ON THE VALIDITY OF EMPLOYING CHATGPT FOR DISTANT READING OF MUSIC SIMILARITY

Arthur Flexer
Institute of Computational Perception
Johannes Kepler University Linz, Austria
arthur.flexer@jku.at

ABSTRACT

In this work we explore whether large language models (LLM) can be a useful and valid tool for music knowledge discovery. LLMs offer an interface to enormous quantities of text and hence can be seen as a new tool for 'distant reading', i.e. the computational analysis of text including sources about music. More specifically we investigated whether ratings of music similarity, as measured via human listening tests, can be recovered from textual data by using ChatGPT. We examined the inferences that can be drawn from these experiments through the formal lens of validity. We showed that correlation of ChatGPT with human raters is of moderate positive size but also lower than the average human inter-rater agreement. By evaluating a number of threats to validity and conducting additional experiments with ChatGPT, we were able to show that especially construct validity of such an approach is seriously compromised. The opaque black box nature of ChatGPT makes it close to impossible to judge the experiment's construct validity, i.e. the relationship between what is meant to be inferred from the experiment, which are estimates of music similarity, and what is actually being measured. As a consequence the use of LLMs for music knowledge discovery cannot be recommended.

1. INTRODUCTION

When developing and validating hypotheses in musicology, relevant information very often is obtained from written documents. This information from collections, anthologies, compilations, biographies, reviews, journals, etc is today often available in digitized formats, enabling usage of methods from natural language processing (NLP) for music knowledge discovery [1]. In the humanities such an approach is also known as 'distant reading' [2, 3], i.e. the computational analysis of large quantities of books and texts 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) [4–6] offer a convenient interface to enormous quantities of text and hence can be seen as a new tool for distant reading. In our previous work [7] we have evaluated the use of LLMs for distant reading of music similarity. Our results showed 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. However, it also already became clear that the black box nature of LLMs, and especially of ChatGPT, presents a problem for the validity of such an approach.

In this article we therefore critically appraise our own previous work by utilizing an established framework of validity by Shadish et al. [8]. Validity is the truth of an inference made from evidence gathered through an experiment and as such an integral pillar of working scientifically. We will question our approach to music knowledge discovery concerning its statistical conclusion, internal, construct and external validity. All four types of validity have recently been discussed by applying Shadish et al. [8] to the context of music information research [9], which we will use as a guideline in this article. Reformulating our work in the general framework of validity will allow us to draw conclusions going beyond our particular music similarity setting to the general problem of using LLMs for music knowledge discovery.

We present related work in section 2 and explain the experimental setting (including preceding work we build on) in section 3. In sections 4 to 7 we critically appraise a primary study on human perception of music similarity [10], our own previous work with ChatGPT [7], as well as a number of additional ChatGPT experiments conducted for this work, all concerning four types of validity. We discuss our main findings and conclude in section 8.

2. RELATED WORK

Large language models (LLM) are deep neural models that learn representations of text by trying to predict the next word given a textual context. State of the art approaches are based on transformer architectures [4, 5], implementing an attention mechanism which learns to reweight parts of the textual input in relation to its importance for the task under consideration. ChatGPT (https://openai.com/blog/chatgpt) is a chatbot imitating a human conversational partner and is also based on a 'Generative Pre-trained Transformer' (GPT). We have used GPT-3.5, which itself is a fine-tuned version of GPT-3 [5], for our experiments in this paper and in our previous work [7]. A problem common to all members of the GPT family (including ChatGPT) is that exact details of models, training sets, parameters, etc are not known. A non peer reviewed report [6] 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.

ChatGPT has already been used in a music context rating instrument sounds on a set of 20 semantic scales [11]. It was found that ChatGPT's answers are only partially correlated with human ratings, with Pearson correlations above 0.80 only achieved for clearly defined dimensions of musical sounds such as brightness (bright–dark) and pitch height (deep–high). This is closely related to another approach trying to extract psychophysical information from text by aligning GPT-4 results with human auditory experience [12]. Further applications of LLMs to music include lyrics summarization [13] and usage as ranking models for music recommendation [14]. Applying LLMs to music data is also reminiscent of preceding approaches computing music similarity from textual sources, including web-based data [15], semantic music tags [16] or lyrics [17]. There also exists related work in the text domain, e.g. on using LLMs for evaluation of jokes [18, 19].

Previous work on distant reading in music information research (MIR) includes automatic band member detection and automatic recognition of all their released records from internet text sources [20], sentiment analysis of a large corpus of Pop music reviews [1], discovery of social and professional networks from Wikipedia articles on Renaissance musicians [21], extraction of semantic information from an online discussion forum on Carnatic music [22], or very detailed cross-linking of references to musical passages in musicological texts [23].

3. EXPERIMENTAL SETUP

In our previous research [7] we explored whether ChatGPT can be used to 'distant read' the similarity between songs and compared the results to a study employing human listening tests on the same pairs of songs [10]. ChatGPT therefore has to recover music similarity, as judged by humans listening to audio, solely from textual data. Textual sources could also provide complementary information like cultural connotations, or other forms of so-called music context [24]. Such information is of course not present in music audio alone, but the mere knowledge of such contextual facts may nevertheless influence human listeners in their judgement. Our major hypothesis therefore was:

> "Music similarity estimated with ChatGPT correlates positively with human perception of music similarity"

In order to being able to properly discuss the validity of conclusions drawn from the respective experiment, we must identify its components. *Treatments* are the things applied to *units* in order to cause an effect. In our case the participants and ChatGPT are the treatments, while the set of questions (pairs of songs to be evaluated) are the units. The effect we want to cause is to gain an estimate of music similarity. *Observations* are what is measured on a unit, in our case the music similarity ratings ranging from 0 to 100.

3.1 Human evaluation of music similarity

The primary study conducted a series of listening tests with human participants [10], with the age of participants ranging from 26 to 34 years with an average of 28.2 (three females and three males, called graders S1 to S6 from here on). The 5×18 songs belonged to five genres (for a full list see section A of the appendix of the original article (`https://doi.org/10.5334/tismir.107.s1`)): (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. All songs had limited popularity with under 50.000 accesses on Spotify at the time of the study. The authors validated genres via respective Wikipedia artist pages as well as by listening to all songs. Fifteen seconds of a representative part of every song (usually the refrain) were presented in the listening tests 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).

3.2 ChatGPT evaluation of music similarity

For our initial experiments [7] conducted on the 5th and 6th of April 2023 we used the "Free Research Preview" of the ChatGPT Mar 23 Version (`https://openai.com/blog/chatgpt`). 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".

We asked ChatGPT the following question for the exact same $15 \times 5 = 75$ song pairs as used in the human listening test:

> "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]?"

Interestingly, 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". The following additional input sentences (in that order) provided by us in ensuing dialogues always resulted in ChatGPT providing a similarity score:

1. "Please just make a guess based on the information you have already"
2. "Please try anyway"
3. "Then please just speculate"

Such additional persuasion was necessary for 8 out of 75 questions, mostly at the beginning of ChatGPT sessions. Experiments had to be split over three separate sessions due to restriction of the free ChatGPT version.

4. STATISTICAL CONCLUSION VALIDITY

Statistical conclusion validity is "the validity of inferences about covariation between two variables" [8]. Here the main concern is with *statistical significance*, i.e., that an observed covariation between treatment and effect is not likely to arise by chance.

In accordance with the initial study [10], we recorded the music similarity ratings and then, to gain an estimate of the level of agreement between human participants and ChatGPT, we analysed degrees of inter-rater agreement. Specifically, we computed the Pearson correlations ρ_{listen} between graders S1 to S6 as well as ρ_{gpt} between graders S1 to S6 and ChatGPT for the 75 pairs of query/candidate songs (see table 1 for an overview of all results). The human listening test had been conducted twice at time points t1 and t2 with a two week time lag [10]. The 15 plus 15 correlations ρ_{listen} (t1 and t2) between the six graders range from 0.59 to 0.86, with an average of 0.74. The 6 plus 6 correlations ρ_{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 correlation ρ_{gpt} is statistically significant, i.e. the probability that we observe such a positive correlation by chance is basically zero (t(898)=$|-17.83|$, p=0.00). Hence, a valid statistical conclusion is that we observe a significant covariation between the human participants and ChatGPT in the observed estimates of music similarity. This result therefore seems to corroborate our hypothesis that music similarity estimated with ChatGPT correlates positively with human perception

	five genres		one genre
agreement	inter	intra	inter
ρ_{listen}	0.74	0.80	0.24
ρ_{gpt}	0.58	0.68	0.06

Table 1. Overview of results for five and one genre experiments. Shown are levels of inter- and intra-rater agreement between human participants (ρ_{listen}) and between human participants and ChatGPT (ρ_{gpt}).

of music similarity. In addition, the differences in correlation between ρ_{listen} and ρ_{gpt} are also statistically significant (t(40)=6.05, p=0.00).

5. INTERNAL VALIDITY

Internal validity is “the validity of inferences about whether the observed covariation between two variables is causal” [8]. It is therefore focused on the *cause* of a particular response to the treatment, going beyond statements concerning only the strength of covariation. A typical threat to internal validity is *confounding*, which is the confusion of the treatment with other factors, often arising from poor operationalisation in an experiment.

For our specific experiment we are interested in estimates of music similarity, either via listening tests with humans or from text sources via ChatGPT. One explanation for the observed level of rater agreement ρ_{gpt} is that indeed human perception of audio music similarity is positively correlated with ChatGPT estimates of music similarity. This is certainly one explanation consistent with our observations, but is it the only one? The internal validity of this conclusion relies on the key assumption that the observed positive correlation can *only* be explained in terms of music similarity and that there is no other way to arrive at the observations.

However, already in the initial study [10], participants commented that the genre of the songs was an important factor when evaluating the similarity of songs. When both query and candidate songs belonged to the same genre, similarity ratings were on average higher (within genres: 43.09) when compared to song pairs from different genres (between genres: 30.10). For our initial ChatGPT experiments [7] we have observed something related in the explanations provided by ChatGPT together with the similarity ratings. We provide some exemplary ChatGPT explanations with different levels of detail in table 2. As is typical for the answers ChatGPT provided, genre, instrumentation or era of recording are being discussed. A standard definition [25] of music genre states that it is "a set of musical events ... governed by a definite set of socially accepted rules", with musical events being "any type of activity performed around any type of event involving sound". Sound events are of course also linked to the concept of music similarity, making it clear that music genre and similarity are related but not synonymous concepts.

We therefore repeat the analysis of similarity ratings re-

`"These two songs are from very different genres and have distinct musical styles."`
`"[...] I can attempt to speculate based on the artist's genre and the era of the music"`
`"They are from different musical genres, different eras, and have different rhythms, melodies, instrumentation, and lyrics."`
`"Both songs share some similarities in terms of their musical genres, but they are likely to have different arrangements, melodies, and lyrics."`
`"While both songs are in the broad category of popular music, they come from different genres (soul/R&B for Major Harris and reggae for The Heptones) and have different rhythms, melodies, instrumentation, and lyrics."`
`"[...] 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"`

Table 2. Typical explanations provided by ChatGPT.

garding genres of query/candidate songs and get the following results: within genres: 43.13, and between genres: 13.42. Just as for human ratings, ChatGPT ratings seem to rely at least in part on genre information and not music similarity alone. This then calls into question how the design of our experiment relates to what we actually want to measure, which is music similarity. This is where construct validity becomes relevant.

6. CONSTRUCT VALIDITY

Construct validity is "the validity of inferences about the higher order constructs that represent sampling particulars" [8]. This concerns the operationalisation of the experimentalist's intention, i.e. the relationship between what is meant to be inferred from an experiment and what is actually measured. In our case the higher order construct is music similarity.

An important part of the operationalisation of our experiment is the exact form of questions the participants ("assess the similarity between the query song") and ChatGPT ("how similar is the song") are being asked. Both questions clearly aim at the similarity between songs but do not specify what exact aspect of similarity is meant. Many possibilities come to mind, e.g. similar in terms of melody, tempo, instrumentation, time of publishing, or maybe genre? Indeed, as has already been explained above, human participants commented that the genre of the songs was an important factor when evaluating the similarity of songs. Many of the explanations provided by ChatGPT were also about music genre or instrumentation, with the latter being an indirect indication of genre. A decisive difference is however that we of course trust in the honesty of human participants when answering post-experiment questions concerning their strategies, while with ChatGPT such trust seems unwarranted. ChatGPT has been criticized for sometimes 'hallucinating' [6] 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 by listening to the audio. Nevertheless the black box nature of LLMs and especially ChatGPT is a problem for judging construct validity. Since the exact training data and modeling approach are unknown [6], we have no way to judge whether ChatGPT really used genre clues for providing music similarity scores. One possibility is that respective webpages about artists and songs, often including genre information, have been part of ChatGPT's training data, allowing ChatGPT to reproduce this content when being queried accordingly. Indeed recent results indicate that LLMs seem to memorize large parts of their training data [26].

One way to test the hypothesis that ChatGPT uses genre information when judging music similarity is to repeat the experiment with music from a single genre. The initial study [10] repeated the listening tests with 90 songs all belonging to the genre **Power Pop** (for a full list see section B of the appendix of the original article (`https://doi.org/10.5334/tismir.107.s1`)) with 28 participants of an average age of 25.6. The average inter-rater agreement between human participants ρ_{listen} dropped from 0.74 to 0.24 when the song material was restricted to a single genre (see table 1). We now repeat the ChatGPT experiments with the restriction to **Power Pop** songs only. The average inter-rater agreement between human participants and ChatGPT ρ_{gpt} drops from 0.58 to 0.06 (see table 1). It seems that without the possibility to resort to genre information, ChatGPT has severe problems to rate music similarity. In the explanations provided by ChatGPT, it is often correctly stated that both songs "were part of the power pop genre during the same era", but sometimes also subgenres are being named when justifying certain scores, e.g. "... leans towards a pop-rock sound ... while ... tends to blend progressive and art-rock elements" or "... was associated with power pop and new wave music, while ... was known for its indie rock and power pop sound". Nevertheless the scores provided by ChatGPT remain very restricted, essentially consisting of three values (30, 40, 50) around the middle of the possible range.

From these results it seems evident that the poor operationalisation of the experiment, essentially not asking a clear enough question, has led to a lack of construct validity: we were aiming for music similarity as a higher order construct but music genre seems to also have been a relevant aspect for both human participants and ChatGPT when answering questions during the experiments. For the human participants this problem became already evident during post-experiment questioning and was then only corroborated with the restricted single genre experiment. The lack of trust due to ChatGPT's black box nature however made the same experiment inevitable to clarify construct validity of the ChatGPT experiment.

Another way to question construct validity is to assess the outcomes of different experiments which are supposed to measure the same higher order constructs. We could for instance study correlations of results from different LLMs being queried with identical prompts. Low correlations between LLM outputs could point to problems of construct validity. This kind of testing already points to the concept of external validity.

7. EXTERNAL VALIDITY

External validity is "the validity of inferences about the extent to which a causal relationship holds over variations in experimental units, settings, treatment variables and measurement variables" [8]. Therefore, external validity is the truth of a generalised causal inference made from an experiment. It is clear that if a causal inference we draw from an experiment already lacks internal validity, then generalising that inference to variations not even tested will not have external validity. In addition, a major threat is that variation of components of an experiment might dismantle the causal inference that holds in the experiment.

One component that could be varied are the annotators, i.e. the human participants or the type of LLM employed. Already in the initial experiment [10] it became clear that human annotators only agree to a certain extent in their evaluation of music similarity (average ρ_{listen} of 0.74). This is because human perception of music is highly subjective with personal taste, listening history, familiarity with the music, current mood, etc, being important influencing factors [24, 27]. Such a lack of inter-rater agreement presents a problem of external validity because inferences from the experiment do not generalize from users or annotators in the experiment to the intended target population of arbitrary users/annotators. It would be interesting to test the level of agreement between ChatGPT-3.5, which has been used for the experiments in this paper, and newer versions like ChatGPT-4 [6] or even alternative LLMs like Google's Gemini (`https://gemini.google.com/`), LLaMA [28] or Alpaca [29]. There already is evidence that ChatGPT's responses differ between different versions [30]. In case we want the conclusions drawn from our experiment to have external validity beyond one specific type of LLM, such additional experiments would of course be necessary.

One could even ask the question what the level of agreement within one person is when faced with identical annotation tasks at different points in time. Results from the initial experiment already showed that such an intra-rater agreement, tested two weeks apart, is only slightly higher than inter-rater agreement [10] at 0.80 versus 0.74. We therefore also repeated our ChatGPT five genre experiment on January 5, 2024, nine month after the first experiment. Although the LLM used was supposedly still a ChatGPT-3.5 version, the intra-rater agreement was only at 0.68. This is actually not much higher than the inter-rater agreement between human participants and ChatGPT ρ_{gpt} at 0.58. It therefore seems that there is a lack of external validity when generalizing ChatGPT results to different points in time.

Another problem of external validity is the influence of prompt engineering on LLM results. It is known that slight variations in prompt formulation can lead to quite different results, which brought about a whole new 'science' of so-called 'prompt engineering' [31]. One example is 'chain-of-thought prompting', where a few chain of thought demonstrations provided to an LLM as prompts lead to improved results [32]. 'Positive thinking' prompts like "You are an expert mathematician" also improve LLM performance and automatic prompt optimization sometimes produces quite bizarre results [33]: answer prefixes with an affinity to the science fiction show Star Trek (e.g.: "Captain's Log, Stardate [insert date here]: We have successfully plotted a course through the turbulence and are now approaching the source of the anomaly.") are able to boost some LLM's proficiency in mathematical reasoning. The reproducibility of LLM experiments seems doubtful given that seemingly irrelevant variations in prompting can have such big influence on results.

8. DISCUSSION AND CONCLUSION

In this work we applied the formal framework of validity [8] to music knowledge discovery, thereby enabling a critical appraisal of using Large Language Models (LLMs) for 'distant reading' of music knowledge. This was demonstrated for the extraction of psychophysical information from text by comparing GPT-3.5 results to human auditory experience. Specifically we re-evaluated our own previous results [7] of using ChatGPT to gain 'distant reading' estimates of music similarity. By evaluating a number of threats to validity and conducting additional experiments with ChatGPT, we were able to show that internal, construct and external validity of our approach are seriously compromised.

A re-analysis of music similarity ratings separate for different or similar pairs of music genres showed that in our experiment music similarity is confounded with genre. Both human participants and ChatGPT at least partly rely on the confounding factor of genre when judging music similarity, which is a clear breach of internal validity. This lead us to scrutinize the operationalisation of the experiment and its construct validity. A closer assessment of the exact questions being asked during the experiment made it evident that they are not precise enough concerning what is actually meant with music similarity. Post-experiment interviews with human participants made clear that they indeed used genre as indication of music similarity. Because of the blackbox nature of ChatGPT, and its doubtful relation to factuality, we had to conduct an additional experiment to corroborate that ChatGPT also relies on genre information. This additional experiment with music from a single genre lead to a complete breakdown of the correlation between human and ChatGPT estimates of music similarity. We also appraised external validity by asking whether our results would generalize to variations in the experimental setting like employing different LLMs or versions thereof or making slight changes to prompts. We conducted a repetition of the ChatGPT experiment with a nine month time lag and showed that correlation of results is moderate at best, although the LLM is supposedly still based on the same version of GPT-3.5.

The overarching question we wanted to answer with this work is whether LLMs can be used as a distant reading tool of music knowledge. The main obstacle seems to be the opaqueness of systems like ChatGPT which make it very hard to judge their construct validity. This opaqueness is evident from the developing team's own statements concerning their unwillingness to share details about their algorithm [6]. This has lead researchers to state that "it is particularly hard to perform scientific experiments, especially since human feedback causes their behaviours to change at a rapid pace" [34]. The latter statement points to the additional problem of constant re-training of models, which might explain the lack of external validity we observed when repeating our experiment with a nine month time lag. It has also lead to speculations as to how ChatGPT actually works, e.g. showing that it performs better when the correct output is a high-probability word sequence, indicating that one should be careful in low-probability situations [35]. This might be connected to the fact LLMs seem to memorize large parts of their training data [26]. It has also been pointed out that the "reasoning process" of LLMs is fundamentally different from humans, as LLMs basically just sample from a probability distribution [34]. As they are not embodied agents in the physical world, their understanding and knowledge lacks symbol grounding [36]. LLMs do not experience the world directly but model the world of text, which of course is a very indirect representation of the real world.

As a concluding comment we want to state that ChatGPT is not a suitable tool for distant reading of music knowledge because of its essentially black box nature which entails severe problems of judging its construct validity. Future work should explore whether open source alternatives like LLaMA [28], Alpaca [29] or Open-Assistant (https://github.com/LAION-AI/Open-Assistant) will be able to change assessment of the usefulness of large language models for distant reading.

9. ACKNOWLEDGMENTS

This research was funded in whole by the Austrian Science Fund (FWF) [10.55776/P36653]. For open access purposes, the authors have applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

10. ETHICS STATEMENT

For all experiments with human involvement, informed consent to participate in the respective studies was obtained from participants in accordance with university and international regulations.

As a potential societal implication we like to mention the fact that it is not known what precise data any of the ChatGPT versions have been trained on. There is however a reasonable suspicion that OpenAI, the company behind ChatGPT, did not obtain legal consent from all creators of text it used during training procedures. As a consequence, anyone using ChatGPT for their own purposes, inlcuding distant reading of music knowledge, would be implicated with the corresponding ethical issues.

11. REFERENCES

[1] S. Oramas, L. Espinosa-Anke, F. Gómez, and X. Serra, "Natural language processing for music knowledge discovery," *Journal of New Music Research*, vol. 47, no. 4, pp. 365–382, 2018.

[2] F. Moretti, "Conjectures on world literature," *New left review*, vol. 2, no. 1, pp. 54–68, 2000.

[3] ——, *Graphs, maps, trees: abstract models for a literary history*. Verso, 2005.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[5] 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*, vol. 33, pp. 1877–1901, 2020.

[6] OpenAI, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[7] A. Flexer, "Can ChatGPT be useful for distant reading of music similarity?" in *HCMIR23: 2nd Workshop on Human-Centric Music Information Research, Milan, Italy*, 2023.

[8] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin, 2002.

[9] B. L. T. Sturm and A. Flexer, "A review of validity and its relationship to music information research," in *Proceedings of the 24th International Society for Music Information Retrieval Conference*, 2023, pp. 47–55.

[10] A. Flexer, T. Lallai, and K. Rašl, "On evaluation of inter- and intra-rater agreement in music recommendation," *Transactions of the International Society for Music Information Retrieval*, vol. 4(1), pp. 182–194, Nov 2021.

[11] K. Siedenburg and C. Saitis, "The language of sounds unheard: Exploring musical timbre semantics of large language models," *arXiv preprint arXiv:2304.07830*, 2023.

[12] R. Marjieh, I. Sucholutsky, P. van Rijn, N. Jacoby, and T. L. Griffiths, "Large language models predict human sensory judgments across six modalities," *arXiv preprint arXiv:2302.01308*, 2023.

[13] Y. Zhang, J. Jiang, G. Xia, and 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.

[14] Y. Hou, J. Zhang, Z. Lin, H. Lu, R. Xie, J. McAuley, and W. X. Zhao, "Large language models are zero-shot rankers for recommender systems," *arXiv preprint arXiv:2305.08845*, 2023.

[15] P. Knees, E. Pampalk, and G. Widmer, "Artist classification with web-based data," in *Proceedings of the 5th International Conference on Music Information Retrieval*, 2004.

[16] D. Turnbull, L. Barrington, D. Torres, and 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.

[17] B. Logan, A. Kositsky, and P. Moreno, "Semantic analysis of song lyrics," in *IEEE International Conference on Multimedia and Expo (ICME)*, vol. 2. IEEE, 2004, pp. 827–830.

[18] F. Góes, Z. Zhou, P. Sawicki, M. Grzes, and D. G. Brown, "Crowd score: A method for the evaluation of jokes using large language model ai voters as judges," *arXiv preprint arXiv:2212.11214*, 2022.

[19] L. F. Góes, P. Sawicki, M. Grzes, D. Brown, and M. Volpe, "Is GPT-4 good enough to evaluate jokes?" in *Proceedings of the 14th International Conference on Computational Creativity*, 2023.

[20] P. Knees and M. Schedl, "Towards semantic music information extraction from the web using rule patterns and supervised learning," in *Workshop on music recommendation and discovery*, 2011, pp. 18–25.

[21] I. Fujinaga and S. F. Weiss, *Digital prosopography for renaissance musicians: Discovery of social and professional networks*. NEH White Paper, 2016.

[22] M. Sordo, J. Serrà Julià, G. K. Koduri, and X. Serra, "Extracting semantic information from an online carnatic music forum," in *Proceedings of the 13th International Society for Music Information Retrieval Conference*, 2012.

[23] R. F. E. Sutcliffe, T. Crawford, C. Fox, D. L. Root, E. H. Hovy, and R. Lewis, "Relating natural language text to musical passages," in *Proceedings of the 16th International Society for Music Information Retrieval Conference*, 2015, pp. 524–530.

[24] M. Schedl, A. Flexer, and J. Urbano, "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 523–539, 2013.

[25] F. Fabbri, "A theory of musical genres. two applications," in *Popular music perspectives*, D. Horn and P. Tagg, Eds. Göteborg and Exeter, International Association for the Study of Popular Music, 1982, vol. 1, pp. 52–81.

[26] M. Nasr, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee, "Scalable extraction of training data from (production) language models," *arXiv preprint arXiv:2311.17035*, 2023.

[27] A. Flexer and T. Grill, "The problem of limited inter-rater agreement in modelling music similarity," *Journal of New Music Research*, vol. 45, no. 3, pp. 239–251, 2016.

[28] 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, and G. Lample, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[29] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.

[30] L. Chen, M. Zaharia, and J. Zou, "How is chatgpt's behavior changing over time?" *arXiv preprint arXiv:2307.09009*, 2023.

[31] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," *arXiv preprint arXiv:2310.14735*, 2023.

[32] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[33] R. Battle and T. Gollapudi, "The unreasonable effectiveness of eccentric automatic prompts," *arXiv preprint arXiv:2402.10949*, 2024.

[34] M. Peeperkorn, D. Brown, and A. Jordanous, "On characterizations of large language models and creativity evaluation," in *Proceedings of the 14th International Conference on Computational Creativity*, 2023.

[35] R. T. McCoy, S. Yao, D. Friedman, M. Hardy, and T. L. Griffiths, "Embers of autoregression: Understanding large language models through the problem they are trained to solve," *arXiv preprint arXiv:2309.13638*, 2023.

[36] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.