

# Can ChatGPT Be Useful for Distant Reading of Music Similarity?

Arthur Flexer<sup>1</sup>

<sup>1</sup>Johannes Kepler University Linz, Austria

## Abstract

We explore whether large language models (ChatGPT) can be used as a ‘distant reading’ tool to estimate music similarity between songs from textual information, complementing experiments with human participants. We compare degrees of rater agreement to previous results from a listening test, showing that correlation of ChatGPT with human raters is significantly lower than the average human inter-rater agreement, but nevertheless still of moderate positive size. We discuss whether an approach based on the largely opaque ChatGPT model can be scientifically valid and to what extent it allows transparent evaluation of music information research experiments.

## Keywords

music similarity, large language models, transparency, evaluation

## 1. Introduction

The accessibility of vast amounts of text in digital form has enabled humanities to add ‘distant reading’ of thousands of books via computational analysis as a new research tool to its repertoire of methods [1, 2]. Distant reading is usually applied to large collections of text often of a magnitude which cannot be handled by individual scholars in what is known as traditional ‘close reading’, i.e. very careful and detailed expert reading of only comparably few texts. Large language models (LLM) [3, 4, 5] are also trained on vast amounts of text, with recent extensions like ChatGPT (<https://openai.com/blog/chatgpt>) providing conversational interfaces which can be used to query and probe the knowledge contained in these texts. In this paper we explore whether ChatGPT can be used to ‘distant read’ the similarity between songs and discuss the results in comparison to previous results obtained via human listening tests [6]. The textual information could provide complementary information like cultural connotations, or other forms of so-called music context [7], which are not present in music audio alone.

Large language models (LLM) are deep neural models that obtain representations of text by learning to predict the next word given a textual context. Most successful approaches are based on the so-called transformer architecture [3, 4], implementing an attention mechanism which learns to reweight parts of the textual input in relation to its importance for the task under consideration. ChatGPT, which we have used for our experiments in this paper, is a

---

*HCMIR23: 2nd Workshop on Human-Centric Music Information Research, November 10th, 2023, Milan, Italy*


✉ [arthur.flexer@jku.at](mailto:arthur.flexer@jku.at) (A. Flexer)

🌐 <https://www.jku.at/en/institute-of-computational-perception/about-us/people/arthur-flexer-dr/> (A. Flexer)

🆔 0000-0002-1691-737X (A. Flexer)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

chatbot imitating a human conversational partner and is also based on a ‘Generative Pre-trained Transformer’ (GPT). More specifically it is based on GPT-3.5, which itself is a fine-tuned version of GPT-3 [4]. A problem common to all members of the GPT family (including ChatGPT) is that the exact models, training sets, parameters, etc are not known. A non peer reviewed report [5] by the developing team about the latest version (GPT-4) even states that "[...] no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar" can be given due to "safety implications" and the "competitive landscape" of LLM research.

Most related to our work is an experiment using ChatGPT to rate musical instrument sounds on a set of 20 semantic scales [8]. ChatGPT’s results are only partially correlated with human ratings, most pronounced for clearly defined dimensions of musical sounds such as brightness (bright–dark) and pitch height (deep–high) with Pearson correlations above 0.80. This is related to work on distilling psychophysical information from text by aligning GPT-4 results with human auditory experience [9]. Other applications of LLMs to music include lyrics summarization [10] and usage as ranking models for music recommendation [11]. Applying ChatGPT to music data is also reminiscent of earlier approaches to compute music similarity using textual sources, most notably web-based data [12], but also semantic music tags [13] or lyrics [14].

## 2. Methods

### 2.1. Human evaluation of music similarity

We will compare our ChatGPT results to a previous study that conducted a series of listening tests with human participants [6]. In this study the age of participants ranged from 26 to 34 years with an average of 28.2. The sample consisted of three females and three males, which we call graders S1 to S6 from here on. The  $5 \times 18$  songs belonged to five genres (for a full list see section A of the appendix of the original article): (i) **American Soul** from the 1960s and 1970s with only male singers singing; (ii) **Bebop**, the main jazz style of the 1940s and 1950s, with excerpts containing trumpet, saxophone and piano parts; (iii) **High Energy** (Hi-NRG) dance music from the 1980s, typically with continuous eighth note bass lines, aggressive synthesizer sounds and staccato rhythms; (iv) **Power Pop**, a Rock style from the 1970s and 1980s, with chosen songs being guitar-heavy and with male singers; (v) **Rocksteady**, which is a precursor of Reggae with a somewhat soulful basis. It is worth noting that one aim of the original study was to find less known songs by making sure that every song had under 50.000 accesses on Spotify. Genres were validated via respective Wikipedia artist pages as well as by listening to all songs. For presentation in the listening test, 15 seconds of a representative part of every song (usually the refrain) were chosen and participants were asked to “assess the similarity between the query song and each of the five candidate songs by adjusting the slider” (ranging from 0 to 100 %) and “to answer intuitively since there are no wrong answers”. Based on randomly chosen 15 query songs, comparisons of five pairs had to be made for every query group yielding a total of  $15 \times 5 = 75$  pairs, with every song appearing exactly once in the whole questionnaire (15 as query songs, 75 as candidate songs).

## 2.2. ChatGPT evaluation of music similarity

For our experiments conducted on the 5th and 6th of April 2023 we used the "Free Research Preview" of the ChatGPT Mar 23 Version. The service came with a warning that "ChatGPT may produce inaccurate information about people, places, or facts" and the information that "ChatGPT is fine-tuned from GPT-3.5, a language model trained to produce text. ChatGPT was optimized for dialogue by using Reinforcement Learning with Human Feedback (RLHF) – a method that uses human demonstrations and preference comparisons to guide the model toward desired behavior".

For the exact same  $15 \times 5 = 75$  song pairs as used in the human listening test we asked ChatGPT the following question: "On a scale of 1 to 100, how similar is the song [s\_i] by [artist\_A] to the song [s\_j] by [artist\_B]?". It is worth noting that ChatGPT sometimes needed persuasion to provide an answer at all, stating e.g. that "As an AI language model, I do not have the ability to directly listen to music or interpret subjective qualities such as similarity between songs", or that any answer would be "merely speculation". However, the following additional input sentences provided by us in ensuing dialogues always resulted in ChatGPT answering with a similarity score: "Please just make a guess based on the information you have already", "Please try anyway", "Then please just speculate". This kind of persuasion was necessary for 8 out of 75 questions, mostly at the beginning of ChatGPT sessions. Due to restriction of the free ChatGPT version experiments had to be split over three separate sessions.

## 3. Results

In accordance with the previous study [6], we analyse the degree of inter-rater agreement by computing the Pearson correlations<sup>1</sup>  $\rho_{listen}$  between graders S1 to S6 as well as  $\rho_{gpt}$  between graders S1 to S6 and ChatGPT for the 75 pairs of query/candidate songs. Please note that the human listening test had been conducted twice at time points t1 and t2 with a two week time lag [6]. The 15 plus 15 correlations  $\rho_{listen}$  (t1 and t2) between the six graders range from 0.59 to 0.86, with an average of 0.74 (see Table 1 for individual correlations). The 6 plus 6 correlations  $\rho_{gpt}$  (t1 and t2) between the six graders and ChatGPT are considerably lower, with a range from 0.39 to 0.72 and an average of 0.58. The differences in correlation between  $\rho_{listen}$  and  $\rho_{gpt}$  are statistically significant ( $t(40)=6.05$ ,  $p=0.00$ ).

When ChatGPT answers the similarity questions it always also provides some form of explanation. Sometimes they are very brief: "Based on the available information, I would speculate [...]". Often genre, instrumentation or era of recording are being debated: "[...] given that both artists were active in the same time period and were part of the Jamaican music scene, it is possible that there may be some similarities in terms of instrumentation, rhythm, or vocal style" or "They are from different musical genres, different eras, and have different rhythms, melodies, instrumentation, and lyrics". Some explanations are quite detailed: "Both songs are characterized

---

<sup>1</sup>One of the advantages of Pearson correlation is that it implicitly normalizes for different rating styles because it is invariant under separate changes in location and scale in the two correlated variables. For example a Pearson correlation is perfect if two raters give identical answers, but also if all answers of one of the raters are always shifted by e.g. 10 units. Other measures of rater agreement like Fleiss' Kappa are only defined for the categorical scale, but applicable alternatives for the interval scale like Krippendorff's alpha exist.

	time point t1								time point t2							
	S1	S2	S3	S4	S5	S6	GPT	S1	S2	S3	S4	S5	S6	GPT		
S1	1.00	0.77	0.74	0.72	0.74	0.82	0.61	1.00	0.79	0.73	0.77	0.74	0.83	0.60		
S2		1.00	0.72	0.75	0.62	0.83	0.53		1.00	0.73	0.74	0.75	0.86	0.57		
S3			1.00	0.70	0.67	0.76	0.49			1.00	0.69	0.69	0.80	0.59		
S4				1.00	0.64	0.80	0.64				1.00	0.59	0.79	0.72		
S5					1.00	0.64	0.45					1.00	0.73	0.39		
S6						1.00	0.72						1.00	0.64		

**Table 1**

Inter-rater correlation between graders S1 to S6 at time points t1 and t2, with mean  $0.74 \pm .064$  standard deviation. Inter-rater correlation between graders S1 to S6 at times t1 and t2 and ChatGPT, with mean  $0.58 \pm .096$  standard deviation.

by smooth, soulful vocal performances and feature catchy melodies and memorable hooks. Additionally, both songs deal with themes of love and relationships, which further underscores their similarities". ChatGPT has been criticized for sometimes ‘hallucinating’ [5] facts that sound plausible but are actually incorrect. We verified that ChatGPT’s argumentation seems to be correct basically all the time by searching and reading respective online sources (e.g. Wikipedia or Discogs), or, in case these cannot be found, by listening to the audio.

## 4. Discussion and Conclusion

Although the mean Pearson correlation of ChatGPT with human raters is significantly lower (0.58) than the average inter-rater agreement (0.74), it still remains at a relatively high level. Even the ranges of correlations  $\rho_{listen}$  and  $\rho_{gpt}$  are overlapping, hence some of the raters agree to a higher degree with ChatGPT than with some of the other human raters.

Many of the explanations provided by ChatGPT are about music genre or instrumentation, which can be seen as an indirect indication of genre. Interestingly, previous self-reports by human raters [6] have already indicated that genre is an important aspect when rating similarity of songs. When comparing ChatGPT similarity scores within genres to scores obtained when query and candidate songs are of different genres, on average within genre scores are always higher than between genre scores. This effect is most pronounced for ‘Soul’ and least for ‘Power Pop’. It would be interesting to explore whether correlations between ChatGPT and human raters would decrease if all song material were part of the same genre. For human raters, agreement for such a single single genre study indeed was lower [6] since genre of songs seems a major factor in judging similarity between songs.

Although it seems that music similarity, as measured via human listening tests, can to a certain degree be recovered from textual data by using ChatGPT as a distant reading tool, the black box nature of LLMs and especially ChatGPT remains a problem. Since the exact training data and modeling approach are unknown (see sec. 1), it remains unclear what enabled ChatGPT to provide answers that at least partially are in alignment with human feedback to listening tests. One very likely possibility is that respective webpages about artists and songs have

been part of ChatGPT's training data, allowing ChatGPT to reproduce this content when being queried accordingly. Another possibility is that ChatGPT is actually able to reason about musical concepts like genre, instrumentation, harmony, etc. How ChatGPT arrives at a similarity score for a pair of songs remains completely mysterious. Because the full nature of ChatGPT has not been disclosed we are left to speculate about these matters. The future will show whether such opaque tools will nevertheless become part of science's repertoire or if they will be confined to practical applications where measurable success is sufficient and scientific rigour not necessary. This could mean that music similarity measured via ChatGPT could e.g. be useful for music recommendation systems but not for research on human semantic music similarity concepts.

As a final comment it will also be interesting to see how open source alternatives to ChatGPT like LLaMA [15], Alpaca [16] or Open-Assistant (<https://github.com/LAION-AI/Open-Assistant>) will change assessment of the usefulness of large language models for distant reading.

## Acknowledgments

This research was funded in whole, or in part, by the Austrian Science Fund (FWF) [P 36653]. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## References

- [1] F. Moretti, Conjectures on world literature, *New left review* 2 (2000) 54–68.
- [2] F. Moretti, *Graphs, maps, trees: abstract models for a literary history*, Verso, 2005.
- [3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [5] OpenAI, Gpt-4 technical report, 2023. [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- [6] A. Flexer, T. Lallai, K. Rašl, On evaluation of inter- and intra-rater agreement in music recommendation, *Transactions of the International Society for Music Information Retrieval* 4(1) (2021) 182–194.
- [7] M. Schedl, A. Flexer, J. Urbano, The neglected user in music information retrieval research, *Journal of Intelligent Information Systems* 41 (2013) 523–539.
- [8] K. Siedenbueg, C. Saitis, The language of sounds unheard: Exploring musical timbre semantics of large language models, 2023. [arXiv:2304.07830](https://arxiv.org/abs/2304.07830).
- [9] R. Marjeh, I. Sucholutsky, P. van Rijn, N. Jacoby, T. L. Griffiths, Large language models predict human sensory judgments across six modalities, 2023. [arXiv:2302.01308](https://arxiv.org/abs/2302.01308).
- [10] Y. Zhang, J. Jiang, G. Xia, S. Dixon, Interpreting song lyrics with an audio-informed pre-trained language model, in: *Proceedings of the 23rd International Society for Music Information Retrieval Conference, 2022*, pp. 19–26.
- [11] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, W. X. Zhao, Large language models are zero-shot rankers for recommender systems, 2023. [arXiv:2305.08845](https://arxiv.org/abs/2305.08845).

- [12] P. Knees, E. Pampalk, G. Widmer, Artist classification with web-based data, in: Proceedings of the 5th International Conference on Music Information Retrieval, 2004.
- [13] D. Turnbull, L. Barrington, D. Torres, G. Lanckriet, Towards musical query-by-semantic-description using the cal500 data set, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 439–446.
- [14] B. Logan, A. Kositsky, P. Moreno, Semantic analysis of song lyrics, in: IEEE International Conference on Multimedia and Expo (ICME), volume 2, IEEE, 2004, pp. 827–830.
- [15] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. [arXiv:2302.13971](https://arxiv.org/abs/2302.13971).
- [16] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.