Evaluating the Multilingual Rhyming capabilities of Open-Source Large Language Models

Aadit Deshpande
aaditd@andrew.cmu.edu

Abstract

Phonology is the systematic study of sounds and their interactions with other groups of sounds. With recent advances in Large Language Models (LLMs), there is growing interest in testing their phonological capabilities, as LLMs are used in many tasks that require joint phonological and orthographic understanding of language. This involves creative writing tasks like poetry generation, lyric and melody composition, and even machine translation for certain language families. In particular, we focus on the rhyming capabilities of LLMs on two Germanic languages - English and Dutch. We quantify 'rhyming' into five subcategories - Single Perfect rhymes, Double Perfect rhymes, Assonance, Alliteration, and Consonance, and present a dataset for both languages corresponding to these categories. We test three open-source LLMs - Llama2-7b, Llama3-8B-Instruct, and CrystalChat-7b. We also investigate the effects of different lengths of prompts (title vs. description-level) as single-prompt evaluations tend to be brittle. The models all lag behind human performance, but both our prompt evaluations reveal that Llama3 performs the best overall (with a 0.73 F-1 score for single perfect rhymes) and shows a sophisticated understanding of phonological concepts in its explicit reasoning. Finally, we also propose a typology of errors about prediction and implicit/explicit reasoning mismatch.

1 Introduction

Phonology is the field of linguistics concerned with the study of how groups of sounds systematically behave and their interactions with other groups of sounds. Humans who have a strong understanding of these phonological rules (implicitly through experience and intentionally through practice) are capable of performing complex writing tasks such as writing poetry, song lyrics and translate between languages (if they speak multiple languages). Recent advances in large language models (LLMs) (Touvron et al. (2023), OpenAI (2023), Taori et al. (2023)) have produced models that can generate very fluent text and apparently seem to have a good understanding of the textual components of linguistics such as morphology and syntax. These impressive text generations beg the question - whether LLMs that have been pre-trained solely on textual data with no explicit phonological representation can also perform the aforementioned writing tasks that require joint orthographic and phonological understanding.

LLMs have been benchmarked for a variety of text understanding tasks like news summarization (Zhang et al. (2024)), pragmatic tasks (Sravanthi et al. (2024)) , and user-centric factual QA (Wang et al. (2024)). However, far fewer studies have examined the phonological capabilities of LLMs. Phonological understanding can unlock a dearth of new tasks that LLMs can gain proficiency in - such as lyric and melody composition (Ding et al. (2024)), language learning, and improved poetry generation. In particular, machine generated poetry is highly criticized as it often fails to employ poetic devices and simply relies on surface level orthographic rhymes or memorized poems which occurred in its pre-training data. (D'Souza and Mimno (2023)). Rhyming is one aspect of phonological understanding that has been shown to be important as a teaching tool for young language learners (Bruck and Genesee (1995)) and thus could benefit LLMs improve their linguistic capabilities. In particular, a comprehensive understanding of rhyming is crucial because it tests key phonological skills such as differentiating between graphemes and phonemes, recognizing stress patterns, and longer range dependencies (as in rhyme schemes).

Previous studies that have examined the phonological understanding of LLMs such as Suvarna et al. (2024) focus on multiple tasks such as Grapheme-to-Phoneme (G2P), syllable counting

	English Dataset	Dutch Dataset
Single Perfect	goodwill-instill	distantieert-getierceerd
Double Perfect	consummation-avocation	slierend-mierend
Alliteration	circuit-saunders	doopheffer-domkapittel
Assonance	twilight-highrise	weert-sneep
Consonance	tokio-attica	fleuren-foulard

Table 1: Examples of rhyming word-pairs for English and Dutch corresponding to the five types of rhymes considered in this study.

and generating rhyming words, and thus are able to pay limited attention to the complexities of the rhyming subtask. Other studies explore the recognition and recall of LLMs by using rhyming and non-rhyming cue words to test whether the models would pass the Tulving test (Chauvet (2024a)). Moreover, almost all these studies only evaluate LLM performance on English. In this study, we expand upon the premise of previous phonological benchmarks of LLMs by focusing on testing the rhyming capabilities of three open-source LLMs. In essence, the research question we ask is - does pre-training purely on text data allow LLMs to acquire phonological capabilities with respect to different kinds of rhymes. Our main contributions are [1]:

- We present rhyme datasets for English and Dutch, each with five lists of one thousand rhyming words corresponding to five types of rhymes - Single Perfect, Double Perfect, Assonance, Consonance, and Alliteration.

- We examine three open-source models in the 7B parameter range - Llama2-7b-chat-hf (Touvron et al. (2023)), Llama3-8B-Instruct (AI@Meta (2024)), and CrystalChat-7b (Liu et al. (2023)). Inspired by the field of information retrieval, we also examine the effects of inferencing the models using two different prompts - a short 'title' prompt and a more detailed 'description'-level prompts.

- Our findings suggest that LLMs still lag behind human performance on the rhyming word pair classification task. Llama3-8B-Instruct shows the best performance relative to the other models. The Single and Double Perfect rhyming tasks are ostensibly the most difficult for all three models and seem to benefit the most during the switch from title prompts to description level prompts.

[1]Code and dataset available at: https://github.com/Aadit3003/llm-rhyme

2 Related Studies

2.1 Phonological Understanding

Suvarna et al. (2024) present PhonologyBench, an English phonology benchmark consisting of three tasks - G2P, syllable counting, and rhyme word generation. They compare open-source and closed-source LLM performance to a human annotator baseline. They find that no one model outperforms others consistently and that LLMs significantly underperform humans on syllable counting and rhyme word generation. Thus the authors recommend selecting LLMs that perform well on the phonological task which closely relates to the downstream task. In another study, Suvarna et al. suggest that LLMs can perform well on phonological tasks purely by learning associations from orthography (i.e. they do not require deep phonological understanding). They study tasks like phonetic transcript generation, rhyming word classification/generation and also report that models struggle far more with rarer words than frequent words that might have appeared in their pre-training corpora.

2.2 Tulving Test

Chauvet (2024b) studies the properties of memory traces in LLMs compared to human memory test performance by employing the Tulving memory models: GAPS (General Abstract Processing System (Chauvet, 2024a)) and the SPI (Serial-Parallel Independent) model. The Tulving test involves presenting a word to the LLM to be remembered in the test and then repeatedly probing with successive cues to estimate the valences of these cues. The authors found that rhyming cue words seemed to be more effective in enhancing recall than unrelated distractor cues, but less effective than copied or associated cue words. Thus, prompting studies to

	Title Prompt	Description Prompt
Single Perfect + Double Perfect	Do these words rhyme form a perfect rhyme?	Do these words rhyme i.e. have different consonants followed by identical vowel and consonant sounds?
Alliteration	Do these words show assonance?	Do these words have identical vowel sounds but different consonant sounds?
Assonance	Do these words show consonance?	Do these words have identical consonant sounds but different vowel sounds?
Consonance	Do these words show alliteration?	Do these words begin with the same consonant sound?

Table 2: Title and Description-level prompts corresponding to the five types of rhymes considered in this study.

examine the rhyming capabilities of LLMs in more detail.

2.3 Prompting

Finally, Mizrahi et al. (2023) examines the limitations of existing LLM evalaution benchmarks that use single-prompt evaluations. The authors instead suggest the use of multiple instruction paraphrases specifically tailored for different downstream use cases to enable a fairer assessment. They examine instruction-tuned models in the 8B-13B parameter range on 39 tasks and find that single prompt evaluations are inconsistent, and instead use instruction template rephrasing, chain-of-thought prompting and gradual template generation. In this study, we borrow some ideas from the template rephrasing in the form of 'title' and 'description'-level prompting.

3 Dataset Creation

3.1 Characteristics of Germanic Languages

In the present study, we choose to focus on two Germanic languages, namely modern English and modern Dutch. Selecting two languages from the same language family allows us to refer to the differences in morphology and phonology of English and Dutch while comparing the downstream LLM performance on rhyming word classification. For instance, Dutch and English are both fusional languages, however, Dutch is far more synthetic than English, as it shows a much higher degree of compounding and polymorphemic words (Wikipedia contributors: Germanic Languages, 2024). An interesting point to note is the differences in poetic devices frequently employed in both languages. Old English (which also happened to be more inflectional than its modern counterpart) heavily relied on alliteration, whereas proto-Germanic poetry such as Skaldic poetry frequently featured assonance. It is also worth noting that the strong initial stress in Germanic language (particularly English) words triggered attrition of other syllables, reducing most words to monosyllables.

3.2 Five Types of Rhymes

We consider five kinds of rhymes in this study to provide a comprehensive view of the rhyming capabilities of LLMs. We rely on definitions from Greene and Cushman (2016) and operationalize the different rhymes between a pair of words as follows:-

- Single Perfect Rhymes: The stressed vowel sound and subsequent phonemes must be identical, but the onset of the stressed syllable must differ. The words must show final stress. For e.g. "billionaire-hare".
- Double Perfect Rhymes: The stressed vowel sound and subsequent phonemes must be identical, but the onset of the stressed syllable must differ. The words must show penultimate stress. For e.g. "suppressant-effervescent".
- Assonance: The vowel sounds and stresses must be identical in both words (and in the same order), and all consonants must differ (to prevent overlap with the perfect rhymes). For e.g. "fairlow-petko".
- Consonance: The consonant phonemes must be identical in both words (and in the same or-

Model	Title prompts				
	Single Perfect	Double Perfect	Assonance	Consonance	Alliteration
Llama-2-7b	38.87	66.52	**66.84**	66.80	66.76
Llama3-8b	**73.60**	**68.41**	65.24	**67.16**	**69.12**
CrystalChat-7b	55.45	60.60	63.32	64.17	63.40

Table 3: Results for the English dataset with title-level prompts. (F-1 score as percentage)

Model	Description prompts				
	Single Perfect	Double Perfect	Assonance	Consonance	Alliteration
Llama-2-7b	65.98	66.84	**66.71**	**66.76**	67.02
Llama3-8b	**67.90**	**68.24**	65.90	**66.62**	**67.50**
CrystalChat-7b	64.96	65.63	60.63	63.41	61.07

Table 4: Results for the English dataset with description-level prompts. (F-1 score as percentage)

der), and no vowel sounds (marked by stress) should overlap. For e.g. "avalons-villines".

- Alliteration: Both words must have initial stress with the same vowel and the same onset phoneme. For e.g. "molded-midair".

3.3 Dataset Mining

Both datasets (English and Dutch) were mined using pronunciation dictionaries for the respective languages. The English dataset was created using the CMU Pronunciation Dictionary which contains over 130K words and their pronunciations (North American English) in ARPAbet symbols with stresses indicated. We created five lists of 1000 word pairs (no overlap) corresponding to the five types of rhymes defined above. Additionally, a list of 5000 non-rhyming words was formed to randomly sample from, in order to create a balanced dataset for testing. We constructed the Dutch dataset in a similar fashion, using the CELEX-2 expanded lexical and phonological database for Dutch. Table 1 shows examples of entries from both datasets.

4 Method

4.1 Models

We examine three state-of-the-art open-source LLMs in the 7B parameter range - Llama2-7b-chat-hf (Touvron et al. (2023)), Llama3-8B-Instruct (AI@Meta (2024)), and CrystalChat-7b (Liu et al. (2023)). We focus on the popular Llama models by Meta, whose specific training data is largely unknown (including the recently released Llama3-8B model) and LLM360's CrystalChat which was trained on publicly available datasets as well as an HTML coding instruction dataset "WebAlpaca".

4.2 Task Definition

We frame the task as zero-shot binary classification with the goal of testing the model's ability to discriminate 1000 pairs of between rhyming and non-rhyming words for five kinds of rhymes. Not only does this task test the LLMs' knowledge of phonemes, but also their knowledge of the typology of rhymes. Initial tests with only rhyming words revealed misleadingly high scores, as the models would predict "yes" for the majority of pairs. Hence, we decided to randomly sample (fixed seed) 500 non-rhyming words from our dataset and 500 rhyming words for each type of rhyme.

4.3 Prompt Variation

Finally, we introduce variation in the prompt style for the models, to examine whether they already possess knowledge of the concept (e.g. what is a perfect rhyme/assonance) and if supplementing this with a definition (as in the 'description'-level prompts) affects their performance. The various prompts are provided in Table 2. We compare the models using their F1 scores on each type of rhyme and prompt type.

5 Results

In this section, we discuss the performance of the LLMs on the various types of rhyme word pair classificationt task. Table 3 and Table 4 show the F-1 scores of the models on the English dataset for title and description-level prompts respectively. Similarly, Table 5 and Table 6 show the F-1 scores

Model	Title prompts				
	Single Perfect	Double Perfect	Assonance	Consonance	Alliteration
Llama-2-7b	28.78	50.33	**45.19**	**52.37**	62.58
Llama3-8b	**65.00**	**66.90**	**67.08**	**59.56**	**67.61**
CrystalChat-7b	55.39	56.48	64.69	57.53	63.59

Table 5: Results for the Dutch dataset with title-level prompts. (F-1 score as percentage)

Model	Description prompts				
	Single Perfect	Double Perfect	Assonance	Consonance	Alliteration
Llama-2-7b	**66.26**	66.09	**55.65**	**60.15**	43.90
Llama3-8b	**65.57**	**66.76**	**66.40**	**59.43**	**66.58**
CrystalChat-7b	64.62	65.32	64.94	56.81	54.86

Table 6: Results for the Dutch dataset with description-level prompts. (F-1 score as percentage)

on the Dutch dataset for title and description-level prompts.

5.1 The effect of language

The LLMs naturally perform better on the English dataset than on the Dutch datasets, as they have likely seen a far greater amount of English rhyming data than Dutch data. However, the difference is not as stark as expected. One possible reason could be that the models can apply their knowledge of English phonology to Dutch as both are Germanic. Another explanation could be the prevalence of monosyllabic/monomorphemic words could make the classification task slightly easier. The single and double perfect tasks seem considerably more difficult for Dutch than for English. This could be attributed to the fact that Dutch is more synthetic than English and that multiple morphemes might trick/confuse the models. However, the models show almost identical performance on what we refer to as the 'poetic device rhymes' (alliteration, assonance and consonance) across both languages, indicating that the models show good recognition of vowel and consonant phonemes regardless of language.

5.2 The effect of different prompts on LLMs

The overall order of performance across both languages and all rhyme types seems to be Llama3 (best performing) followed by Llama2, which is closely followed by CrystalChat. Llama3 particularly shines in the Single and double perfect tasks where the other models lag behind by a large margin. However, all three models seem to perform identically on the poetic device rhymes (assonance, alliteration, consonance). All three models seem to cap out at an upper bound of about 70-75% F-1, seemingly much lower than human performance as reported by related studies.

The prompt variation seem to have an unexpected effect on the models. The description level prompts yield a dramatic improvement in performance for the seemingly more difficult "perfect rhyme tasks" (especially single perfect), whereas assonance, alliteration, and consonance are nearly unchanged (or slightly worsened). In particular, Llama2 benefits the most from this kind of longer prompt style. CrystalChat also benefits from description-level prompts but to a smaller degree. Surprisingly, Llama3-8B's performance is almost entirely unaffected by changing the prompt style. In fact, it reduces significantly for English single perfect rhymes, indicating that it might already possess an understanding of the concept of perfect rhymes and that the 'helpful description' may not contribute any new useful information to the model.

5.3 Qualitative Analysis of Models

We realize that while drawing conclusions, it's crucial to keep in mind that the models' generated 'explicit reasoning' might not match the internal logic it used to predict an answer (the 'internal reasoning'). Based on this fact and upon examining the reasoning generated by the models, we propose two major kinds of errors. They are:

- Type-A: The LLM's answer contradicts its explicit reasoning. This is easy to see in the LLM generations, as the model provides a certain reasoning, but its classification strays from this very reasoning
- Type-B: The LLM's answer contradicts its

implicit reasoning. This is much harder to see and the only feasible way to detect such an inconsistency is by examining the models' pre-training data.

In this section, we also briefly discuss the qualitative performance of the models to examine the kinds of errors they routinely make. Appendix A shows examples of these kinds of errors for all three LLMs. Here is a brief discussion of the general characteristics of the responses provided by the LLMs:

- Llama2-7b: The model's performance severely degrades when the language is switched from English to Dutch. For the Dutch single perfect rhymes, it overwhelmingly answers "no" leading to a very low F-1 score (0.287). This could be either due to a lack of Dutch training data that it could draw from or an inability to generalize English phonological concepts to Dutch. For the rest of the rhyme types, the model heavily relies on orthography and the ending sounds and rarely generates IPA representations of the words to reason. This leads to a lot of Type-A errors as mentioned above.

- Llama3-8B: The model no longer suffers from the drop in performance due to switching the language, suggesting that Llama3 saw far more multilingual data in its pre-training phase. It outperforms the other models by a significant margin on single and double perfect rhymes and its explicit reasoning shows a sophisticated understanding of concepts like perfect rhyme and stress patterns. It employs IPA notation far more than Llama2 and correspondingly, its performance is much stronger. Llama3 still suffers from the Type-A mistakes, but to a lesser extent. Finally, it gains almost no boost from the description-level prompts, indicating, that it may rely on its implicit reasoning far more than the information provided in the prompt.

- CrystalChat-7b: This model has a tendency to predict "yes" for the majority of the examples. The high number of false positives leads to the worst performance overall (only slightly behind Llama2). CrystalChat does not seem to possess understanding of concepts like assonance or perfect rhymes and thus benefits from the description prompts. It often goes on tangents and brings in unrelated concepts when it is unfamiliar with the concept. Naturally, it suffers from the Type-A mistakes, and hallucinates a lot. For example, in the case of alliteration, the starting sounds for $Word_1$ are forgotten immediately after observing $Word_2$. Further analysis on its training dataset - WebAlpaca is required to make conclusive remarks about the model's ability.

6 Conclusion

In this study, we investigate the phonological capabilities of LLMs by testing whether they are able to discern rhyming and non-rhyming word-pairs in two Germanic languages (Dutch and English). To this end, we construct datasets for both languages and operationalize five kinds of rhymes to capture the phonological understanding at different positions within words. We test three open-source LLMs: Llama2-7b, Llama3-8B-Instruct and CrystalChat-7b on the English and Dutch datasets. Based on the F-1 scores, the overall performance order seems to be Llama3-8B > Llama2-7b > CrystalChat-7b. Performance is highly task dependent, as the perfect rhymes seem to be difficult for both languages, whereas assonance, consonance, and alliteration are nearly identical. We also investigate the effect of different levels of detail in the prompt and find that the description level prompts benefit models like Llama2 or CrystalChat, but Llama3 is virtually unaffected. Finally we discuss two kinds of mismatch errors between the models' predicted answers and their implicit or explicit reasoning used to reach those answers. This is an exciting avenue for future studies, where models such as OLMO-7b could be investigated, with a deeper analysis on its pre-training dataset (which is unavailable for the Llama models). Additionally, language families such as Arabic could be studied which feature far more challenging orthography than the Latin script of English and Dutch.

References

AI@Meta. 2024. Llama 3 model card.

Maggie Bruck and Fred Genesee. 1995. Phonological awareness in young second language learners. *Journal of child Language*, 22(2):307–324.

Jean-Marie Chauvet. 2024a. Memory gaps: Would llm pass the tulving test? *arXiv preprint arXiv:2402.16505*.

Jean-Marie Chauvet. 2024b. Memory traces: Are transformers tulving machines? *arXiv preprint arXiv:2404.08543*.

Shuangrui Ding, Zihan Liu, Xiaoyi Dong, Pan Zhang, Rui Qian, Conghui He, Dahua Lin, and Jiaqi Wang. 2024. Songcomposer: A large language model for lyric and melody composition in song generation. *arXiv preprint arXiv:2402.17645*.

Lyra D'Souza and David Mimno. 2023. The chatbot and the canon: Poetry memorization in llms. *Proceedings http://ceur-ws. org ISSN*, 1613:0073.

Roland Greene and Stephen Cushman. 2016. *The Princeton handbook of poetic terms*. Princeton University Press.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. 2023. Llm360: Towards fully transparent open-source llms. *Preprint*, arXiv:2312.06550.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of what art? a call for multi-prompt llm evaluation. *arXiv preprint arXiv:2401.00595*.

R OpenAI. 2023. Gpt-4 technical report. *ArXiv*, 2303.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *arXiv preprint arXiv:2401.07078*.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. Should large language models go to school for phonology?

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. Phonologybench: Evaluating phonological skills of large language models. *arXiv preprint arXiv:2404.02456*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. A user-centric benchmark for evaluating large language models. *arXiv preprint arXiv:2404.13940*.

Wikipedia contributors: Germanic Languages. 2024. Germanic languages. [Online; accessed 03-May-2024].

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

A Appendix

Qualtiative examples of errors made by the three models on the different datasets.

Rhyme Type	Rhyme Word Pair	LLM Reasoning	Notes
Single Perfect (EN)	**stare, boisclair**, Gold: 1, Pred: 0	No, the words "stare" and "Boisclair" do not form a perfect rhyme. A perfect rhyme is when two words have the same ending sound. The word "stare" ends in /re/, while "Boisclair" ends in /air/.</s>	Relies on orthography
Single Perfect (NL)	**deposant, fondant**, Gold: 1, Pred: 0	No, the words "deposant" and "fondant" do not form a perfect rhyme in Dutch. In Dutch, two words are considered to be perfect rhymes if they have the same ending sound and vowel sound(s). For example, "rood" (red) and "bloed" (blood) are perfect rhymes because both endings are pronounced with the exact same sounds	Brings in irrelevant example (Relies on definition), relies on gloss
Single Perfect (NL + Desc)	**toppenant, fondant**, Gold: 1, Pred: 1	Yes, the two words "toppensant" and "fondant" do rhyme in Dutch because they both end with the same sequence of consonant (n) followed by an identically pronounced vowel sound ("a").</s>	Uses Description prompt to improve!

Table 7: Example predictions and corresponding generations from Llama2-7b on some rhyme word pairs. EN and NL refer to the English and Dutch Datasets respectively.

<table>
<tr><th>Rhyme Type</th><th>Rhyme Word Pair</th><th>LLM Reasoning</th><th>Notes</th></tr>
<tr><td>Single Perfect (EN)</td><td>bared, repaired, Gold: 1, Pred: 1</td><td>Yes.The justification is that the ending sounds of "bared" (rhyming with "aired", "fired", etc.) match exactly to those of "repared" in their vowel quality, consonant sound, and syllable stress pattern ("-ed"). This meets the criteria for a perfect rhyme according to phonological rules of English language rhymes.<|eot_id|></td><td>Explicit reasoning shows richer understanding of Rhyme concepts!</td></tr>
<tr><td>Single Perfect (NL)</td><td>luck, jam, Gold: 0, Pred: 1</td><td>Yes.The justification is that the ending sounds of "luck" (uck) and "jam" (am) match exactly, which meets the criteria for rhyming in English language poetry or music lyrics. Perfect rhymes require identical final syllable sounds, and both "luck" and "jam" meet this requirement. Therefore, I conclude that they do indeed form a perfect rhyme.<|eot_id|></td><td>Type-A mistake! Long explanation but mismatch!</td></tr>
<tr><td>Single Perfect (NL + Desc)</td><td>planimetreer, navel, Gold: 0, Pred: 1</td><td>Yes.The justification is as follows:* "planimetree" (the correct spelling) has the initial sound /pl/, which consists of two similar stops, both voiced (/p/) followed by a voiceless lateral approximant (/l/).* The word "navel", on the other hand, starts with the same sequence: /n-av-el/. Although they share some similarities in their spellings, such as having'm' or 'v', respectively <|eot_id|></td><td>Type-A mistake!</td></tr>
</table>

Table 8: Example predictions and corresponding generations from Llama3-8B on some rhyme word pairs. EN and NL refer to the English and Dutch Datasets respectively.

Rhyme Type	Rhyme Word Pair	LLM Reasoning	Notes
Single Perfect (EN)	**gettin', locatelli**, Gold: 0, Pred: 1	Yes, the words "gettin'-locatelli" form a near-perfect rhyme (also known as a near-doublereative or near-antipodal rhyme). The sound at the end of the first word, /n/, is very close to the /ltl/ sound of the second word, making it a near-rhyme. Although the two words don't exactly rhyme, their sound is very similar, indicating that the syllables in the two words share the same underlying vowel, but with a small change in the vowel's length. This is considered to be a near-sound change which makes the two words be considered near-rhymes.</s>	Explicit reasoning shows richer understanding of Rhyme concepts!
Single Perfect (NL)	**reveal, neel**, Gold: 1, Pred: 1	Yes. The words "reveal" and "neel" rhyme because they have the same consonant sounds (both start with "n") and the same vowel sound (both end with "ail").</s>	Boost with Description level prompting! Wrong reasoning
Single Perfect (NL + Desc)	**loodwitfabriek, levensverandering**, Gold: 1, Pred: 1	Yes.The justification is that "lood" and "leven" both start with the /l/ sound, which creates an alliterative effect in the phrase "loowitfabriek-levensverandering". Alliteration occurs when two or more consecutive sounds within a word repeat identical initial consonant(s) followed by different vowel sounds (or no vowels). In this case, we have repeated "/l/" sounds creating an instance of alliteration.</s>	Defaults to predicting positive, but works well even with compound/longer words!

Table 9: Example predictions and corresponding generations from CrystalChat-7b on some rhyme word pairs. EN and NL refer to the English and Dutch Datasets respectively.