

# Evaluating the Rhyming capabilities of Large Language Models

Aadit Deshpande

April 23 2024

# Motivation and Research Questions

- LLMs are used in many downstream tasks that require both textual and phonological understanding
- Tasks
  - Poetry Generation (Highly criticized!!)
  - Song Writing
  - Machine Translation
- **Does pre-training only on text data allow acquisition of phonological capabilities?**
- More specifically - Rhyming
  - Critical to creative writing tasks
  - Early age language learner studies
  - Poetic Devices - Alliteration, Assonance, Consonance

# What is Rhyming exactly?

- **Single Perfect**

- Identical after Stressed vowel (Final). Different Onset. **Billionaire-Hare**
- billionaire: [B, IH2, L, Y, AH0, N, EH1, R] hare: [HH, EH1, R]

- **Double Perfect**

- Identical after Stressed vowel (Penultimate). Different Onset. **Suppressant-Effervescent**
- Suppressant: [S, AH0, P, R, EH1, S, AH0, N, T] effervescent: [EH2, F, ER0, V, EH1, S, AH0, N, T]}

- **Assonance**

- Identical Vowels. **Fairlow-Petko**
- fairlow: [F, EH1, R, L, OW0] petko: [P, EH1, T, K, OW0]

- **Consonance**

- Identical Consonants **Avalons-Villines**
- avalons: [AE1, V, AH0, L, AA2, N, Z] villines: [V, IH1, L, AY2, N, Z]

- **Alliteration**

- Initial Stress, Same Onset **Molded-Midair**
- molded: [M, OW1, L, D, AH0, D] midair: [M, IH1, D, EH1, R]}



# Germanic Language Characteristics

- Grimm's Law - Intro of Fricatives to proto-German
- Verner's law - Strong initial Stress - attrition of vowels in unstressed syllables
- Vowels -> merger of long, short /a/ and /o/ -> Affects Diphthongs (many)
- Modern English has lost almost all of the complex inflectional morphology
- Dutch and English are both fusional
- Dutch more synthetic than English! (Highly compounded + Several morphemes)
- Old English was More inflectional!
- Modern Sound Poetic Devices - Alliteration, Assonance, Consonance
- Old English - heavily relied on Alliteration! (best performing)
- Skaldic poetry - Poetry forms - Lyric / Sonnet Assonance!! (best performing)

# Experimental Setup

- Three open-source Models
  - **Llama2-7b**-chat-hf
  - **Llama3-8b**-Instruct
  - **CrystalChat-7b**
- Languages
  - ENGLISH
  - DUTCH
- Prompt Types - *Title / Description*
- Five Rhyme Types
- Task Definition - Zero-shot binary classification - classify 1000 pairs of randomly drawn (balanced) rhyming, non-rhyming words (F-1 score)

# Prompt Variation

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

# Dataset Creation

## English Dataset

- Five lists of 1000 rhyme pairs, a list of 5000 non-rhyming words
- Mined from CMU Pronouncing Dictionary (130K)
- word1 | rep1 | word2 | rep2

## Dutch Dataset

- Five lists of 1000 rhyme pairs, a list of 5000 non-rhyming words
- Mined from Celex2 dpw set (Dutch Phonology Word Forms)
- Created CMU-Dutch-Dict (340K) [22V, 23C]
- *geneert*: [Z, @0, n, e1, r, t] *geciseleerd*: [x, @0, s, i0, z, @0, l, e1, r, t]

# Data Examples

		
Single Perfect	goodwill instill seared appeared	distantieert getierceerd
Double Perfect	consummation avocation vary contrary	slierend mierend
Alliteration	lag litters circuit saunders'	doopheffer domkapittel
Assonance	habitats caravans twilight highrise	weert sneep
Consonance	tokio attica lovitz elevates	fleuren foulard



# Results: English

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

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

# Results: Dutch

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

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

# Discussion 1: Language

- English ofc does better (very well represented in the training data!), but Dutch is not too far behind!
  - Both are Germanic Languages!
  - Could be because binary classification is easier (especially with description prompts!)
  - Several Monosyllabic / Monomorphemic Words!
- Dutch Single Perfect and Double Perfect seem particularly difficult!
- Alliteration and Consonance seem language independent (Latin script!)
- Vowel Length + Unique Dutch Diphthongs could make assonance trickier!
- Modern Dutch is more Synthetic than Modern English!

# Discussion 2: Which model came out on top?

## Models

- Overall:- **Llama3** > Llama2 > CrystalChat
- Single/Double Perfect - By a big margin
- Exceptions
  - Assonance + Consonance - Rarer concepts (Nearly equal performance)
- Mostly 0 / Mostly 1 performance (Unpredictable, rhyme type dependent!)
- Overall performance upper bounded by ~70% (Human performance is likely far higher)

## Effects of prompting

- Most improved - Llama2! (Title -> Description)
  - Especially for Dutch
- Crystal and Llama3 slightly worse! (with description prompts)
- Task-specific
  - Single/Perfect really benefit
  - Assonance/Consonance - Same/slightly worsened!



# Discussion 3: Two Kinds of Errors

## Types of Errors

- **Disclaimer:** *Model's explicit reasoning may not match its internal reasoning!*
- Type-A: Answer contradicts explicit reasoning [easy to see]
- Type-B: Answer contradicts implicit reasoning [need to examine data]
- Task Difficulty
  - Single/Double Perfect most difficult! Biggest gains from longer prompts
  - Assonance, Consonance seem easier!
  - Alliteration - Naturally easiest!

## Models

- Llama2 - Poor application of explicit conceptual understanding of rhymes
  - Many Type-A errors! (especially Dutch single perfect - mostly predicted 0!)
  - Relies on glosses ??
- Llama3 - Bridges answer-reasoning gap!
- CrystalChat - Explicit reasoning mentions a lot of Phonological properties



# Qualitative analysis of Model behaviour

Not to anthropomorphize, but let's anthropomorphize!

# Llama2-7b-chat

- *Disobedient - Unearned confidence!*
- **Performance really degrades when we switch over to Dutch!**
- General Characteristics of responses
  - Very negative for Single Perfect
  - Overwhelmingly “Yes” for Alliteration/Assonance/Consonance
- Reasoning Used
  - Orthography
  - Ending Sounds
  - With Description - Sometimes brings in IPA
- Qualitative Examples

# Llama2 Examples

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/.	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



# Llama3-8b-Instruct

- *Student who studied for the exam and will let you know it!*
- **No longer suffers from the drop in performance due to language change!**
- Does Far better on Single/Double Perfect
- Same performance on Assonance/Consonance/Alliteration
- General Characteristics of responses
- Reasoning Used
  - **More sophisticated understanding of perfect rhymes / concepts**
  - Uses IPA notation and phonological features in explicit reasoning!
- Lesser mistakes, but still mismatch b/w judgement and reasoning
- No gains from description (strong pre-training)

# Llama3 Examples

Single Perfect (EN)	<b>bared, repaired</b> , Gold: 1, Pred: 1    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 >	<b>Explicit reasoning shows richer understanding of Rhyme concepts!</b>
Single Perfect (EN)	<b>luck, jam</b> , Gold: 0, Pred: 1    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 >	<b>Type-A mistake! Long explanation but mismatch!</b>
Consonance (NL)	<b>planimetreer, navel</b> , Gold: 0, Pred: 1    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, after	<b>Type-A mistake!</b>

# CrystalChat-7b

- Highly Positive, and goes on tangents
- WebAlpaca (Code) Dataset - Instruction tuning
- Overall performs worst, middle of the pack on Single perfect
- General Characteristics of responses
  - Perfect Gets a big boost from Description prompts
  - Assn/Cons/Allit seem to suffer a bit!
- Reasoning Used
  - **Very often just defaults to saying “yes”**
  - Uses unrelated concepts if unfamiliar
  - Hallucinates a lot (starting sounds for word1 are forgotten after seeing word2!)
- Many False positives

# Crystal Mistakes

Single Perfect (EN + Title)	<b>gettin', locatelli</b> , 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, /ɛn/, is very close to the /lɑtələ/ 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>	<b>Concept Mismatch (Didn't ask for near perfect rhyme)</b>
Single Perfect (EN + Desc)	<b>reveal, neel</b> , 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>	<b>Boost with Description level prompting! Wrong reasoning</b>
Alliteration (NL)	<b>loodwitfabriek, levensverandering</b> , 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 "loodwitfabriek-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.< eot_id >	Defaults to predicting positive, but works well even with compound/longer words!

# Summary

## Takeaways

- Llama3 relatively best model, still lags behinds humans.
- Performance on Dutch is not too far behind that on English.
- Effect of prompts is pronounced for models struggling on Perfect rhyming!
- Task difficulty:- *Alliteration* < *Assn* ~ *Consn* < *Double* << *Single*
- Explicit reasoning often does not match implicit reasoning!

## Future Directions

- Entropy of Suffixes
- Effects of Word rarity in the dataset
- More analysis of pre-training data

