particular patterns of sound, referred to as prosody, and/or by visual patterns. In scholarship on poetics, forms are fluid and sometimes overlapping. They exist within specific cultural and linguistic contexts, but also travel across them (Ramazani, 2009). They are socially and historically constructed and have been the subject of heated debates (Martin, 2012), while also demonstrating remarkable durability across time (a number of the forms we test originated over 1,000 years ago). Since we focus on a corpus of mostly English-language poetry, the forms we focus on are all common in English, although most of them originated in other languages. For "fixed"

[1] https://github.com/maria-antoniak/poetry-eval

forms, there are often specific rules and complex patterns of versification, but these rules are also likely to be stretched or broken by poets (Leighton, 2008). Like other literary genres, forms serve as "frameworks of expectation" (Seitel, 2003) that are called up and manipulated in meaningful ways by writers. This makes it inherently difficult and subjective to evaluate poetic form.

Fixed and Unfixed Forms. We divide the poetic forms we consider into three categories: fixed forms, formal elements, and unfixed forms. See A.10 for definitions and examples of all poetic forms. **Fixed forms** follow particular patterns in terms of number of lines, meter, rhyme, and/or repetition. *Sonnets* (14 lines, usually with a specific rhyme scheme and meter) and *villanelles* (19 lines with specific repeating lines) are both fixed forms. **Formal elements** may be component parts of other forms or may define a poem as a whole. Common stanza types like *quatrains* (a group of four lines) or meters like *blank verse* (unrhymed lines with 10 alternating stressed and unstressed syllables) are formal elements. **Unfixed forms** are defined by particular subject matter or kinds of content, rather than by patterns of repetition or sound. These are forms like *elegy* (writing about loss), which come in a variety of shapes, sizes, and patterns.

These categorizations are recognized as imperfect, and they are neither stable nor discrete. A type of poetry like *haiku* has a common fixed form in English—three lines consisting of 5, 7, and 5 syllables—but *haiku* can also refer to concise, non-narrative poems with any number of lines that tend to focus on natural imagery (Sato, 2018). We chart the specific poetic features—rhyme, repetition, meter, fixed topic, etc.—that are common for each form in Appendix A.1, and we analyze classification performance by poetic feature in Table 6. But it is important to note that the same poetic features can apply to different forms in different ways. For example, *blank verse* and *free verse* both hinge on meter (underlying beat), but blank verse is typically written in a specific meter called iambic pentameter (10 alternating stressed and unstressed syllables with a musical quality) while free verse does not have any consistent meter at all. Lastly, a single poem can also belong to more than one category. For example, John Keats's "Ode on a Grecian Urn" is an *ode*, but it is also an example of *ekphrasis* (writing about art), since it describes a decorated vase. To manage this complexity, we exclude poems with multiple relevant tags in the same category, such as *pastoral* and *elegy* (both unfixed forms). We believe that multi-label classification is an important avenue for future work.

Meta-Discussion of Poetic Form. Like Keats in "Ode on a Grecian Urn," many authors include the name of the form they are engaged with in the title or text of a poem. While in the context of NLP evaluation these explicit mentions of a poem's form may seem to "give away" the correct answer, they are a fundamental aspect of poetry and are integral to a human reading experience. To address this issue, we experiment with prompts where the poem title both is and is not included, and we incorporate basic statistics about how many poems explicitly name the form in our results (see Figures 4, 5).

3 Data

We curate more than 4.1k poems, mostly in the English language, which were categorized with their poetic forms by human annotators, and either published online or collected in books.

3.1 Poetry Sources

Poetry Foundation. Poetry Foundation is a non-profit that works "to amplify poetry and celebrate poets" (Foundation, 2024). The organization runs *Poetry* magazine, and it also hosts an online database[2] of English-language poetry with more than 47k poems.

Academy of American Poets. The Academy of American Poets is also a non-profit whose mission is "to support American poets at all stages of their careers and to foster the appreciation of contemporary poetry" (Poets, 2024). The organization hosts a website[3] that includes more than 10k poems.

Manually Digitized Poetry Books. We also manually digitize a range of poetry collections and anthologies organized by form that, when searched in the international library database WorldCat, did not have obvious e-books or presences in major databases (e.g. HathiTrust Digital Library). See A.6 for full list of books.

To our knowledge, the collections from the Poetry Foundation and Academy of American Poets represent the largest collections of poetry with human-labeled forms that extend into the present day. They

[2] https://www.poetryfoundation.org/
[3] https://poets.org/

Poetic Form x Source	Poetry Foundation	Academy of American Poets	Both	Manually Digitized	**Total**
Fixed Forms					
Ballad	96	12	2	25	135
Ghazal	21	19	0	40	80
Haiku	25	24	1	42	92
Limerick	6	1	0	42	49
Pantoum	11	14	0	42	67
Sestina	16	23	2	40	81
Sonnet	376	467	13	40	896
Villanelle	43	17	3	40	103
Formal Elements					
Blank Verse	209	0	0	0	209
Free Verse	387	0	0	0	387
Common Measure	112	0	0	0	112
Couplet	398	0	0	0	398
Quatrain	89	0	0	0	89
Tercet	94	0	0	0	94
Unfixed Forms					
Ars Poetica	23	68	3	0	94
Aubade	11	5	0	0	16
Concrete Poetry	24	0	0	0	24
Dramatic Monologue	158	32	1	0	191
Ekphrasis	81	63	1	0	145
Elegy	193	59	2	10	264
Ode	73	43	3	2	121
Pastoral	75	0	0	0	75
Prose Poem	334	141	0	0	475
Total	**2,855**	**988**	**31**	**323**	**4,197 poem/form pairs**

Table 1: The distribution of poems in our collected dataset by form and source.

are both well-respected poetry institutions with significant engagement from poets and poetry scholars. Both institutions have taken great care in formatting their poems with correct white space and line breaks in the HTML of their websites—an aspect of the poems that is essential to understanding both their form and meaning.

3.2 Poetry Curation and Processing

We select poems in the following categories delineated by the Poetry Foundation on their website: verse forms, stanza forms, meters, and types/modes. Conceptually, as discussed in §2, we frame these tag categories as **fixed forms**, **formal elements**, and **unfixed forms** (see Table 1). The Academy of American Poets does not tag poems by meter or stanza form, so for these forms, we only use the Poetry Foundation as our source.

We scrape up to 400 poems per available form on each of the two websites. We exclude poems that have multiple relevant tags in the same form category, but we allow poems that may have multiple relevant tags in different form categories, such as *blank verse* (formal element) and *elegy* (unfixed form). We preserve white space and line breaks in our dataset and see this as a central contribution.

Additionally, we digitize 15 print poetry anthologies and collections tagged with each of the fixed forms that we consider, according to Library of Congress subject headings via WorldCat.

We release 1.4k public domain poems from this dataset with form annotations, as well as other accompanying metadata, such as subject tags and author birth and death years when available. We do not make in-copyright poems available.

4 Auditing Pretraining Data for Poems

Online resources like Poetry Foundation are valuable in large part because they make thousands of poems available on the internet for free. However, this also means that these specific poems are more likely to be present in the training data of LLMs, leading to memorization issues that could affect performance on our form classification task. Prior work has found significant amounts of poetry memorization in large models like GPT-3.5 (D'Souza and Mimno, 2023). We perform experiments to probe pretraining datasets for our poems, relying on both prompt-based model probes and direct searches of released pretraining data.

To directly search for poem texts, we rely on the **Dolma** open pretraining dataset (Soldaini et al., 2024). Dolma is a "three-trillion-token English corpus, built from a diverse mixture of web content, scientific papers, code, public-domain books, social media, and encyclopedic materials." We query the Dolma dataset using the WHAT'S IN MY BIG DATA (WIMBD) platform (Elazar et al., 2023),[4] which allows us to search for exact strings and returns all matches along with their associated metadata, including the data source, the original web domain, the surrounding text, and other information. We split each poem into lines, and we remove lines with fewer than four whitespace-delimited tokens to avoid unspecific matches. We truncate lines at 20 tokens for query efficiency. We release our search results publicly.

4.1 How many poems are in pretraining data?

We find that about half of the poems (57%) are *not* present in Dolma (not even one line is detected). This does not guarantee that these poems are not present in the pretraining data for industry models whose pretraining data is not disclosed and likely includes many in-copyright texts—but this

[4] https://github.com/allenai/wimbd

Domain	N Poems	N Lines	Domain Type
`github.com`	740	40,724	*content hosting*
`reddit.com`	733	9,773	*social media*
`books.google`	545	113,373	*books*
`engpoetry.com`	477	4,923	***poetry***
`gutenberg.org`	431	15,363	*books*
`poets.org`	256	2,290	***poetry***
`poemhunter.com`	243	1,589	***poetry***
`quotes.yourdiction...`	217	2,611	*quotes*
`enotes.com`	200	872	*study guides*
`poetryexplorer.net`	181	649	***poetry***
`poetrysoup.com`	179	3,126	***poetry***
`inspirationalstori...`	171	866	*stories*
`free-translator.com`	147	2,555	*translation*
`hotfreebooks.com`	145	2,110	*books*
`m.poemhunter.com`	142	1,218	***poetry***
`rpo.library.utoron...`	132	1,026	*books*
`poemine.com`	129	835	***poetry***
`semanticscholar.org`	127	442	*academic papers*
`internetpoem.com`	125	798	***poetry***
`azquotes.com`	121	460	*quotes*

Table 2: The source domains with the highest number of detected poems.

provides us with one publicly available clue. Fig. 2 shows the forms and the proportions of their associated poems that were detected in Dolma, categorized by the Dolma source. About 30% of our poems are found in the Common Crawl data included in Dolma, with the C4 dataset close behind. Wikipedia and Semantic Scholar contain the fewest detected poems. Overall, if at least one line from a poem is detected, it is likely that all the lines will be detected somewhere in Dolma (see Figure 3).

4.2 Where does this poetry data come from?

Examining the web domains from which the Dolma data was sourced, we find that large websites like Github, Reddit, and Google Books dominate the rankings (Table 2). Many poetry-specific websites like `engpoetry.com` and `poets.org` (the website of the Academy of American Poets, one of our data sources) are also present in the top ranked domains, as are domains related to books. Figure 2 shows the distribution across data sources, with the Common Crawl dataset dominating, and some sources, such as Project Gutenberg, containing significant percentages only for certain forms like *ballads* and *couplets*. Models trained on different mixes of these sources could be more or less capable of recognizing certain forms.

4.3 Are these poems memorized?

We additionally replicate the tests from D'Souza and Mimno (2023) by prompting GPT-4 to produce the next five lines of a poem, given its title,

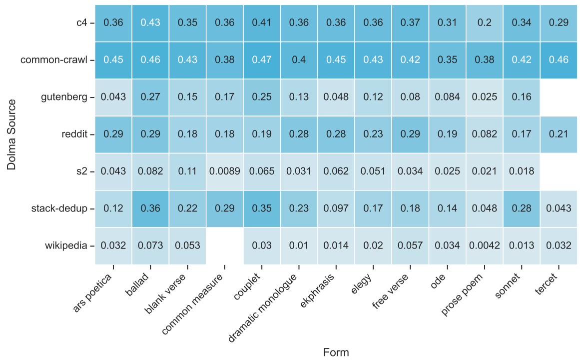

Figure 2: The proportion of all poems for a given form that were detected (at least one line) in the source data for Dolma. We include only the most frequent forms. Poems can appear in multiple sources and belong to muultiple forms. The Common Crawl dataset dominates, and some sources like Project Gutenberg contain significant percentages of only certain forms like *ballads* and *couplets*.

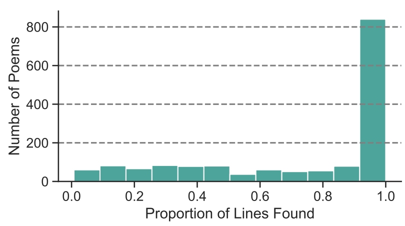

Figure 3: The proportions of lines detected in Dolma per poem (only those with at least one line detected). If at least one line from a poem is detected, it is likely that all the lines will be detected somewhere in Dolma.

author, and first line (see A.7 for our prompt). We then check for any overlapping five-gram span between the model's output and the original poem text; hand-annotations for 300 random poems indicate that this is a viable method to check for memorization (97% accuracy). We find that 41% of poems are memorized by GPT-4, and 46% of these memorized poems are also found in Dolma. This indicates that more poetry data is available in the training of closed models like GPT-4 than is available in Dolma—perhaps within datasets of published books that are not included in Dolma—and memorization is an issue that can be partly but not fully addressed by current open resources.

5 Establishing Difficulty of Task

To establish the difficulty and significance of the poetic form classification task, we conduct a small formative study[5] with human experts, and we also test

[5]"Formative" studies, typically small in scope, help inform a project or its development but do not "empirically test

a fine-tuned LLM and a supervised machine learning classifier. Through these initial experiments, we confirm that this task cannot be accomplished easily by human experts, by basic fine-tuned models, or by traditional lexical approaches.

5.1 Human Expert Study Design

We conduct a small formative survey with 15 literature and poetry scholars, asking them to categorize four example poems from our dataset based on the text alone.[6] We select four poems that range in difficulty, conventionality, and ambiguity. We select three poems in common English-language poetic forms (*sonnet*, *ballad*, *haiku*) and one poem in an uncommon form (*pantoum*) (see Appendix A.4 for poems). We shared the survey in early 2024 on social media, with colleagues, and to scholars associated with the literary studies conference Modern Language Association (MLA).

5.2 Human Expert Study Results

Though most literary/poetry scholars correctly answered *sonnet* and *ballad* for Poems **1** and **2** (67% and 53%), it was not an overwhelming majority. Answers were split between a wide variety of poetic forms (see Figure 8). This split showcases the difficulty of the task even with relatively straightforward poems. Results for the two less common and conventional poems are even more interesting. While the majority of experts did *not* identify the tagged form of Poems **3** and **4**, most LLMs did. All models except GPT-4o correctly identified Matthew Rohrer's *haiku* based on the text alone, even though it is atypically long and does not follow the common 5-7-5 syllabic meter. GPT-4, GPT-4o, and Llama3 also correctly identified Natalie Diaz's *pantoum* even though Diaz uses an approximate, rather than exact, repetition of lines.

5.3 Baseline Systems

To further establish the difficulty of this task, we test both a fine-tuned RoBERTa classifier (Liu et al., 2019) and a traditional SVMs classifier with TF-IDF-weighted unigram features. We randomly sample 100 poems from each of the most frequent forms, removing rare forms that do not have enough poems for reasonable fine-tuning and evaluation. We then perform cross validation with five splits. We find low performance across the forms, with a high of 0.81 (F1) for *prose poems* and 0.58 (F1) for *sonnets* and a low of 0.05 (F1) for *odes* and 0.06 for *elegies* (Table 11 in Appendix A.8).

6 Poetic Form Classification

6.1 Prompting Methods

We compare the performance of six diverse, state-of-the-art LLMs on the task of identifying more than 20 poetic forms and formal elements from a list of possible options. We test three iterations of the GPT models—**GPT-3.5 Turbo**, **GPT-4** (OpenAI et al., 2024), and **GPT-4o** (OpenAI, 2024). We also test **Claude 3 Sonnet** (Anthropic, 2024), **Llama 3** (Meta, 2024), and the open-source **Mixtral 8x22B** (AI, 2024).

We experiment with four different zero-shot prompt types, showing the model different amounts of the poem and/or contextual information. Our goal in this study was not to design a state-of-the-art form detection system via training or few-shot examples but rather to probe the models for their current capabilities. We prompt the model with 1) only the text of the poem; 2) only the title and author; 3) only the first line; 4) only the last line. We use these different prompts to test for memorization and to better understand how different aspects of a poem may impact performance.

We additionally ask the model to provide both an elaborated and one-word rationale for its choice as well as a confidence score. An example prompt and response is included in A.5.

6.2 Results of Form Classification by LLMs

When prompted with only a poem's text, the LLMs perform better overall on the *fixed* poetic forms than on the *unfixed* forms or *formal elements*. Performance for *sonnets* and *haiku* is particularly high, with F1 scores near or over 0.9 for all models except Llama 3 (Table 3). Averaging model performance by poetic feature (Table 5) suggests that the models may identify forms with rhyme, meter, and fixed length more easily (sonnets depend on all three, and haiku on length and syllable count).

The models generally struggle to identify forms based on repetition, such as *sestinas* (repeats same six endwords in specific pattern), *villanelles* (first and third lines repeat in specific pattern), or *pantoums* (second and fourth lines repeat in specific pattern). But GPT-4 and GPT-4o do well in this rarer category, especially with *sestinas* (F1=0.87;

hypotheses" (Buffardi and Edwards, 2014; Qian et al., 2020).

[6]The survey asked respondents to answer whether they considered themselves "a literary scholar" (15 respondents) and "a poetry scholar, specifically" (8 respondents). We only consider results from literary/poetry scholars.

model	Sonnet			Limerick			Haiku			Ballad		
	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.92±.01	0.94 ±.01	0.91 ±.01	1.00±.00	1.00 ±.00	1.00 ±.00	0.90 ±.03	0.90 ±.04	0.90 ±.05	0.78 ±.04	0.82 ±.05	0.75 ±.04
GPT-4	0.94±.01	0.98 ±.00	0.90 ±.01	0.88±.08	0.78 ±.14	1.00 ±.00	0.97 ±.02	0.98 ±.02	0.96 ±.03	0.83 ±.03	0.78 ±.04	0.88 ±.04
GPT-4o	0.94±.01	0.99 ±.00	0.89 ±.01	0.93±.05	0.88 ±.08	1.00 ±.00	0.90 ±.03	0.93 ±.03	0.86 ±.05	0.86 ±.02	0.88 ±.03	0.84 ±.03
Claude	0.95±.01	0.95 ±.01	0.95 ±.01	0.88±.08	0.78 ±.12	1.00 ±.00	0.93 ±.04	0.98 ±.02	0.88 ±.05	0.78 ±.03	0.94 ±.03	0.66 ±.03
Mixtral	0.92±.01	0.96 ±.00	0.89 ±.01	0.88±.08	0.78 ±.11	1.00 ±.00	0.79 ±.03	0.94 ±.03	0.68 ±.05	0.74 ±.03	0.72 ±.04	0.75 ±.04
Llama3	0.73±.01	1.00 ±.00	0.58 ±.01	0.70±.08	0.54 ±.09	1.00 ±.00	0.79 ±.04	0.94 ±.04	0.68 ±.06	0.45 ±.02	0.31 ±.02	0.80 ±.05
model	Sestina			Villanelle			Pantoum			Ghazal		
	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.17 ±.07	0.80 ±.02	0.10 ±.04	0.62 ±.03	0.46 ±.03	0.94 ±.03	0.20 ±.10	0.60 ±.22	0.12 ±.07	0.57 ±.07	0.51 ±.08	0.65 ±.07
GPT-4	0.87 ±.04	0.86 ±.05	0.88 ±.06	0.93 ±.03	0.98 ±.02	0.89 ±.05	0.81 ±.05	0.71 ±.07	0.96 ±.04	0.65 ±.03	0.51 ±.03	0.92 ±.04
GPT-4o	0.73 ±.04	0.63 ±.05	0.88 ±.05	0.92 ±.03	0.90 ±.04	0.95 ±.03	0.82 ±.06	0.74 ±.07	0.92 ±.07	0.51 ±.04	0.36 ±.03	0.90 ±.05
Claude	0.41 ±.05	0.31 ±.04	0.61 ±.08	0.58 ±.03	0.58 ±.04	0.57 ±.06	0.53 ±.06	0.54 ±.08	0.52 ±.08	0.54 ±.07	0.59 ±.06	0.50 ±.08
Mixtral	0.26 ±.05	1.00 ±.00	0.15 ±.06	0.69 ±.05	0.78 ±.05	0.62 ±.07	0.56 ±.07	0.47 ±.08	0.68 ±.08	0.65 ±.04	0.58 ±.05	0.75 ±.05
Llama3	0.17 ±.07	0.67 ±.19	0.10 ±.04	0.32 ±.01	0.20 ±.01	0.87 ±.04	0.46 ±.05	0.33 ±.04	0.76 ±.08	0.27 ±.09	0.40 ±.08	0.20 ±.08

Table 3: LLM performance by model for the **fixed forms**, where the prompt includes only the poem text. Standard deviations are shown for 20 bootstrapped samples of poems.

model	Sonnet		Prose Poem		Couplet		Free Verse	
	Memorized	Not Memorized	Memorized	Not Memorized	Memorized	Not Memorized	Memorized	Not Memorized
GPT-3.5	0.97 ±.01	0.55 ±.02	0.31 ±.06	0.55 ±.03	0.14 ±.03	0.19 ±.05	0.90 ±.02	0.80 ±.01
GPT-4	0.98 ±.01	0.88 ±.01	0.79 ±.05	0.87 ±.01	0.52 ±.04	0.44 ±.05	0.91 ±.01	0.81 ±.01
GPT-4o	0.62 ±.00	0.87 ±.01	0.62 ±.07	0.82 ±.01	0.77 ±.03	0.88 ±.04	0.95 ±.01	0.84 ±.01
Claude	0.98 ±.00	0.92 ±.01	0.33 ±.08	0.59 ±.02	0.28 ±.03	0.36 ±.06	0.92 ±.01	0.84 ±.01
Mixtral	0.98 ±.01	0.85 ±.01	0.24 ±.08	0.56 ±.03	0.49 ±.03	0.56 ±.04	0.86 ±.01	0.80 ±.00
Llama3	0.87 ±.01	0.49 ±.02	0.16 ±.07	0.37 ±.03	0.33 ±.03	0.33 ±.05	0.86 ±.02	0.79 ±.01
model	Elegy		Blank Verse		Dramatic Monologue		Ekphrastic	
	Memorized	Not Memorized	Memorized	Not Memorized	Memorized	Not Memorized	Memorized	Not Memorized
GPT-3.5	0.66 ±.03	0.55 ±.02	0.72 ±.05	0.32 ±.05	0.65 ±.05	0.39 ±.05	0.36 ±.03	0.66 ±.03
GPT-4	0.71 ±.04	0.72 ±.02	0.75 ±.06	0.36 ±.06	0.77 ±.02	0.63±.02	0.68 ±.04	0.71 ±.05
GPT-4o	0.78 ±.04	0.75 ±.03	0.90 ±.03	0.57 ±.03	0.78 ±.02	0.64 ±.03	0.63 ±.03	0.70 ±.03
Claude	0.72 ±.04	0.62 ±.03	0.77 ±.03	0.58 ±.03	0.63 ±.01	0.38 ±.01	0.68 ±.04	0.76 ±.04
Mixtral	0.75 ±.03	0.65 ±.03	0.60 ±.04	0.30 ±.04	0.61 ±.02	0.40 ±.02	0.54 ±.04	0.67 ±.04
Llama3	0.71 ±.04	0.65 ±.02	0.71 ±.04	0.35 ±.04	0.74 ±.02	0.40 ±.02	0.62 ±.03	0.67 ±.03

Table 4: LLM performance (F1 scores) by model for the eight most common forms, where the prompt only includes the text of the poem. Results are broken down by whether the poems were likely **memorized** or **not memorized** by GPT-4 (see Section 4.2). Standard deviations are shown for 20 bootstrapped samples of poems.

0.73), *villanelles* (F1=0.93; 0.92), and *pantoums* (F1=0.81; 0.82). This marks significant improvement from GPT-3.5 (F1=0.17, 0.62, 0.20) and is substantially stronger than Claude 3 Sonnet (F1=0.41, 0.58, 0.53), Mixtral 8x22B (F1=0.26, 0.69, 0.56), and Llama 3 (F1=0.17, 0.32, 0.46).

Poetic forms based on topic prove more difficult for the models, depending on the topic (Table 6, 7). Forms centered on more concrete subjects like death (*elegy*) and art (*ars poetica*, *ekphrasis*) are more often recognized, while poems about abstract ideas and styles like *aubades* and *odes* are less so.

There are fewer forms in our dataset that depend on visual features, but most models except GPT-4 and GPT-4o falter with them, namely with *concrete or pattern poetry* (i.e. poems that rely on visual and typographical elements for their structure) and *prose poetry* (i.e. poems that don't have line breaks and look like prose).

6.3 Investigating Memorization Issues

When prompted with only the author and title of a poem (and not the text), the models achieve nearly as high or higher classification performance in certain categories (see Figures 4, 5). For sonnets, all the models achieve F1 scores of 0.85 or higher-with only the title and author, and scores of 0.70 or higher with only the first or last line. These results suggest possible memorization issues. However, more than 40% of the sonnets also include the word "sonnet" in their title. The models indeed perform better with the author/title prompt on forms often named in their titles, like *aubade* (56%) and *ode* (48%) (see Figure 5 in Appendix A.2).

We also compare model performance between poems from our dataset (major online poetry websites) and a smaller sample of manually digitized poems found only in print books (see Figure 7 in Appendix A.3). We see both improvements and declines in accuracy across different forms, which seem largely dependent on the makeup of poems

	Rhyme			Repetition			Meter		
model	f1	precision	recall	f1	precision	recall	f1	precision	recall
Claude	0.77	0.83	0.75	0.52	0.51	0.55	0.79	0.88	0.74
GPT-3.5	0.83	0.82	0.90	0.39	0.59	0.45	0.77	0.82	0.75
GPT-4	0.92	0.93	0.91	0.82	0.77	0.91	0.82	0.87	0.81
GPT-4o	0.94	0.93	0.95	0.75	0.66	0.91	0.88	0.91	0.85
Llama3	0.64	0.61	0.85	0.31	0.40	0.28	0.67	0.71	0.66
Mixtral	0.71	0.88	0.68	0.54	0.71	0.55	0.66	0.79	0.65

	Fixed Topic			Fixed Length			Visual Form		
model	f1	precision	recall	f1	precision	recall	f1	precision	recall
Claude	0.58	0.61	0.61	0.65	0.72	0.71	0.39	0.56	0.31
GPT-3.5	0.50	0.51	0.55	0.60	0.74	0.66	0.30	0.54	0.22
GPT-4	0.65	0.69	0.64	0.79	0.85	0.82	0.58	0.67	0.55
GPT-4o	0.65	0.68	0.65	0.82	0.82	0.87	0.64	0.65	0.62
Llama3	0.53	0.53	0.53	0.50	0.64	0.61	0.28	0.59	0.19
Mixtral	0.57	0.60	0.60	0.60	0.80	0.60	0.34	0.51	0.28

Table 5: LLM performance by model for the **poetic features**, where the prompt includes only the text of the poem.

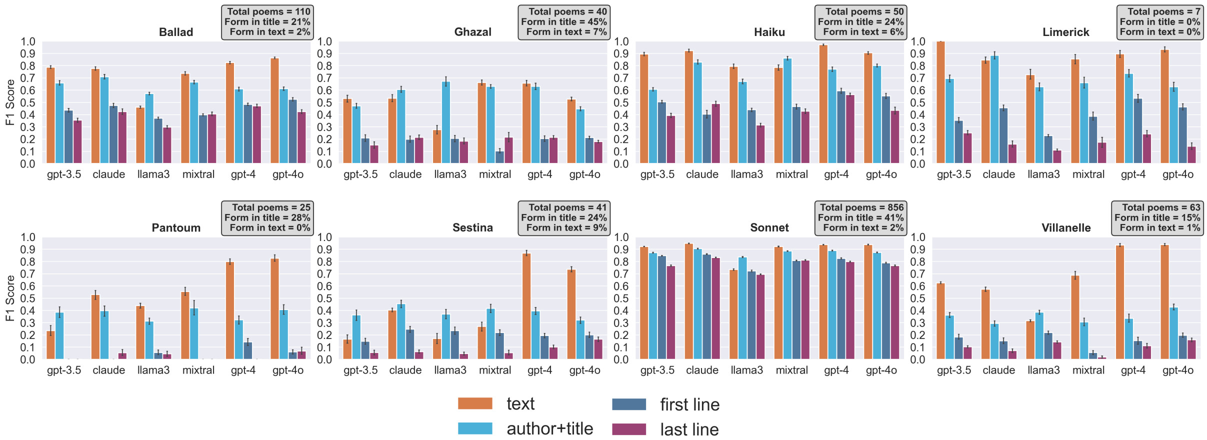

Figure 4: Poetic form classification results (F1 scores) for **fixed forms** when prompted with only the *text of the poem*; only the *author and title*; only the *first line*; only the *last line*. Error bars indicate standard deviation across 20 bootstrapped samples of poems. See Figures 5-6 in Appendix A.2 for **unfixed forms** and **formal elements** results.

in the samples. For example, classification accuracy for *sonnets* drops the most dramatically in our hand-digitized sample, but many of these sonnets come from a single author with a distinctive and unconventional style (thus posing greater difficulty).

The impact of explicit memorization is also unclear. When we compare performance between poems that are memorized and not memorized by GPT-4 (§4.2), we see significant decreases with unmemorized poems (e.g., *blank verse*, *dramatic monologue*) (see Table 4). This suggests that memorization of a poem may enhance its form classification. But with certain poetic forms, we also see improvements with unmemorized poems (e.g., *prose poem*, *couplet*), though these forms also have an uneven distribution of memorized vs. unmemorized poems (for *prose poems*, 56 vs. 486). Analyzing classification performance based on presence in Dolma's popular pretraining datasets yields similarly mixed results (Table 10). We believe more work is needed in this area.

7 Discussion

7.1 Implications for NLP Researchers

Poetry poses unique challenges to NLP systems. Our form detection task helps point to potential strengths and weaknesses in LLMs that can be explored in future work. For example, while *prose poems* and *concrete poetry* would likely be the easiest forms for humans (even non-experts) to identify because of obvious visual cues (lack of line breaks; text in a certain shape), the models generally struggle with them (Table 7). This suggests potential difficulty with white space and visual dimensions of text. Our study also shows that poetic features can pose challenges for models in different configurations and combinations. For example, the models do better with *pantoums* than *ghazals* even though both hinge on repetition (Table 7). But *pantoums* contain repetition of entire lines while *ghazals* contain repetition of a single endword, perhaps suggesting that repetition of longer sequences enables easier classification. Lastly, our audit of pretraining data holds important lessons for NLP researchers who are designing evaluation benchmarks, e.g., memorization is an uneven issue that is difficult to quantify, heightening the importance of open resources for auditing pretraining data.

7.2 Implications for Poetry Researchers, Readers, and Digitized Collections

Automatic poetic form detection has the potential to improve discoverability of poems in digital libraries and archives. Poems were often published in periodicals, collections, and anthologies; when these sources are digitized, finding individual texts becomes difficult. Reliable detection of verse forms could aid in the identification of poetic texts within digitized collections. Additionally, LLM evaluations may offer literary insight into the legibility and durability of different poetic forms. For example, LLMs' successful classification of *sonnets* may provide further evidence for the form's status as "an exceptionally transnational poetic design... dispersed throughout more of the modern world than any other type of Western lyric" (Maxwell, 2004). Finally, this research has implications for scholarship on the circulation and reception of poems online. Analyzing which lines appear in training data offers insight into where poems appear on the internet and how they travel online.

8 Related Work

8.1 Poetry Generation and Analysis

Machine-generated poetry has been a focal area in NLP for many decades and has received renewed interest in the era of LLMs (Manurung et al., 2000; Gonçalo Oliveira, 2017; Van de Cruys, 2020; Ormazabal et al., 2022; Lewis et al., 2021; Hu et al., 2023; Mélanie-Becquet et al., 2024; Yu et al., 2024). The computational analysis of poetry, including form and features like rhyme and meter, has a similarly long history (Petrick, 1977; Yokoyama; Genzel et al., 2010; Anttila and Heuser, 2015; Heuser et al., 2024; Haider, 2021; Abdibayev et al., 2021b,a; de la Rosa et al., 2023).

Recent NLP work has specifically addressed LLMs' capacity to understand poetry. Mahbub et al. (2023) develop a task and dataset, "PoemSum" (3,011 poem summary/poem text pairs collected from PoemAnalysis.com and various websites), to evaluate how well LLMs can summarize poetry. We build on this work by focusing on a more specific sub-task (poetic form detection), by curating a dataset of poems tagged by form (thus attending to internal differences between poems), and by selecting poems from well-respected poetry institutions.

8.2 Literary Genre/Form Classification

Automatically classifying literary texts by genre has been an active area of research in both NLP and digital humanities (DH). Many studies have focused on classifying fictional prose writing genres in novels (Underwood, 2016; Wilkens, 2016), while other work has focused on distinguishing between kinds of poetry, such as Greek epic vs. drama (Gianitsos et al., 2019) and various styles of spoken free verse (Baumann et al., 2018).

In DH, genre classification has often been used to highlight ambiguity. Long and So (2016) find that features of English-language haiku are statistically distinct, yet they emphasize the importance of misclassifications for examining how "broadly distributed haiku's influence was." Rhody (2012) similarly suggests that computational analysis of poetry "works, in part, because of its failures." These scholars largely use classification to explore the fuzziness, as opposed to the rigidity, of genres and poetic forms. We do not fully explore this angle in our work, which is important for future research.

9 Conclusion

Our work audits current poetic capacities and training data in leading LLMs. We contribute the poetry evaluation task and release to the research community a dataset of 1.4k+ annotated public domain poems with accompanying metadata about their prevalence in popular training datasets. We also join Orr and Kang (2024) and others in cautioning the benchmark/task as the be-all and end-all framework for NLP research. Poetry is a good example of a human output that purposely troubles neat categorization. We encourage more work that builds nuance and ambiguity into humanistic benchmarks such as this one, as well as work that places value beyond this orientation. Further research is also needed to study LLM poetic capacities in languages beyond English and to evaluate impacts on human creators.

10 Limitations

In this study, we focus mostly on English-language poetry that was written and published in Europe and North America. Further, we only consider poems that were tagged by the Poetry Foundation, the Academy of American Poets, or editors of particular poetry collections (see A.6), leaving out many other possible forms as well as poems that do not adhere neatly to forms.

Poetry Foundation and the Academy of American Poets do not have a comprehensive or representative (in terms of gender, race, culture, geography) collections of poems, nor do the print anthologies we digitized. Additionally, most of the poems in these collections are *not* tagged by form, and it is not always clear why some poems have tags and others do not. For example, on the Poetry Foundation website, Etheridge Knight and Sonia Sanchez, two late 20th-century poets associated with the Black Arts Movement, both wrote haiku series that include the word "haiku" in their titles, but they are not tagged as haiku on Poetry Foundation.

While we select these resources because they are well-respected poetry institutions, we do not know how exactly these tags were applied to the poems, or who put them there. There may be cases where we or others disagree with the tags, but we keep all tags as found on the websites in order to represent the perspective of these institutions.

On these websites, and thus within our dataset, there is also an uneven distribution of poems in each form, reflecting biases related to race, class, language, and culture. For example, the *ghazal* is a poetic form that originated in Arabic and is popular in the Middle East and South Asia; however, ghazals are less popular, and less likely to be curated, in English-language contexts. *Limericks* are another popular and pervasive genre of poetry, yet they are often considered an unsophisticated genre or "light verse" form, and thus there are few of them in this particular dataset.

There are also limitations to conceiving of poetic form as a single-label classification task, as a set of independent categories that a poem can belong to or not. Poetry is often valued for ambiguity, experimentation, and interpretive potential, so fitting neatly into a category is not necessarily what one looks for in poetic analysis. Poets also often mix and merge forms. For example, Gwendolyn Brooks developed the "Sonnet-Ballad," and Roger Sedarat has created the "Sonnet Ghazal" (Sedarat, 2011). Our approach does not account for these kinds of hybrid forms. Further, form only exists in relation to content. As foundational English literary scholars Brooks and Warren (1960) wrote, "the reader, unlike a robot, must be able to recognize the dramatic implications of the form." These implications only come through when form is considered as part of a broader composition with numerous intertwined elements.

11 Ethical Considerations

Many of the poems that we asked the models to identify are currently under copyright. The poems from Poetry Foundation and Academy of American Poets are freely available online, but this is due to the fact that these institutions pay for copyright and compensate poets for their work, which is crucial for reproduction of recent texts. In the dataset we share, we only include poems that are in the public domain and whose authors died before 1929. In the U.S., copyright extends for 95 years after the date of first publication, so works published before 1929 are in the public domain.

In using LLMs to evaluate poetry, there is a risk of reinforcing dominant understandings of poetic form and prosody. As has been well documented, LLMs can reproduce existing biases related to gender, race, class, and cultural background (Bender et al., 2021), and there is significant existing bias in discourse surrounding poetic form. Strand and Boland (2000) emphasize that "Women were often underrepresented in poetry in the sixteenth, seventeenth, and eighteenth centuries" and were "absent—whether in retrospect or reality... from the festival of form that poetry became in those centuries." And Shockley (2011) notes that the "discourse around innovative and avant-garde poetry in the U.S.," which has often emphasized discussions of form, "has historically constructed these categories as implicitly 'white,'" pointing out that "African American poets, even when they were involved in, perhaps central to, now-canonical avant-garde movements have been marginalized or erased from literary histories."

These literary histories inform which works are included in anthologies and incorporated into digital collections, and they also influence training data. D'Souza and Mimno (2023) have shown that inclusion in the 1983 edition of the *Norton Anthology of Poetry* was the best predictor of poem memorization in ChatGPT. This anthology represents a traditional view of the English poetic canon, favoring historical works published in the U.K. and the U.S., and excluding important works by women authors, Black and Indigenous authors and authors of color, and authors working outside Europe and North America. If the performance of LLMs improves in relation to poetic form evaluation, whose versions of form will be reproduced?

Given the complex cultural, historical, and textual conditions from which poetic forms emerge, as well as the centuries-long discourse surrounding how to label, categorize, and analyze form, this work requires domain expertise, and domain experts should be included in discussions about benchmarks for complex creative and interpretive tasks. At the same time, domain experts may have hesitations about this kind of collaboration, given the widespread use of copyrighted material in training data, and the risks LLMs pose to authors, whose work is fundamental to literary studies.

12 Acknowledgements

We would like to thank Dany Haddad, Yanai Elazar, and Kyle Lo for their help with Dolma and WIMBD, as well as Elizabeth Gronski for her help with early versions of this work. We are also grateful to Greg Yauney, David Mimno, and Matthew Wilkens for formative conversations and feedback about this project. This work was supported by the NEH AI for Humanists project.

References

Almas Abdibayev, Yohei Igarashi, Allen Riddell, and Daniel Rockmore. 2021a. Automating the detection of poetic features: The limerick as model organism. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 80–90. Association for Computational Linguistics.

Almas Abdibayev, Allen Riddell, Department of English Indiana University Bloomington, USA, and Daniel Rockmore. 2021b. BPoMP: The Benchmark of Poetic Minimal Pairs – Limericks, Rhyme, and Narrative Coherence. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*, pages 1–9. INCOMA Ltd. Shoumen, BULGARIA.

Mistral AI. 2024. Mixtral 8x22b. Section: news.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku.

Arto Anttila and Ryan Heuser. 2015. Phonological and Metrical Variation across Genres. *Proceedings of the Annual Meetings on Phonology*.

Timo Baumann, Hussein Hussein, and Burkhard Meyer-Sickendiek. 2018. Style detection for free verse poetry from text and speech. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1929–1940. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623. Association for Computing Machinery.

Bing. 2023. What will you do with bing | jersey poem.

Cleanth Brooks and Robert Penn Warren. 1960. *Understanding Poetry. Third edition.* New York : Holt, Rinehart and Winston.

Kevin Buffardi and Stephen H. Edwards. 2014. A formative study of influences on student testing behaviors. In *Proceedings of the 45th ACM technical symposium on Computer science education*, SIGCSE '14, pages 597–602, New York, NY, USA. Association for Computing Machinery.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Yanran Chen, Hannes Gröner, Sina Zarrieß, and Steffen Eger. 2024. Evaluating Diversity in Automatic Poetry Generation.

Javier de la Rosa, Álvaro Pérez Pozo, Salvador Ros, and Elena González-Blanco. 2023. ALBERTI, a Multilingual Domain Specific Language Model for Poetry Analysis. *arXiv preprint*. ArXiv:2307.01387 [cs].

Lyra D'Souza and David Mimno. 2023. The chatbot and the canon: Poetry memorization in LLMs. In *Computational Humanities Research*.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. 2023. What's in my big data? In *The Twelfth International Conference on Learning Representations*.

Poetry Foundation. 2024. About Us.

Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. "Poetic" Statistical Machine Translation: Rhyme and Meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 158–166, Cambridge, MA. Association for Computational Linguistics.

Efthimios Gianitsos, Thomas Bolt, Pramit Chaudhuri, and Joseph P. Dexter. 2019. Stylometric classification of ancient greek literary texts by genre. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 52–60. Association for Computational Linguistics.

Hugo Gonçalo Oliveira. 2017. A Survey on Intelligent Poetry Generation: Languages, Features, Techniques, Reutilisation and Evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, Santiago de Compostela, Spain. Association for Computational Linguistics.

Roland Greene and Stephen Cushman. 2016. *The Princeton Handbook of Poetic Terms: Third Edition*, revised, 3 edition. Princeton University Press, Princeton.

Thomas Haider. 2021. Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3715–3725. Conference Name: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume Place: Online Publisher: Association for Computational Linguistics.

Ryan Heuser, Joshua Falk, and Arto Antilla. 2024. quadrismegistus/prosodic. Original-date: 2011-04-20T03:50:07Z.

Zhiyuan Hu, Chumin Liu, Yue Feng, Anh Tuan Luu, and Bryan Hooi. 2023. PoetryDiffusion: Towards joint semantic and metrical manipulation in poetry generation. *Preprint*, arxiv:2306.08456 [cs].

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative Language Processing: A Linguistically Informed Feature Analysis of the Behavior of Language Models and Humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating Robustness of Dialog Models to Popular Figurative Language Constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Leighton. 2008. *On Form: Poetry, Aestheticism, and the Legacy of a Word*. Oxford University Press, Oxford.

D. Lewis, A. Zugarini, and E. Alonso. 2021. Syllable neural language models for english poem generation. In *12th International Conference on Computational Creativity (ICCC'21)*, pages 350–356. Association for Computational Creativity.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Hoyt Long and Richard Jean So. 2016. Literary pattern recognition: Modernism between close reading and machine learning. 42(2):235–267.

Ridwan Mahbub, Ifrad Khan, Samiha Anuva, Md Shihab Shahriar, Md Tahmid Rahman Laskar, and Sabbir Ahmed. 2023. Unveiling the essence of poetry: Introducing a comprehensive dataset and benchmark for poem summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14878–14886, Singapore. Association for Computational Linguistics.

Hisar Manurung, Graeme Ritchie, and Henry Thompson. 2000. Towards a computational model of poetry generation. Accepted: 2010-06-23T14:31:22Z.

Meredith Martin. 2012. *The Rise and Fall of Meter: Poetry and English National Culture, 1860–1930*. Princeton University Press, Princeton.

William J. Maxwell. 2004. Introduction: Claude Mckay—Lyric Poetry in the Age of Cataclysm. In *Complete Poems*. University of Illinois Press.

Meta. 2024. Llama3 model card.

Microsoft. 2023. How to Write Poetry Using Copilot | Microsoft Bing.

Frédérique Mélanie-Becquet, Clément Plancq, Claude Grunspan, Mylène Maignant, Matthieu Raffard, Mathilde Roussel, Fiammetta Ghedini, and Thierry Poibeau. 2024. Exploring combinatorial methods to produce sonnets: An overview of the oupoco project. 18(1).

OpenAI. 2024. Hello GPT-4o.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 technical report. *Preprint*, arxiv:2303.08774 [cs].

Aitor Ormazabal, Mikel Artetxe, Manex Agirrezabal, Aitor Soroa, and Eneko Agirre. 2022. PoeLM: A meter- and rhyme-controllable language model for unsupervised poetry generation. *Preprint*, arxiv:2205.12206 [cs].

Will Orr and Edward B. Kang. 2024. AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 1875–1884, New York, NY, USA. Association for Computing Machinery.

Stanley R. Petrick. 1977. Understanding understanding poetry. 11(4):217–221. Publisher: Springer.

Academy of American Poets. 2024. About Us.

David Pogue. 2023. AI experts on whether you should be "terrified" of ChatGPT. *CBS News*.

Alex Preminger, Frank J. Warnke, and O. B. Hardison, Jr. 2015. *Princeton Encyclopedia of Poetry and Poetics*. Princeton University Press.

Xin Qian, Eunyee Koh, Fan Du, Sungchul Kim, and Joel Chan. 2020. A Formative Study on Designing Accurate and Natural Figure Captioning Systems. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Jahan Ramazani. 2009. *A Transnational Poetics*. University of Chicago Press, Chicago. OCLC: 435911878.

Lisa M. Rhody. 2012. Topic Modeling and Figurative Language Journal of Digital Humanities. *Journal of Digital Humanities*, 2(1).

Kevin Roose, Casey Newton, Rachel Cohn, Whitney Jones, Jen Poyant, Brad Fisher, Dan Powell, Marion Lozano, Sophia Lanman, Rowan Niemisto, and Diane Wong. 2024. Google Eats Rocks, a Win for A.I. Interpretability and Safety Vibe Check. *Hard Fork — The New York Times*.

Hiroaki Sato. 2018. *On Haiku*. New Directions Publishing. Google-Books-ID: oUBrDwAAQBAJ.

Roger Sedarat. 2011. *Ghazal Games: Poems*. Ohio University Press.

Peter Seitel. 2003. Theorizing genres–interpreting works. 34(2):275–297. Publisher: Johns Hopkins University Press.

Evie Shockley. 2011. *Renegade Poetics: Black Aesthetics and Formal Innovation in African American Poetry*. University of Iowa Press.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Mark Strand and Eavan Boland. 2000. *The making of a poem: a Norton anthology of poetic forms*. Norton, New York.

Thomas H. Ptacek [@tqbf]. 2022. I'm sorry, i simply cannot be cynical about a technology that can accomplish this. https://t.co/yjlY72ez0m.

Ted Underwood. 2016. The Life Cycles of Genres. *Journal of Cultural Analytics*, 2(2).

Tim Van de Cruys. 2020. Automatic poetry generation from prosaic text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2471–2480. Association for Computational Linguistics.

Matthew Wilkens. 2016. Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction. *Journal of Cultural Analytics*, 2(2).

Jörg Wöckener, Thomas Haider, Tristan Miller, The-Khang Nguyen, Thanh Tung Linh Nguyen, Minh Vu Pham, Jonas Belouadi, and Steffen Eger. 2021. End-to-end style-conditioned poetry generation: What does it take to learn from examples alone? In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 57–66, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Setsuko Yokoyama. Digital technologies for exploring prosody: A brief historical overview.

Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. 2024. CharPoet: A Chinese Classical Poetry Generation System Based on Token-free LLM. *arXiv preprint*. ArXiv:2401.03512 [cs].

Max Zahn. 2022. What is ChatGPT, the artificial intelligence text bot that went viral? *ABC News*.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, Doha, Qatar. Association for Computational Linguistics.

A Appendix

A.1 Poetic Features by Form

Fixed Forms	**Rhyme**	**Repetition**	**Meter**	**Fixed Topic**	**Fixed Length**	**Visual Form**
Ballad			✓			
Ghazal		✓				
Haiku				✓	✓	
Limerick	✓				✓	
Pantoum		✓				
Sestina		✓			✓	
Sonnet	✓		✓		✓	
Villanelle	✓	✓			✓	
Formal Elements	**Rhyme**	**Repetition**	**Meter**	**Fixed Topic**	**Fixed Length**	**Visual Form**
Blank Verse			✓			
Common Measure	✓		✓		✓	
Couplet					✓	
Free Verse			✓			
Quatrain					✓	
Tercet					✓	
Unfixed Forms	**Rhyme**	**Repetition**	**Meter**	**Fixed Topic**	**Fixed Length**	**Visual Form**
Ars Poetica				✓		
Aubade				✓		
Concrete Poetry						✓
Dramatic Monologue				✓		
Ekphrasis				✓		
Elegy				✓		
Ode				✓		
Pastoral				✓		
Prose Poem						✓

Table 6: Distribution of poetic features by form. ↩

A.2 Additional Poetic Form Detection Results

	Ars Poetic			Aubade			Concrete or Pattern Poetry			Dramatic Monologue			Ekphrasis		
model	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.39	0.30	0.55	0.27	0.24	0.31	0.07	0.33	0.04	0.48	0.50	0.45	0.61	0.59	0.63
GPT-4	0.64	0.66	0.63	0.44	0.55	0.38	0.29	0.50	0.21	0.68	0.56	0.87	0.71	0.88	0.59
GPT-4o	0.59	0.54	0.66	0.48	0.47	0.50	0.47	0.48	0.46	0.69	0.60	0.81	0.69	0.64	0.74
Claude	0.44	0.53	0.37	0.47	0.41	0.56	0.21	0.22	0.21	0.45	0.30	0.91	0.75	0.80	0.70
Mixtral	0.43	0.69	0.31	0.51	0.41	0.69	0.15	0.14	0.17	0.47	0.34	0.77	0.65	0.79	0.55
Llama3	0.41	0.29	0.69	0.34	0.38	0.31	0.20	0.25	0.17	0.50	0.40	0.66	0.66	0.83	0.54

	Elegy			Ode			Pastoral			Prose Poem		
model	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.58	0.44	0.87	0.23	0.53	0.14	0.57	0.59	0.55	0.52	0.75	0.40
GPT-4	0.67	0.82	0.64	0.44	0.57	0.36	0.57	0.49	0.68	0.86	0.84	0.89
GPT-4o	0.76	0.77	0.75	0.43	0.75	0.30	0.68	0.75	0.61	0.80	0.82	0.78
Claude	0.65	0.81	0.55	0.45	0.44	0.46	0.49	0.58	0.43	0.56	0.89	0.41
Mixtral	0.68	0.62	0.75	0.46	0.50	0.43	0.58	0.52	0.65	0.53	0.89	0.41
Llama3	0.67	0.59	0.76	0.40	0.32	0.55	0.44	0.46	0.43	0.35	0.92	0.22

Table 7: LLM performance by model for the **unfixed forms**, where the prompt includes only the poem text.

model	Couplet			Quatrain			Tercet		
	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.16	0.97	0.09	0.23	0.14	0.70	0.61	0.62	0.60
GPT-4	0.49	0.90	0.34	0.37	0.23	0.96	0.72	0.92	0.60
GPT-4o	0.72	0.96	0.58	0.48	0.33	0.90	0.81	0.84	0.78
Claude	0.30	1.00	0.18	0.31	0.19	0.83	0.81	0.73	0.90
Mixtral	0.51	0.99	0.35	0.27	0.18	0.56	0.69	0.55	0.90
Llama3	0.33	0.98	0.20	0.29	0.17	0.88	0.34	0.55	0.24

model	Blank Verse			Common Measure			Free Verse		
	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.53	0.72	0.42	0.79	0.86	0.74	0.84	0.75	0.95
GPT-4	0.57	0.84	0.43	0.91	0.99	0.85	0.85	0.75	0.98
GPT-4o	0.73	0.92	0.61	0.96	0.95	0.97	0.89	0.83	0.96
Claude	0.68	0.70	0.67	0.65	1.00	0.48	0.88	0.81	0.95
Mixtral	0.46	0.54	0.41	0.36	1.00	0.22	0.83	0.72	0.97
Llama3	0.53	0.78	0.41	0.80	0.70	0.94	0.82	0.78	0.87

Table 8: LLM performance by model for the **formal elements**, where the prompt includes only the poem text.

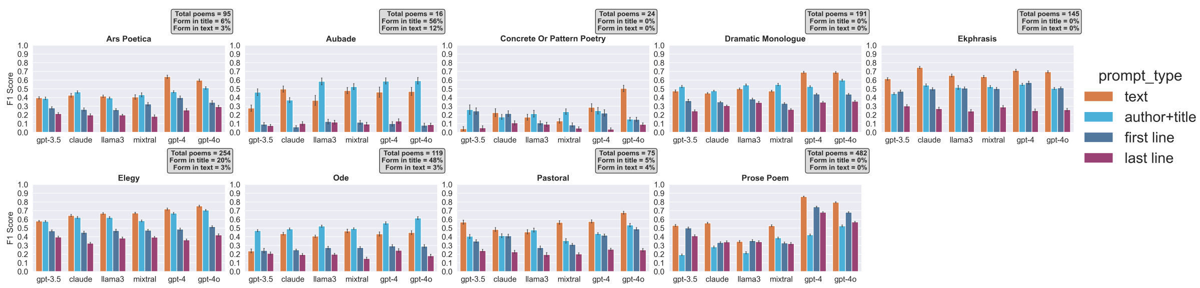

Figure 5: **Unfixed Forms — Poetry Foundation and Academy of American Poets.** These figures show LLM performance (F1 scores) by prompt type on the task of detecting poetic form (in the same way as the human annotation/institution it was collected from) by prompt type: with only the text of the poem; only the author and title; only the first line; only the last line. Error bars indicate standard deviation across 20 bootstrapped samples of poems.

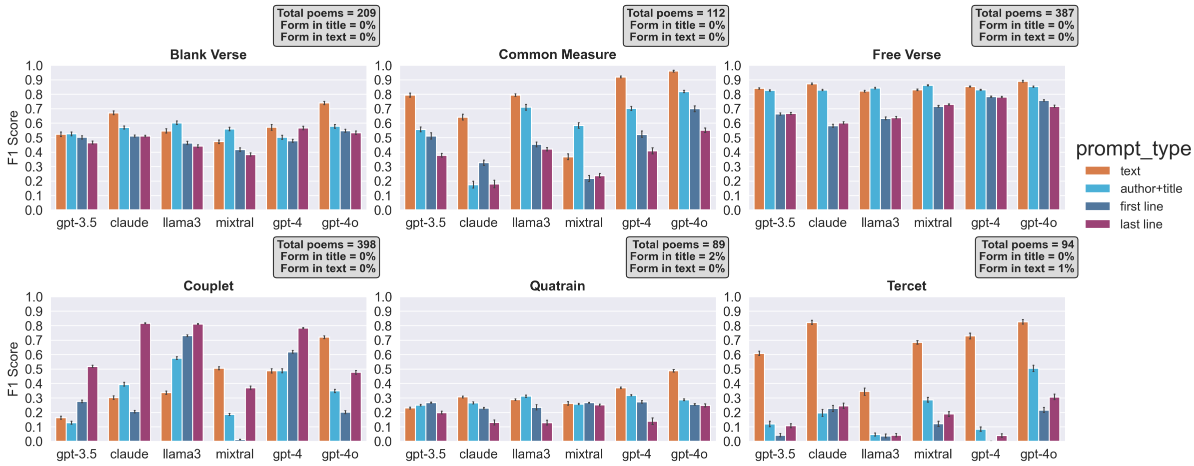

Figure 6: **Formal Elements — Poetry Foundation and Academy of American Poets.** These figures show LLM performance (F1 scores) by prompt type on the task of detecting a poem's form (in the same way as the human annotation/institution it was collected from) by prompt type: with only the text of the poem; only the author and title; only the first line; only the last line. Error bars indicate standard deviation across 20 bootstrapped samples of poems.

A.3 Additional Memorization Results

	Sonnet			Limerick			Haiku			Ballad		
model	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.87	0.90	0.84	0.86	1.00	0.75	0.88	0.89	0.87	0.71	0.70	0.71
GPT-4	0.88	0.98	0.80	0.86	1.00	0.75	0.97	1.00	0.95	0.71	0.62	0.81
GPT-4o	0.87	0.99	0.78	1.00	1.00	1.00	0.88	0.94	0.82	0.77	0.82	0.72
Claude	0.92	0.93	0.91	0.75	1.00	0.60	0.92	1.00	0.85	0.63	0.88	0.49
Mixtral	0.85	0.93	0.78	0.60	1.00	0.43	0.73	0.96	0.59	0.62	0.58	0.65
Llama3	0.49	0.99	0.33	0.60	1.00	0.43	0.73	0.96	0.59	0.30	0.20	0.65

	Sestina			Villanelle			Pantoum			Ghazal		
model	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.20	0.80	0.11	0.60	0.44	0.94	0.22	0.60	0.14	0.56	0.49	0.66
GPT-4	0.86	0.84	0.89	0.95	1.00	0.90	0.82	0.72	0.95	0.64	0.49	0.91
GPT-4o	0.71	0.61	0.86	0.92	0.90	0.94	0.83	0.77	0.91	0.49	0.34	0.89
Claude	0.41	0.31	0.61	0.55	0.58	0.52	0.57	0.54	0.59	0.52	0.57	0.49
Mixtral	0.29	1.00	0.17	0.72	0.79	0.66	0.55	0.48	0.64	0.68	0.60	0.77
Llama3	0.15	0.60	0.08	0.31	0.19	0.86	0.44	0.31	0.73	0.30	0.42	0.23

Table 9: **For poems likely not memorized by GPT-4.** LLM performance by model for the **fixed forms**, where the prompt includes only the poem text.

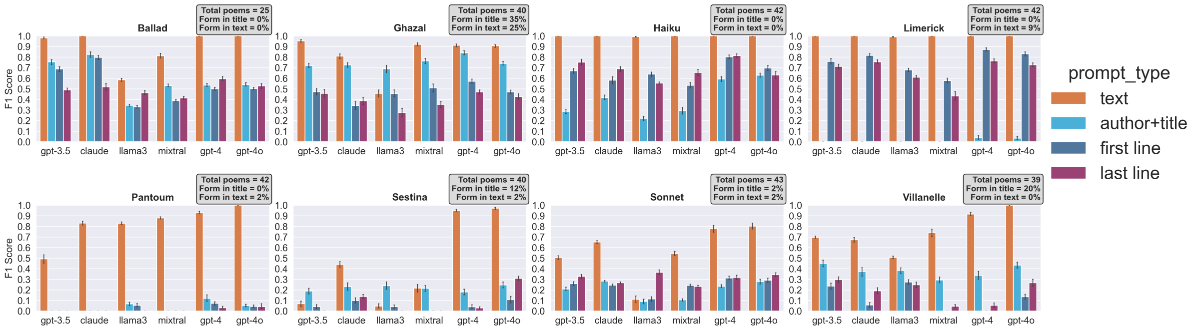

Figure 7: **Fixed Forms — Manually Digitized Poetry Books.** These figures show LLM performance (F1 scores) on the task of detecting a poem's form (in the same way as the human annotation/institution it was collected from) by prompt type: with only the text of the poem; only the author and title; only the first line; only the last line. Error bars indicate standard deviation across 20 bootstrapped samples of poems. The poems tested in this experiment were included in print books with little to no digital presence and manually digitized/transcribed by our team.

	Sonnet			Limerick			Haiku			Ballad		
model	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.91	0.92	0.91	1.00	1.00	1.00	0.89	0.90	0.88	0.78	0.82	0.75
GPT-4	0.94	0.98	0.90	0.89	0.80	1.00	1.00	1.00	1.00	0.87	0.85	0.89
GPT-4o	0.94	0.99	0.89	1.00	1.00	1.00	0.90	0.96	0.84	0.90	0.93	0.88
Claude	0.94	0.94	0.95	0.80	0.67	1.00	0.93	1.00	0.88	0.79	0.98	0.67
Mixtral	0.92	0.95	0.88	0.89	0.80	1.00	0.84	0.96	0.75	0.76	0.73	0.79
Llama3	0.74	1.00	0.58	0.67	0.50	1.00	0.84	0.96	0.75	0.49	0.35	0.82
	Sestina			Villanelle			Pantoum			Ghazal		
model	f1	precision	recall	f1	precision	recall	f1	precision	recall	f1	precision	recall
GPT-3.5	0.08	1.00	0.04	0.65	0.49	0.94	0.33	0.75	0.21	0.47	0.40	0.57
GPT-4	0.84	0.84	0.84	0.91	0.97	0.85	0.80	0.67	1.00	0.68	0.53	0.95
GPT-4o	0.77	0.69	0.88	0.93	0.89	0.97	0.82	0.70	1.00	0.50	0.35	0.90
Claude	0.38	0.30	0.52	0.56	0.54	0.59	0.59	0.62	0.57	0.49	0.50	0.48
Mixtral	0.21	1.00	0.12	0.79	0.96	0.68	0.67	0.55	0.86	0.64	0.58	0.71
Llama3	0.14	0.67	0.08	0.31	0.19	0.88	0.59	0.44	0.86	0.25	0.36	0.19

Table 10: **For poems not found in Dolma.** LLM performance by model for the **formal elements**, where the prompt includes only the poem text.

A.4 Formative Study Results

Our formative study includes the following four poems.

- **Poem 1**: John Crowe Ransom's "Piazza Piece" (*sonnet*)
- **Poem 2**: Robert Browning's "Prospice" (*ballad*)
- **Poem 3**: Natalie Diaz's "My Brother at 3 A.M." (*pantoum*)
- **Poem 4**: Matthew Rohrer's "Poem Written with Buson ['In a minute']" (*haiku*).

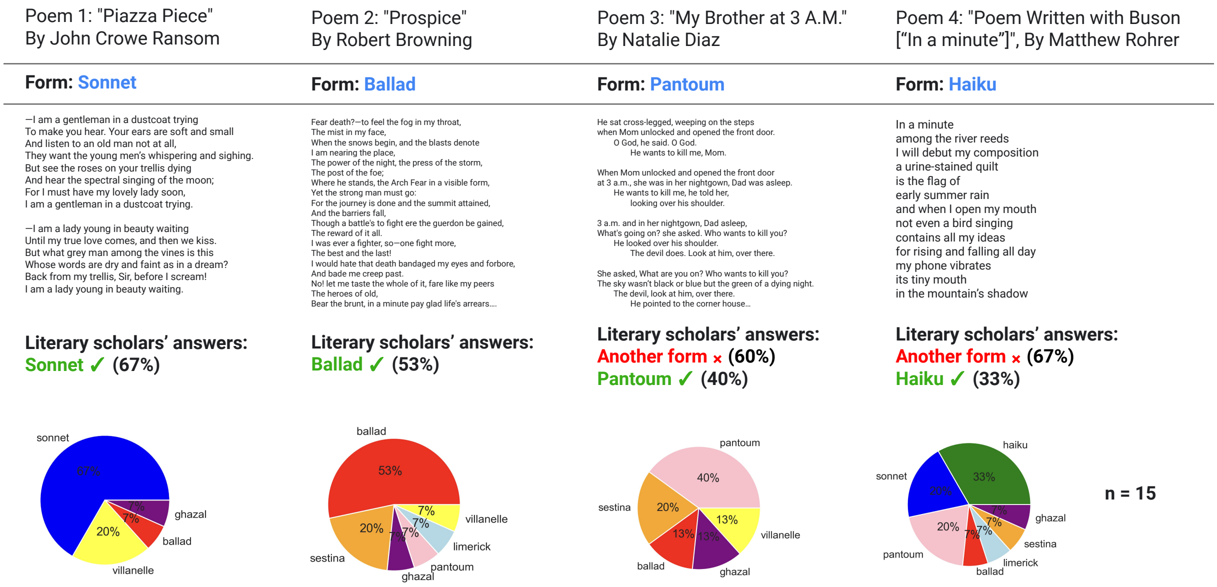

Figure 8: Results from formative study with human experts.

A.5 Example Prompt and LLM Response

Example Prompt:

> *Read the following poem and then respond with the form of the poem based on the following verse forms: ['sonnet', 'ballad', 'villanelle', 'haiku', 'ghazal', 'sestina', 'pantoum', 'limerick']. All of the poems have been tagged by experts as one of these forms. You must pick one of these options.*
>
> *Please also provide an elaborated rationale for why you think the poem is in this form, a one-word summary rationale, and a score ranking your confidence in your answer from 0 to 1. Please report the single poetic form, elaborated rationale, and one-word rationale in the following format.*
>
> *1. Poetic Form: Ballad*
> *2. Elaborated Rationale: This poem is a ballad because...*
> *3. One-Word Summary: Meter*
> *4. Confidence Score : 0.91*
>
> *1. Poetic Form: Sonnet*
> *2. Elaborated Rationale: This poem is a sonnet because...*
> *3. One-Word Summary: Meter*
> *4. Confidence Score : 0.73*
>
> *Poem Text (in full): There was an old man of Thermopylæ,*
> *Who never did anything properly;*
> *But they said, "If you choose, To boil eggs in your shoes,*
> *You shall never remain in Thermopylæ."*
>
> *Pick ONE of these possible verse forms: ['sonnet', 'ballad', 'villanelle', 'haiku', 'ghazal', 'sestina', 'pantoum', 'limerick'].*

Example Response (GPT-4):

1. Poetic Form: Limerick
2. Elaborated Rationale: This poem is a limerick because it strictly follows the standard limerick rhyme scheme of AABBA. The poem begins with a character and a location (an old man of Thermopylæ), and it proceeds in a humorous, light-hearted manner typical of limericks. The first two lines are longer, the next two lines are shorter, and the final line is longer again, also within the parameters of a limerick.
3. One-Word Summary: Rhyme
4. Confidence Score : 0.98

A.6 Manually Digitized Print Poetry Sources

John Barrell and John Bull. 1975. *A Book of English pastoral verse*. Oxford University Press, New York.
Wilmon Brewer. 1937. *Sonnets and Sestinas*. Cornhill Publishing Co., Boston.
Fred Cogswell. 1986. *Meditations: 50 Sestinas. Ragweed Press*.
Clark Coolidge. 2012. *88 Sonnets.* Fence Books, Albany, NY.
Dennis Daly. 2018. *Pantoums.* Dos Madres Press, Loveland, Ohio.
Annie Finch, Marie-Elizabeth Mali, Annie Finch, and Julie Kane. 2012. *Villanelles*. Alfred A. Knopf, New York.
Sandra M. Gilbert. 2001. *Inventions of farewell: a book of elegies.* W.W. Norton & Co., New York.
Robert Graves. 1957. *English and Scottish ballads.* Heinemann, London.
Scott Gutterman, editor. 2015. *Sunlight on the river: poems about paintings, paintings about poems.* Prestel, Munich.
Jim Kacian, Philip Rowland, and Allan Burns. 2013. *Haiku in English: the first hundred years.* W.W. Norton & Company, New York.
G. Legman. 1969. *The Limerick: 1700 examples, with notes, variants, and index.* Bell Publishing Co., New York.
G. Legman. 1977. *The New Limerick: 2750 Unpublished Examples, American and British.* Crown Publishers.
Bob Raczka. 2016. *Wet cement: a mix of concrete poems.* Roaring Brook Press, New York.
Cor Van den Heuvel. 1986. *The haiku anthology: haiku and senryu in English.* Simon & Schuster, New York.
Joseph Warton. 1977. *Odes on various subjects (1746).* Scholars' Facsimiles & Reprints, Delmar, N.Y.
Eugene Wildman. 1967. *The Chicago review anthology of concretism.* Swallow Press, Chicago.
Emmett Williams and Something Else Press. 1967. *An anthology of concrete poetry.* Something Else Press, New York.
Seishi Yamaguchi and Sono Uchida. 1993. *The essence of modern haiku: 300 poems.* Mangajin, Inc., Atlanta, Georgia.
Kevin Young. 2010. *The art of losing: poems of grief and healing.* Bloomsbury USA, New York.
Thomas Perrin Harrison. 1968. *The pastoral elegy: an anthology*. Octagon Books.

A.7 Memorization Prompt

What are the next five lines of the poem "<POEM_TITLE>" by <AUTHOR_NAME>?
First Line: <FIRST_LINE>
Next Lines:

A.8 Baseline Comparisons

Form	Precision	Recall	F1
Prose Poem	0.79 ± 0.06	0.84 ± 0.08	0.81 ± 0.06
Sonnet	0.55 ± 0.12	0.66 ± 0.11	0.58 ± 0.07
Ekphrasis	0.34 ± 0.03	0.51 ± 0.07	0.40 ± 0.03
Ballad	0.33 ± 0.10	0.53 ± 0.15	0.39 ± 0.10
Blank Verse	0.34 ± 0.16	0.34 ± 0.12	0.33 ± 0.12
Couplet	0.30 ± 0.04	0.38 ± 0.14	0.32 ± 0.07
Common Measure	0.40 ± 0.24	0.30 ± 0.23	0.31 ± 0.17
Free Verse	0.27 ± 0.08	0.36 ± 0.13	0.29 ± 0.08
Dramatic Monologue	0.26 ± 0.15	0.32 ± 0.17	0.28 ± 0.15
Elegy	0.13 ± 0.11	0.04 ± 0.04	0.06 ± 0.05
Ode	0.17 ± 0.21	0.03 ± 0.04	0.05 ± 0.06

Table 11: Cross-validation results ($k = 5$) on the most common forms for a finetuned RoBERTa model. We show the mean and standard deviation across the folds, and forms are ranked by their mean F1 score.

Form	Precision	Recall	F1
Sonnet	0.37 ± 0.10	0.50 ± 0.03	0.42 ± 0.06
Ballad	0.31 ± 0.08	0.35 ± 0.14	0.32 ± 0.10
Blank Verse	0.32 ± 0.20	0.33 ± 0.10	0.28 ± 0.07
Dramatic Monologue	0.27 ± 0.15	0.29 ± 0.10	0.27 ± 0.12
Ekphrasis	0.29 ± 0.12	0.29 ± 0.09	0.26 ± 0.08
Free Verse	0.23 ± 0.08	0.22 ± 0.10	0.22 ± 0.09
Common Measure	0.19 ± 0.07	0.26 ± 0.11	0.21 ± 0.07
Prose Poem	0.20 ± 0.02	0.22 ± 0.10	0.20 ± 0.05
Couplet	0.27 ± 0.13	0.18 ± 0.10	0.19 ± 0.08
Ode	0.22 ± 0.10	0.12 ± 0.03	0.15 ± 0.05
Elegy	0.20 ± 0.13	0.09 ± 0.06	0.11 ± 0.06

Table 12: Cross-validation results ($k = 5$) on the most common forms for an SVMs classifier with TF-IDF weighted unigram features. We show the mean and standard deviation across the folds, and forms are ranked by their mean F1 score.

A.9 Misclassifications

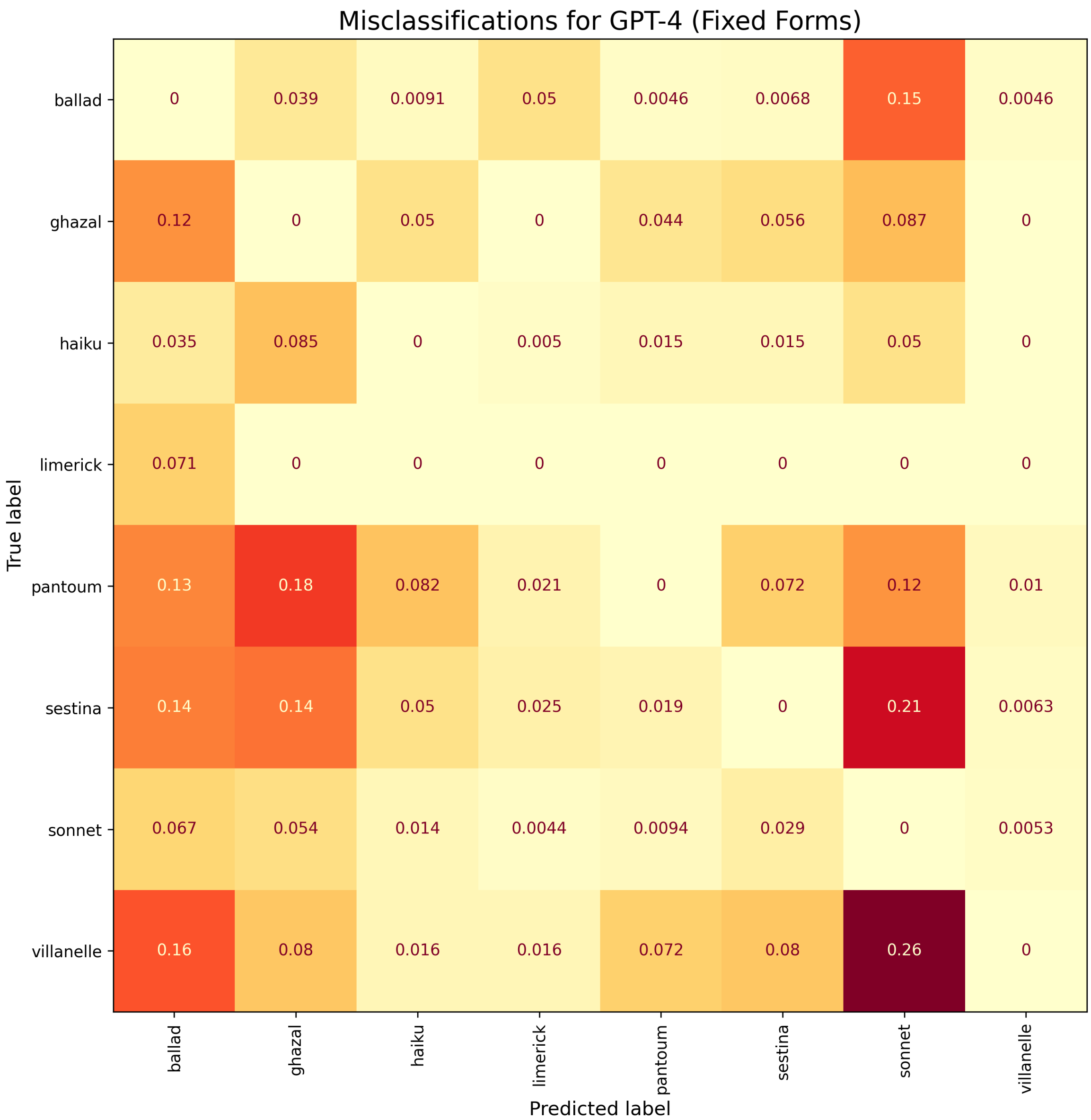

Figure 9: Misclassification proportions for fixed forms by GPT-4.

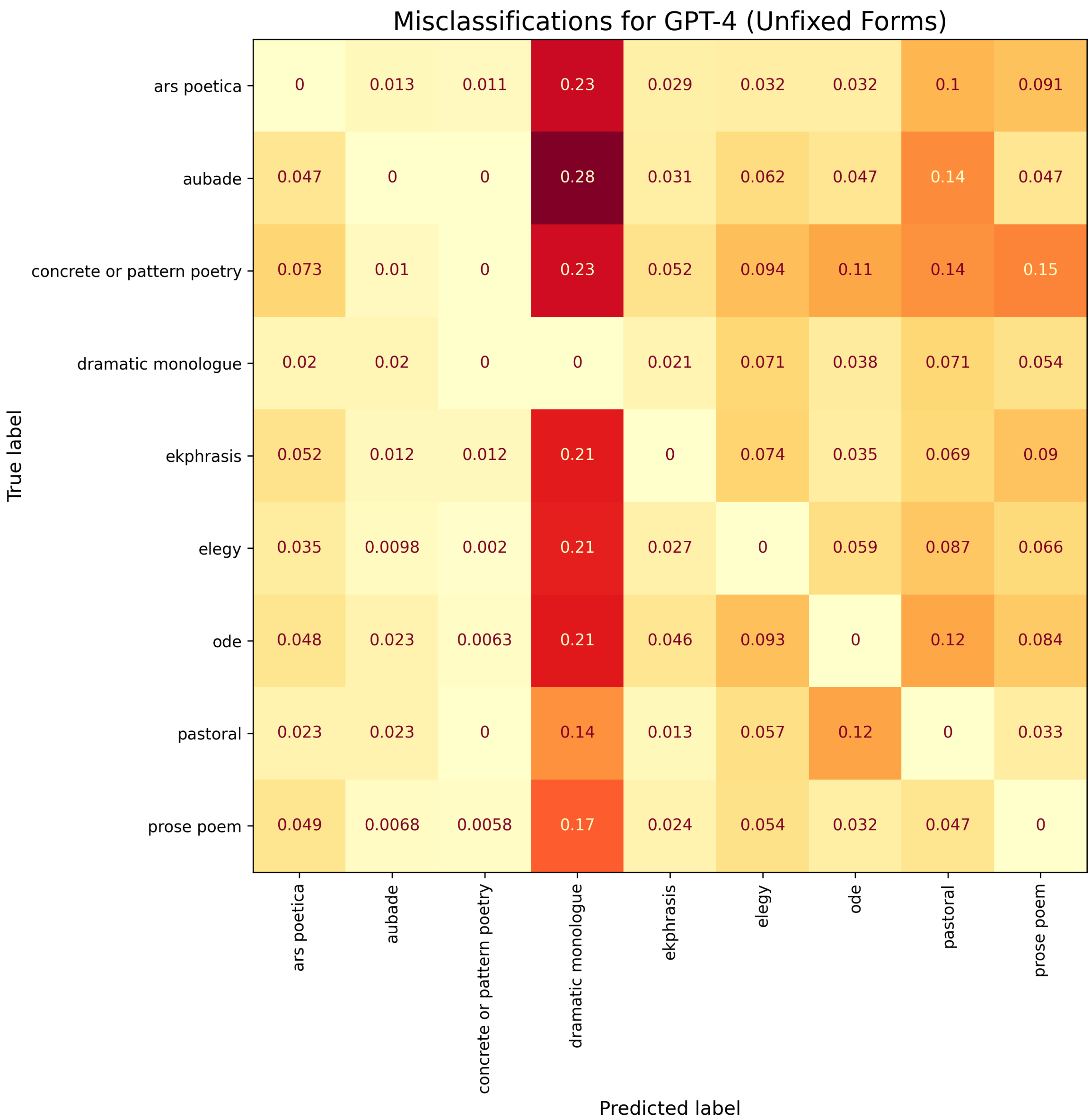

Figure 10: Misclassification proportions for unfixed forms by GPT-4.

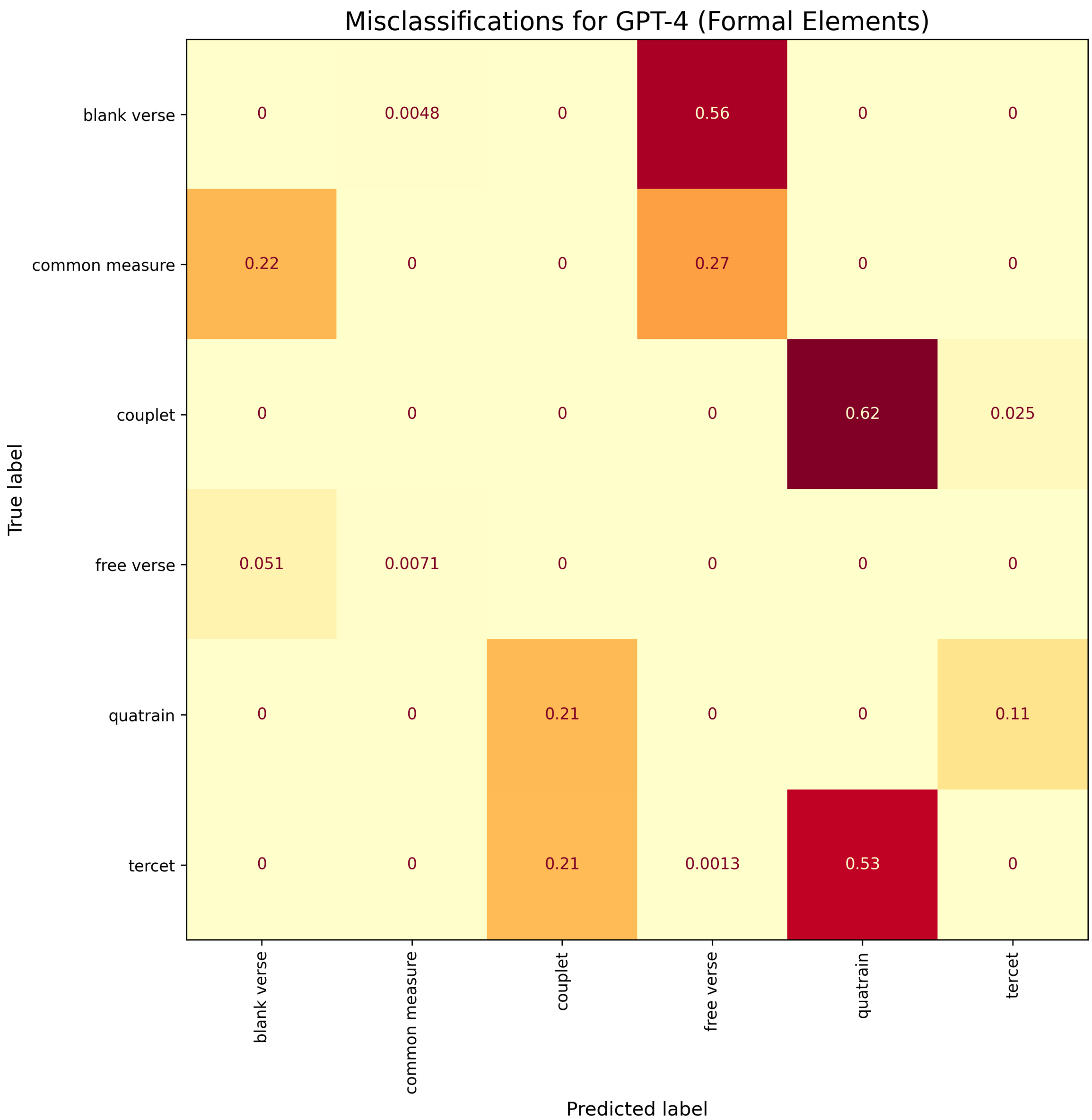

Figure 11: Misclassification proportions for formal elements by GPT-4.

A.10 Poetic Forms

Poetic forms can be defined and categorized in various ways. The definitions of forms and formal elements that we offer here are synthesized from information in glossaries of poetic terms available on the Poetry Foundation and Academy of American Poets websites as well as from widely used poetry resources by Strand and Boland (2000), Greene and Cushman (2016), and Preminger et al. (2015).

A.10.1 Fixed Forms

Ballad ↩ A type of narrative poem with ties to music and oral performance. Traditional ballads often feature regular meter and stanzas. One conventional pattern is "common measure," which consists of quatrains that rhyme ABCB and alternate iambic tetramater and trimeter.

Example Ballad: from "Barbara Allen" (by Anonymous)

In Scarlet town, where I was born,
  There was a fair maid dwellin',
Made every youth cry Well-a-way!
  Her name was Barbara Allen.

All in the merry month of May,
  When green buds they were swellin',
Young Jemmy Grove on his death-bed lay,
  For love of Barbara Allen.

He sent his man in to her then,
  To the town where she was dwellin';
"O haste and come to my master dear,
  If your name be Barbara Allen...

Ghazal ↩ Originally an Arabic verse form, ghazals consists of a series of couplets usually all ending in the same word. Poets may include their name in the final couplet.

Example Ghazal: from "Where did the handsome beloved go?" (by Jalal Al-Din Rumi, translated by Brad Gooch and Maryam Mortaz)

Where did the handsome beloved go?
I wonder, where did that tall, shapely cypress tree go?

He spread his light among us like a candle.
Where did he go? So strange, where did he go without me?

All day long my heart trembles like a leaf.
All alone at midnight, where did that beloved go?

Go to the road, and ask any passing traveler —
That soul-stirring companion, where did he go?

Go to the garden, and ask the gardener —
That tall, shapely rose stem, where did he go?

Go to the rooftop, and ask the watchman —
That unique sultan, where did he go?

Haiku ↩ Originating in Japan, haiku are concise, non-narrative poems that often focus on imagery. In English, haiku often consist of three unrhymed lines with 5, 7, and 5 syllables respectively.

Example Haiku: "In Kyoto" (by Bashō, translated by Jane Hirshfield)

In Kyoto,
hearing the cuckoo,
I long for Kyoto.

Limerick ↩ A light, often comedic verse form consisting of five lines rhymed AABBA. In traditional limericks, lines 1, 2, and 5 are trimeter, while lines 3 and 4 are dimeter, and the dominant meter is anapestic.

Example Limerick: "A Young Lady of Lynn" (by Anonymous)

There was a young lady of Lynn,
Who was so uncommonly thin
That when she essayed
To drink lemonade
She slipped through the straw and fell in.

Pantoum ↩ A Malaysian verse form that was adapted into French and later English, which consists of a series of quatrains in which the second and fourth lines of each quatrain serve as the first and third lines of the next quatrain. Pantoums do not have a determined length.

Example Pantoum: from "Nocturne" (by Sadakichi Hartmann)

Upon the silent sea-swept land
The dreams of night fall soft and gray,
The waves fade on the jeweled sand
Like some lost hope of yesterday.

The dreams of night fall soft and gray
Upon the summer-colored seas,
Like some lost hope of yesterday,
The sea-mew's song is on the breeze.

Upon the summer-colored seas
Sails gleam and glimmer ghostly white,
The sea-mew's song is on the breeze
Lost in the monotone of night.

Sails gleam and glimmer ghostly white,
They come and slowly drift away,
Lost in the monotone of night,
Like visions of a summer-day.

They shift and slowly drift away
Like lovers' lays that wax and wane,
The visions of a summer-day
Whose dreams we ne'er will dream again.

Sestina ↩ A complex verse form consisting of six, unrhymed, six-line stanzas followed by a three-line envoi. Each sestet includes the same six endwords in shifting, but specific patterns (below), and all six endwords also appear in the envoi. Endword pattern:

1: ABCDEF
2: FAEBDC
3: CFDABE
4: ECBFAD
5; DEACFB
6: BDFECA
envoi : ECA or ACE

Example Sestina: from "Sestina: Altaforte" (by Ezra Pound)

I
Damn it all! all this our South stinks peace.
You whoreson dog, Papiols, come! Let's to music!
I have no life save when the swords clash.
But ah! when I see the standards gold, vair, purple, opposing
And the broad fields beneath them turn crimson,
Then howl I my heart nigh mad with rejoicing.

II
In hot summer have I great rejoicing
When the tempests kill the earth's foul peace,
And the lightnings from black heav'n flash crimson,
And the fierce thunders roar me their music
And the winds shriek through the clouds mad, opposing,
And through all the riven skies God's swords clash.

III
Hell grant soon we hear again the swords clash!
And the shrill neighs of destriers in battle rejoicing,
Spiked breast to spiked breast opposing!
Better one hour's stour than a year's peace
With fat boards, bawds, wine and frail music!
Bah! there's no wine like the blood's crimson!

IV
And I love to see the sun rise blood-crimson.
And I watch his spears through the dark clash
And it fills all my heart with rejoicing
And pries wide my mouth with fast music
When I see him so scorn and defy peace,

His lone might ’gainst all darkness opposing.

V
The man who fears war and squats opposing
My words for stour, hath no blood of crimson
But is fit only to rot in womanish peace
Far from where worth’s won and the swords clash
For the death of such sluts I go rejoicing;
Yea, I fill all the air with my music.

VI
Papiols, Papiols, to the music!
There’s no sound like to swords swords opposing,
No cry like the battle’s rejoicing
When our elbows and swords drip the crimson
And our charges ’gainst “The Leopard’s” rush clash.
May God damn for ever all who cry “Peace!”

VII
And let the music of the swords make them crimson!
Hell grant soon we hear again the swords clash!
Hell blot black for always the thought “Peace!”

Sonnet ↩ A fourteen-line verse form, usually in iambic pentameter, and usually following a set rhyme scheme. The most common types of sonnets are Shakespearean/English, which consist of three quatrains followed by a couplet and often rhyme ABABCDCDEFEFGG, and Petrarchan/Italian, which consists of an octave followed by a sestet and often rhyme ABBAABBACDCDCD or ABBAABBACDECDE.

Example Petrarchan sonnet: “When I consider how my light is spent” (John Milton)

When I consider how my light is spent,
Ere half my days, in this dark world and wide,
And that one Talent which is death to hide
Lodged with me useless, though my Soul more bent
To serve therewith my Maker, and present
My true account, lest he returning chide;
“Doth God exact day-labour, light denied?”
I fondly ask. But patience, to prevent
That murmur, soon replies, “God doth not need
Either man’s work or his own gifts; who best
Bear his mild yoke, they serve him best. His state
Is Kingly. Thousands at his bidding speed
And post o’er Land and Ocean without rest:
They also serve who only stand and wait.”

Example Shakespearean Sonnet: “America” (Claude McKay)

Although she feeds me bread of bitterness,
And sinks into my throat her tiger’s tooth,
Stealing my breath of life, I will confess
I love this cultured hell that tests my youth.

Her vigor flows like tides into my blood,
Giving me strength erect against her hate,
Her bigness sweeps my being like a flood.
Yet, as a rebel fronts a king in state,
I stand within her walls with not a shred
Of terror, malice, not a word of jeer.
Darkly I gaze into the days ahead,
And see her might and granite wonders there,
Beneath the touch of Time's unerring hand,
Like priceless treasures sinking in the sand.

Villanelle ↩ A 19-line verse form originating in France, made up of five tercets followed by a quatrain, in which the first and third line of the first stanza are alternatingly repeated as a refrain in the following stanzas. Stanza 1 line 1 repeats as the third line of stanzas 2 and 4, and stanza 1 line 3 repeats as the third line of stanzas 3 and 5. These two lines also appear as the closing lines of the quatrain.

Example Villanelle: "Do not go gentle into that good night" (Dylan Thomas)

Do not go gentle into that good night,
Old age should burn and rave at close of day;
Rage, rage against the dying of the light.

Though wise men at their end know dark is right,
Because their words had forked no lightning they
Do not go gentle into that good night.

Good men, the last wave by, crying how bright
Their frail deeds might have danced in a green bay,
Rage, rage against the dying of the light.

Wild men who caught and sang the sun in flight,
And learn, too late, they grieved it on its way,
Do not go gentle into that good night.

Grave men, near death, who see with blinding sight
Blind eyes could blaze like meteors and be gay,
Rage, rage against the dying of the light.

And you, my father, there on the sad height,
Curse, bless, me now with your fierce tears, I pray.
Do not go gentle into that good night.
Rage, rage against the dying of the light.

A.10.2 Stanza Forms

Couplet ↩ A two-line stanza or two lines of verse, often but not always rhymed.

Example Couplets: "Interview" by Dorothy Parker

The ladies men admire, I've heard,
Would shudder at a wicked word.

Their candle gives a single light;
They'd rather stay at home at night.
They do not keep awake till three,
Nor read erotic poetry.
They never sanction the impure,
Nor recognize an overture.
They shrink from powders and from paints ...
So far, I've had no complaints.

Tercet ↩ A three-line stanza or three lines of verse, often but not always containing a rhyme.

Example Tercets: from "The Convergence of the Twain" (Thomas Hardy)

(Lines on the loss of the "Titanic")

I
In a solitude of the sea
Deep from human vanity,
And the Pride of Life that planned her, stilly couches she.

II
Steel chambers, late the pyres
Of her salamandrine fires,
Cold currents thrid, and turn to rhythmic tidal lyres.

III
Over the mirrors meant
To glass the opulent
The sea-worm crawls — grotesque, slimed, dumb, indifferent.

IV
Jewels in joy designed
To ravish the sensuous mind
Lie lightless, all their sparkles bleared and black and blind.

V
Dim moon-eyed fishes near
Gaze at the gilded gear
And query: "What does this vaingloriousness down here?" ...

Quatrain ↩ A four-line stanza or unit of verse, often, but not always containing rhyme.

Example Quatrains: from "Elegy Written in a Country Churchyard" (Thomas Gray)

The curfew tolls the knell of parting day,
The lowing herd wind slowly o'er the lea,
The plowman homeward plods his weary way,
And leaves the world to darkness and to me.

Now fades the glimm’ring landscape on the sight,
And all the air a solemn stillness holds,
Save where the beetle wheels his droning flight,
And drowsy tinklings lull the distant folds; ...

A.10.3 Meters

Free Verse ↩ Verse that does not follow a particular pattern of meter or rhyme.

Example Free Verse: from “The Waste Land” (T.S. Eliot)

April is the cruellest month, breeding
Lilacs out of the dead land, mixing
Memory and desire, stirring
Dull roots with spring rain.
Winter kept us warm, covering
Earth in forgetful snow, feeding
A little life with dried tubers.
Summer surprised us, coming over the Starnbergersee
With a shower of rain; we stopped in the colonnade,
And went on in sunlight, into the Hofgarten,
And drank coffee, and talked for an hour.
Bin gar keine Russin, stamm’ aus Litauen, echt deutsch.
And when we were children, staying at the archduke’s,
My cousin’s, he took me out on a sled,
And I was frightened. He said, Marie,
Marie, hold on tight. And down we went.
In the mountains, there you feel free.
I read, much of the night, and go south in the winter. ...

Blank Verse ↩ Unrhymed iambic pentameter.

Example Blank Verse: from *Paradise Lost* (John Milton)

Of Mans First Disobedience, and the Fruit
Of that Forbidden Tree, whose mortal tast
Brought Death into the World, and all our woe,
With loss of Eden, till one greater Man
Restore us, and regain the blissful Seat,
Sing Heav’nly Muse, that on the secret top
Of Oreb, or of Sinai, didst inspire
That Shepherd, who first taught the chosen Seed,
In the Beginning how the Heav’ns and Earth
Rose out of Chaos: or if Sion Hill
Delight thee more, and Siloa’s brook that flow’d
Fast by the Oracle of God; I thence
Invoke thy aid to my adventrous Song,
That with no middle flight intends to soar
Above th’ Aonian Mount, while it pursues
Things unattempted yet in Prose or Rhime.

Common Measure ↩ Quatrains consisting of alternating lines of iambic tetrameter and trimeter, rhymed ABAB.

Example Common Measure: from "It was not death for I stood up" (Emily Dickinson)

It was not Death, for I stood up,
And all the Dead, lie down -
It was not Night, for all the Bells
Put out their Tongues, for Noon.

It was not Frost, for on my Flesh
I felt Siroccos - crawl -
Nor Fire - for just my marble feet
Could keep a Chancel, cool -

And yet, it tasted, like them all,
The Figures I have seen
Set orderly, for Burial
Reminded me, of mine - ...

A.10.4 Unfixed forms

Ode ↩ A formal lyric poem, which addresses or celebrates a person, place, object, or concept, usually that is not present. Odes are often longer verse forms, and their stanza patterns vary.

Example Ode: from "Ode on a Grecian Urn" (John Keats)

Thou still unravish'd bride of quietness,
Thou foster-child of silence and slow time,
Sylvan historian, who canst thus express
A flowery tale more sweetly than our rhyme:
What leaf-fring'd legend haunts about thy shape
Of deities or mortals, or of both,
In Tempe or the dales of Arcady?
What men or gods are these? What maidens loth?
What mad pursuit? What struggle to escape?
What pipes and timbrels? What wild ecstasy?

Heard melodies are sweet, but those unheard
Are sweeter; therefore, ye soft pipes, play on;
Not to the sensual ear, but, more endear'd,
Pipe to the spirit ditties of no tone:
Fair youth, beneath the trees, thou canst not leave
Thy song, nor ever can those trees be bare;
Bold Lover, never, never canst thou kiss,
Though winning near the goal yet, do not grieve;
She cannot fade, though thou hast not thy bliss,
For ever wilt thou love, and she be fair!

Pastoral ↩ A type of poetry and a broader creative tradition idealizing rural life.

Example Pastoral: "The Passionate Shepherd to His Love" (Christopher Marlowe)

Come live with me and be my love,
And we will all the pleasures prove,
That Valleys, groves, hills, and fields,
Woods, or steepy mountain yields.

And we will sit upon the Rocks,
Seeing the Shepherds feed their flocks,
By shallow Rivers to whose falls
Melodious birds sing Madrigals.

And I will make thee beds of Roses
And a thousand fragrant posies,
A cap of flowers, and a kirtle
Embroidered all with leaves of Myrtle;

A gown made of the finest wool
Which from our pretty Lambs we pull;
Fair lined slippers for the cold,
With buckles of the purest gold;

A belt of straw and Ivy buds,
With Coral clasps and Amber studs:
And if these pleasures may thee move,
Come live with me, and be my love.

The Shepherds' Swains shall dance and sing
For thy delight each May-morning:
If these delights thy mind may move,
Then live with me, and be my love.

Aubade ↩ A poem or song welcoming or lamenting the arrival of dawn, usually with romantic themes.

Example Aubade: "Break of Day" (John Donne)

'Tis true, 'tis day, what though it be?
O wilt thou therefore rise from me?
Why should we rise because 'tis light?
Did we lie down because 'twas night?
Love, which in spite of darkness brought us hither,
Should in despite of light keep us together.

Light hath no tongue, but is all eye;
If it could speak as well as spy,
This were the worst that it could say,
That being well I fain would stay,

And that I loved my heart and honour so,
That I would not from him, that had them, go.

Must business thee from hence remove?
Oh, that's the worst disease of love,
The poor, the foul, the false, love can
Admit, but not the busied man.
He which hath business, and makes love, doth do
Such wrong, as when a married man doth woo.

Dramatic Monologue ↩ A poem in which a usually fictional speaker addresses a listener, who is also often imagined.

Example Dramatic Monologue: from "My Last Duchess" (Robert Browning)

That's my last Duchess painted on the wall,
Looking as if she were alive. I call
That piece a wonder, now; Fra Pandolf's hands
Worked busily a day, and there she stands.
Will't please you sit and look at her? I said
"Fra Pandolf" by design, for never read
Strangers like you that pictured countenance,
The depth and passion of its earnest glance,
But to myself they turned (since none puts by
The curtain I have drawn for you, but I)
And seemed as they would ask me, if they durst,
How such a glance came there; so, not the first
Are you to turn and ask thus. Sir, 'twas not
Her husband's presence only, called that spot
Of joy into the Duchess' cheek; perhaps
Fra Pandolf chanced to say, "Her mantle laps ...

Elegy ↩ A form of poetry and broader mode of writing expressing grief or loss, often in relation to its subject's death.

Example Elegy: from "Lycidas" (John Milton)

Yet once more, O ye laurels, and once more
Ye myrtles brown, with ivy never sere,
I come to pluck your berries harsh and crude,
And with forc'd fingers rude
Shatter your leaves before the mellowing year.
Bitter constraint and sad occasion dear
Compels me to disturb your season due;
For Lycidas is dead, dead ere his prime,
Young Lycidas, and hath not left his peer.
Who would not sing for Lycidas? he knew
Himself to sing, and build the lofty rhyme.
He must not float upon his wat'ry bier
Unwept, and welter to the parching wind,

Without the meed of some melodious tear.

Concrete Poetry ↩ A type of poetry that is structured by visual effect on the page, and often emphasizes nonlinguistic aspects of writing, including typography, layout, whitespace, etc.

Example Concrete Poetry: "Easter Wings" (George Herbert)

Lord, who createdst man in wealth and store,
Though foolishly he lost the same,
Decaying more and more,
Till he became
Most poore:
With thee
O let me rise
As larks, harmoniously,
And sing this day thy victories:
Then shall the fall further the flight in me.

My tender age in sorrow did beginne
And still with sicknesses and shame.
Thou didst so punish sinne,
That I became
Most thinne.
With thee
Let me combine,
And feel thy victorie:
For, if I imp my wing on thine,
Affliction shall advance the flight in me.

Prose Poem ↩ A poetic composition that is not broken up into lines.

Example Prose Poem: *Gitanjali*, #14 (by Rabindranath Tagore)

My desires are many and my cry is pitiful, but ever didst thou save me by hard refusals; and this strong mercy has been wrought into my life through and through.

Day by day thou art making me worthy of the simple, great gifts that thou gavest to me unasked—this sky and the light, this body and the life and the mind—saving me from perils of overmuch desire.

There are times when I languidly linger and times when I awaken and hurry in search of my goal; but cruelly thou hidest thyself from before me.

Day by day thou art making me worthy of thy full acceptance by refusing me ever and anon, saving me from perils of weak, uncertain desire.

Ars Poetica ↩ A poem about poetry.

Example Ars Poetica: from "Poetry" (Marianne Moore)

I too, dislike it: there are things that are important beyond all this fiddle.
Reading it, however, with a perfect contempt for it, one discovers that there is in
it after all, a place for the genuine.
Hands that can grasp, eyes
that can dilate, hair that can rise
if it must, these things are important not because a

high-sounding interpretation can be put upon them but because they are
useful; when they become so derivative as to become unintelligible, the
same thing may be said for all of us—that we
do not admire what
we cannot understand. The bat,
holding on upside down or in quest of something to

eat, elephants pushing, a wild horse taking a roll, a tireless wolf under
a tree, the immovable critic twinkling his skin like a horse that feels a flea, the base—
ball fan, the statistician—case after case
could be cited did
one wish it; nor is it valid
to discriminate against "business documents and

school-books"; all these phenomena are important. One must make a distinction
however: when dragged into prominence by half poets, the result is not poetry,
nor till the autocrats among us can be
"literalists of
the imagination"—above
insolence and triviality and can present

for inspection, imaginary gardens with real toads in them, shall we have
it. In the meantime, if you demand on the one hand, in defiance of their opinion—
the raw material of poetry in
all its rawness, and
that which is on the other hand,
genuine, then you are interested in poetry.

Ekphrasis ↩ Writing that uses vivid language to respond to or describe a work of visual art.

Example Ekphrasis: "On Seeing the Elgin Marbles" (John Keats)

My spirit is too weak—mortality
Weighs heavily on me like unwilling sleep,
And each imagined pinnacle and steep
Of godlike hardship tells me I must die
Like a sick eagle looking at the sky.
Yet 'tis a gentle luxury to weep,
That I have not the cloudy winds to keep,
Fresh for the opening of the morning's eye.
Such dim-conceived glories of the brain
Bring round the heart an indescribable feud;
So do these wonders a most dizzy pain,
That mingles Grecian grandeur with the rude

Wasting of old Time—with a billowy main—
A sun—a shadow of a magnitude.