Journal of Information and Data Management, 2026, 17:1, doi: 10.5753/jidm.2026.5578

Leveraging LLMs for Topic Modeling and Classification in Brazilian Funk Lyrics

Jesus Daniel Yepez Rojas [**Universidade Federal do Rio Grande do Sul** | ***jesus.rojas@inf.ufrgs.br***]
Bruno Tavares Santos [**Universidade Federal do Rio Grande do Sul** | ***bruno.tsantos@inf.ufrgs.br***]
Fabíola de Carvalho Leite Peres [**Universidade Federal do Rio Grande do Sul** | ***fabioladecarvalholeite@gmail.com***]
Karin Becker [**Universidade Federal do Rio Grande do Sul** | ***karin.becker@inf.ufrgs.br***]

Informatics Institute, Universidade Federal do Rio Grande do Sul, Av. Bento Gonçalves, 9500 - Agronomia, Porto Alegre - RS, 91501-970, Brazil.

Received: 18 February 2025 • **Published:** 13 March 2026

Song lyrics present unique challenges for topic modeling and classification due to their implicit discourse, reliance on figurative and poetic language, and use of slang. As a cultural expression of urban peripheries, Brazilian funk provides a rich social narrative. This work proposes LLMusic, an end-to-end framework for topic extraction and classification of song lyrics, using Brazilian funk as a case study. LLMusic synergistically combines prompt-based Large Language Models (LLMs) with advanced topic modeling techniques such as BERTopic, aiming to address the limitations of traditional methods for identifying subjectively represented topics in texts. Zero-shot prompting is also deployed for unsupervised classification of new lyrics based on the identified topics. Our assessments demonstrate that LLMusic outperforms BERTopic in identifying subjectively expressed topics while achieving strong performance in unsupervised topic classification. The paper describes the components of LLMusic for topic identification and topic classification and illustrates its effectiveness by analyzing the discourse in the most popular funk songs, highlighting its potential for large-scale lyrical analysis.

Keywords: Lyrics, Topic Modeling, BERTopic, Large Language Models, Prompt Engineering,

1 Introduction

The study of music provides a multidimensional view of the complex relationships between culture and society, offering an opportunity to develop interdisciplinary analyses. Through rhythms and lyrics, music serves as a reflection of cultural traditions, beliefs, and values [Serra *et al*., 2013]. Music lyrics express the experiences and perspectives of the individuals within a society, serving as a mirror that reflects their social conditions [Spencer-Espinosa, 2022]. In many musical genres (e.g., blues, folk, and rap), the lyrics reflect a given time's social reality and struggles.

Brazilian funk (BRFunk) is a musical genre born in Rio de Janeiro's slums (*favelas*). It incorporated elements of foreign genres such as soul, funk, American, R&B, and rap, mixed with characteristics of Brazilian genres such as samba. The lyrics combine day-to-day topics in communities, resilience, and social criticism, and often include themes such as violence, drugs, sex, and objectification of women. Despite the prejudice regarding the artistic value of the genre [Lopes, 2011], BRFunk broke through the bubble of parties in Rio de Janeiro's *favelas* and is today one of Brazil's most significant mass cultural manifestations, listened to by all social classes and with international projection.

Various qualitative studies have emphasized the social and cultural facets of the BRFunk music ([Lopes, 2011; Facina, 2012; Lopes and Facina, 2012; Pereira, 2014; Peres, 2023]). These works typically rely on a carefully curated selection of songs that exemplify specific phenomena examined, such as expressions of masculinity [Peres, 2023], the subtle racism present in the Brazilian society [Lopes, 2011], or the connection of this genre to violence [Brilhante *et al*., 2019]. These studies are limited to small data sets and rely on traditional analysis techniques. Computational tools can assist these qualitative studies, allowing, on a larger scale, the identification, analysis, classification, and summary of the discourses present in these texts.

Topic modeling techniques for song lyrics analyses have addressed sexism [Betti *et al*., 2023], common topics in lyrics of distinct musical genres [Calcina, 2022], topic identification [Junior *et al*., 2019], and sentiment analysis [Devi and Saharia, 2020]. LDA [Blei *et al*., 2003] is the most popular topic modeling technique, which relies on word-co-occurrence. LDA is limited in capturing semantic relationships in musical lyrics since they employ metaphors, similes, euphemisms, and figurative language, which can lead to a diverse representation of words in similar contexts. BERTopic [Grootendorst, 2022] is a topic modeling technique that explores embeddings, similarity, and density grouping, resulting in advances in the semantic interpretation of musical lyrics [Calcina, 2022]. However, by focusing on grouping semantically similar documents, it can generate irrelevant groupings from the discourse's point of view, which often is implicit. To illustrate this difficulty, let us consider simplified excerpts of real BRFunk songs. Excerpts like "God, she is so beautiful", "God guided me so far",' and "I'll send this guy to God" tend to be grouped in the same topic due to the similar evocation of God. However, they express different discourses, namely women's appearance, reflections on life, and violence, respectively.

Large Language Models (LLMs) have proven effective in

various natural language processing (NLP) tasks. One way to communicate with LLMs is through prompts. The way one designs prompts, a discipline called Prompt Engineering (PE) [Liu *et al*., 2023], can significantly impact the outcomes. With the proper prompts, LLMs have proven to be as effective as humans in text summarization tasks [Zhang *et al*., 2024]. Trained in vast quantities of text, LLMs may help overcome topic modeling challenges [Pham *et al*., 2024]. We regard it as a promising approach for grasping contextual and semantic nuances of subjective and poetic lyrics in which traditional topic modeling methods fail to achieve good performance [Watanabe and Goto, 2020].

A study [Ziems *et al*., 2024] has addressed the potential of LLMs for text classification using a number of tasks common to Computational Social Science (CSS), such as the classification of emotions, toxic language, mental illness, stance, politeness, among others. They conclude that direct prompts do not achieve the performance of state-of-the-art classification models for these tasks. They also evaluate whether LLMs can substitute human annotation in dataset annotation for supervised learning, suggesting that the traditional annotation process can be redesigned using LLMs as co-annotators of data. Similar conclusions were drawn in [Qin *et al*., 2023].

This work proposes LLMusic, a framework for topic extraction and classification of song lyrics. LLMusic synergistically combines the power of PE and LLMs with advanced topic modeling techniques such as BERTopic, aiming to address the limitations of traditional topic modeling methods for identifying subjectively represented topics in texts. Despite the potential for generalization of the method, our research initially focuses on Brazilian funk. The method requires as input a corpus of lyrics representative of the musical genre. We have leveraged LLMs in three ways: (1) as a component of our topic modeling approach, (2) for unsupervised classification of lyrics excerpts according to the topics identified, and (3) as co-annotator of an improved test dataset to assess the unsupervised classification of lyrics. In our topic modeling approach, first, we use prompts and LLMs to extract themes expressed in song excerpts, exploiting the generative capacity of LLMs for summarization. To create a robust and representative distribution of themes, we randomly combine song excerpts from a reference corpus in multiple iterations. Then, we use BERTopic to summarize this distribution into a non-redundant list of representative topics. In our unsupervised classification approach, we use PE over LLMs to assign the identified topics to music excerpts using (zero-shot prompts), enabling large-scale analysis of new lyrics. Finally, to assess the performance of LLMs in an unsupervised topic classification model, we improved an existing corpus using LLMs as a co-annotator, analyzing the consensus level between the human annotators and the LLM. The annotation process of such subjective and nuanced language, as used in BRFunk music, generated heated debates among the annotators, and the use of LLMs shed light on the challenging excerpts.

Using the curated collection from [Peres, 2023] as a reference corpus for BRFunk, we illustrate the application of LLMusic as a computational strategy to identify meaningful topics and analyze the top 100 most played BRFunk songs, showcasing its ability to identify the discourses embedded in a large set of lyrics. Our assessments demonstrate that LLMusic outperforms BERTopic in identifying subjectively expressed topics while achieving strong performance in unsupervised topic classification. Our results using a test dataset reveal that the unsupervised classification approach achieves a macro-averaged F1 of 81.23%.

This article presents the components of the proposed LLMusic framework, the results of deploying it to interpret Brazilian funk using songs curated in [Peres, 2023] as reference *corpus*, and the assessment of both the topic identification and topic classification tasks. This article extends our previous paper [Yepez *et al*., 2024b], providing a more in-depth discussion of the results. In particular, we describe the LLM-based co-annotation process used to create a test dataset for assessing the classification component of LLMusic. Based on the level of consensus between the LLM and humans as judges, we discuss the insights about the challenges related to nuanced, subjective, and unique language used in Brazilian funk to develop the narratives.

The contributions of this article are:

- the framework LLMusic, an end-to-end approach for identifying topics in lyrics of a given genre and using these topics to classify lyrics from a corpus targeted at analyzing a given phenomenon. Although the approach is generalizable, we focus on Brazilian Funk;
- an illustrative case study showcasing the use of LLMusic for discovering topics in BRFunk, and using these topics to obtain insights about the top-100 most played BRFunk songs;
- an annotated dataset relating lyrics excerpts to the identified topics, with a detailed discussion of the insights obtained about classifying BRFunk lyrics due to nuanced, subjective, and unique language. The excerpts were extracted from selected songs among the top-100 most played BRFunk songs.

The remainder of this paper is structured as follows. Section 2 explores related work. Section 3 presents the LLMusic framework, providing details on how it addresses the topic identification and topic classification tasks. Section 4 presents the results of deploying LLMusic in a case study. Sections 3.1 and 3.2 focus on evaluating the topic identification and classification tasks, respectively. Finally, Section 7 discusses the findings and outlines directions for future work.

2 Related Work

Studies that combine computational techniques and musical analysis are increasing, allowing the extraction of insights from the extensive amount of music available [Oramas *et al*., 2018]. They highlight how music has increasingly been viewed as a linguistic system and can benefit from contemporary NLP techniques, such as LLMs. Related work includes supervised and unsupervised applications for song lyrics analysis.

Supervised modeling techniques have been used in the analysis of large *corpus* of lyrics. In [Betti *et al*., 2023], models based on the BERT architecture were fine-tuned using an annotated database to identify sexism and gender bias. The trained models were then applied to a corpus composed of

377,808 English song lyrics to analyze the relationship between sexism and artist gender. As a supervised analysis, its reproduction is restricted to contexts where annotated reference databases exist for fine-tuning the models. Another limitation is that the topics for annotating the dataset must be known beforehand, which hinders the approach in studies that explore the themes contained in a large corpus.

LDA is the prevalent technique in music studies that uses unsupervised topic modeling techniques. In [Devi and Saharia, 2020], LDA is applied to sentiment classification. LDA is also deployed to investigate how specific terms (e.g., alcohol, relationships) are used in lyrics of Brazilian country music ("sertanejo"), a popular musical genre in Brazil. Since LDA extracts co-occurrence relations of words, the generation of clusters representing topics is limited to contexts in which artists employ common words to represent the same situations.

The use of BERTopic for topic modeling in lyrics is still recent. In [Calcina, 2022], BERTopic is used to discover similarities between lyrics from different musical genres (e.g., *folk* and *rap*), based on the sets of topics addressed in the respective songs (e.g., daily struggles, social problems). Despite an improved contextual interpretation due to embeddings and similarity, the language model tends to capture embbedings that represent the information explicitly contained in the *corpus*. The discourse in song lyrics is often implicit, using nuanced, figurative, and poetic language, metaphors, and unique slang to convey a message, which requires advanced abstraction skills to understand the semantics and its context.

LLMs have not yet been utilized as computational support for lyrics analysis. These models are designed to generate text in an auto-regressive manner, predicting tokens sequentially while taking into account the context provided by preceding tokens [Zhao *et al*., 2023].

Recent work highlights the generative potential of LLMs in text summarization tasks. In [Zhang *et al*., 2024], a human evaluation of ten different LLMs is performed, where it is suggested that the prompt configuration is more important than the model size in zero-shot summarization tasks. The work indicates that the summarization quality of LLMs is equivalent to that of humans.

The use of LLMs exploits the prompt learning paradigm for its parameterization. The basic structure of prompt-based learning is illustrated in Figure 1, containing a *input* $[X]$, a Template containing the prompt's set of instructions, and a *output* $[Y]$. The strategy (e.g., zero-shot, few-shot, chain of thought) chosen in the prompt creation process is fundamental, as it directly impacts the result of the template.

Some of the challenges to using LLMs include hallucinations (i.e., nonsensical or irrelevant responses from the model) [Huang *et al*., 2023] and dealing with the non-deterministic nature of LLMs (i.e., the same prompt can generate different responses). These questions require exploring what best suits the task at hand, such as mixing different types of prompts.

Input (X): *And I, I liked you so much*
Template: [X] Suggest 4 topics for this piece of lyrics: [Y].
Output (Y): Love; End of relationship; Memories; Longing.

Figure 1. Example of prompt to generate topics from a excerpt of music

Studies have addressed the potential of Prompt Engineering PE in topic modeling [Pengfei Liu and Neubig, 2023], [Pham *et al*., 2024]. [Pham *et al*., 2024] propose a *framework* where PE generates a distribution of topics by considering Wikipedia articles and summaries of notes from the American Congress. Topics are then refined and finally assigned to documents through a classification process again using PE. The method was superior to traditional topic modeling models. However, topic refinement requires manual intervention to analyze and define topic merging and deletion criteria. Furthermore, the topic assignment step requires the creation of *prompts* containing examples of each topic (few-shot learning).

Studies have also addressed the potential of LLMs as an alternative to manual annotation due to lower costs and higher scalability. In social science computing, a study [Ziems *et al*., 2024] has evaluated whether LLMs can substitute human annotation in dataset annotation for supervised learning in different domains. They conclude that they cannot replace annotators, suggesting, however, that the annotation process could be redesigned such that LLMs could act as co-annotators. A novel paradigm for Human-LLM co-annotation of unstructured texts was proposed in [Qin *et al*., 2023], proven to be an effective means to allocate work with superior performance compared to random baselines. This trend was also assessed as promising in [Bencke *et al*., 2024].

We contribute to the field by exploring LLMs and prompt engineering for fully unsupervised topic identification and classification of lyrics, specifically Brazilian Funk. We also explore a co-annotation process to extend and improve a test dataset to assess the performance of unsupervised lyrics classification.

3 Framework Overview

We propose LLMusic, an end-to-end approach for identifying topics in lyrics of a given genre and using these topics to classify lyrics from a corpus targeted at analyzing a given phenomenon. The proposed approach for topic modeling combines the text summarization capacity of LLMs through prompt engineering with the topic identification capacity of BERTopic. The summarization capacity of LLMs allows topics generated using LLMusic to be closer to human annotations than topics generated by traditional topic modeling approaches. As a result, LLMs can derive the main topics of music lyrics of a genre of interest.

Topic classification is a complementary aspect of LLMusic in the automated, unsupervised analysis of songs, given the topics that dynamically emerged from the proposed topic modeling approach. Topic classification enables the assignment of multi-topic labels to new, unforeseen songs according to a totally unsupervised process, hence an invaluable tool for larger-scale analyses.

Although LLMusic is agnostic regarding musical genres, we focus on BRFunk as a case study. There are several challenges in analyzing musical lyrics, given that the texts are nuanced, ambiguous, and subjective, with multiple themes and intertwined lyrical structures, making the automatic identification and classification of topics challenging for traditional methods. There are additional challenges in the BRFunk con-

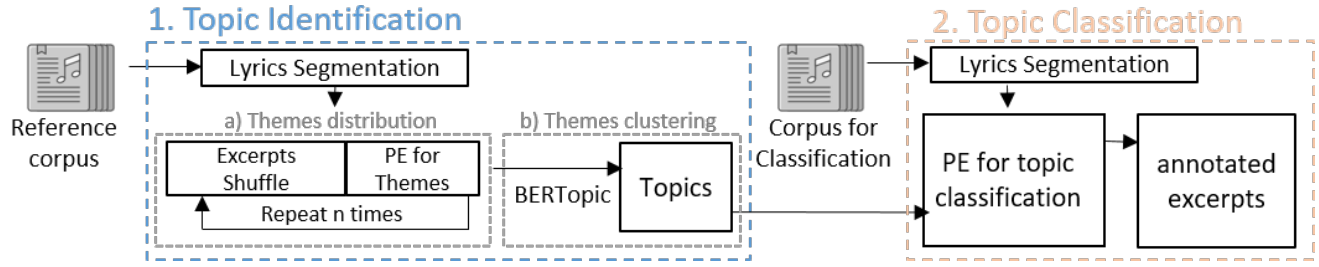

Figure 2. General overview of the LLMusic Framework

a) Template for Topic Identification

```
Given the following song excerpts,
suggest [n] topics that describe the
subjects addressed:
1.  [Excerpt]
2.  [Excerpt]
3.  [Excerpt]
    ...
n.  [Excerpt]
```

b) Template for Topic Classification

```
Evaluation of the Semantic Relationship
between [Topic] and [Excerpt]
1: No semantic relationship.
2: Weak semantic relationship.
3: Moderate semantic relationship.
4: Strong semantic relationship.
5: Very strong semantic relationship.
Respond exclusively with a number
between 1 and 5.
```

Figure 3. PE templates for topic identification and attribution

text since the lyrics use unique lyrics that reflect life on the *favelas* [Lopes, 2011].

Figure 2 provides an overview of the framework LLMusic, which highlights two main tasks: *Topic Identification* and *Topic Classification*, detailed in the remainder of this section. Both are based on zero-shot prompt structures, depicted in Figure 3, in a completely unsupervised process.

The framework assumes as input an unlabeled *corpus* of lyrics representative of the phenomenon or genre to be studied, referred to as Reference Corpus. Lyrics can cover multiple topics, even in the same excerpt, and the interpretation can be subjective and not deterministic. We divided each song of the *corpus* into excerpts such that we reach an analysis granularity that enables us to relate the specific portions of the song that are related to a specific topic.

3.1 Topic Identification

As shown in the left-hand side of Figure 2, the Topic Identification stage is divided in two phases: a) *Themes distribution creation* and b) *Themes clustering*. In the first phase, we rely on the generative ability of LLMs in a process that leverages PE to identify the different themes expressed in the lyrics iteratively. Then, we use BERTopic to summarize the themes into a smaller, nonredundant, more cohesive, and representative set of topics. This process is illustrated in Figure 4.

The strategy of LLMusic for the identification of topics has some advantages. First, it relies in a process that iterates through random shuffle of excerpts, which increases the diversity of themes identified through zero-shot PE, creating a reliable distribution through repetition. As a side effect, the Reference Corpus is reformulated through this combination, emphasizing themes initially marginalized in the lyrics. It mitigates the validity threats due to the size and possible bias of the Reference Corpus. Second, the topics abstract similar themes. Since BERTopic is based on density clustering, it helps eliminate outliers, prioritizing the most frequent themes within the distribution. By describing the musical *corpus* from the themes with higher frequency, we minimize the problems of hallucinations and irrelevant outputs, which are common in applying LLMs.

1) Creation of a Themes Distribution: To create a reliable distribution, we iteratively shuffle the segments to form random groups and use them as input for a PE model, repeating this process several times and aggregating the results at the end, as illustrated in 4. The steps are:

(i) *Excerpts Shuffle*: The excerpts of all the lyrics are mixed and randomly divided into groups. The premise is that the random shuffle of groups will provide a greater diversity of themes while addressing hallucinations and non-determinism. Additionally, it has the side effect of reformulating the initial *corpus*. Furthermore, by organizing the excerpts into smaller groups, the size of these groups can be adjusted to handle the maximum token size limitations of the specific LLM at hand (e.g. ChatGPT, Sabia). In case one needs to work with a smaller number of tokens, the number of segments per prompt can be reduced. On the other hand, a smaller number of prompts decreases the number of requests to the LLM's API.

(ii) *Prompt Engineering for themes extraction*: we explore PE according to the template in Figure 3.(a) to extract several themes for each group of random excerpts. The combination of randomization and repetition guarantees a wide distribution of themes. The final distribution of themes, which is the aggregation of all themes throughout the repetitions, allows the creation of the musical *corpus* from recurring themes, minimizing the influence of possible hallucinations of LLM.

2) Identification of Topics by Clustering of Themes: The results of the previous phase is a rich but possibly redundant distribution of themes. We use BERTopic to summarize these theses into a smaller set of representative topics. Leveraging semantic similarity between themes, BERTopic abstracts redundant and similar themes into topics, while treating low density regions (infrequent or low-representative themes) as outliers.

3.2 Topic Classification

The previous stage (Topic Identification) helps identify the relevant topics addressed in the lyrics of a given genre. *Topic classification* (right-hand side of Figure 2) is a complementary aspect of LLMusic in the automated, unsupervised analysis of songs, given the topics that dynamically emerged from the Reference Corpus. The aim of this stage is to classify a corpus of lyrics to analyze some phenomenon. Hence, this corpus is new, unlabeled, and possibly significantly large, from which one wishes to derive insights and knowledge. Unlike our

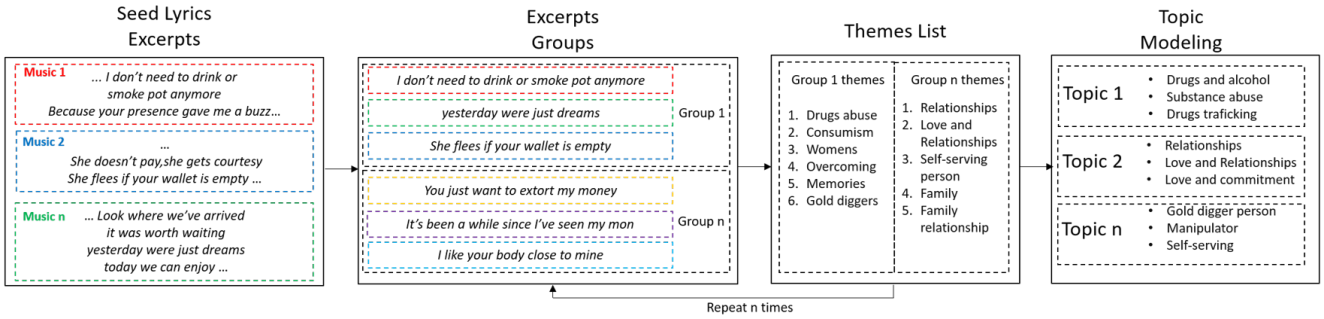

Figure 4. Illustration of the Topic Identification stage

previous approach [Yepez *et al*., 2024a], LLMusic is totally unsupervised, i.e., it does not require labeled data to train the multi-label classifier.

LLMusic adopts a zero-shot prompt-based approach for assigning topics, treating it as a non-supervised classification task. Topic classification also occurs at excerpt level, using the This stage explores an LLM with *template* of Figure 3.(b), where each request each requirement individually compares each topic with each excerpt of music, evaluating the level of semantic relationship between each pair of excerpt and topic.

4 Results and Illustrative Case Study

This section discusses the results obtained in leveraging the LLMusic framework on the BRFunk music genre. We show how topics extracted from a Reference Corpus of BRFunk songs can be used to classify excerpts from other songs of the same genre. We also illustrate the potential for larger-scale analysis by using the top 100 BRFunk songs as a case study. The public repository[1] contains details on the Reference Corpus, the topics found, and the notebooks with implementation details for each stage of LLMusic.

4.1 Reference Corpus

For the Reference Corpus we adopted the 18 songs analyzed in [Peres, 2023], which studies the expression of masculinity in BRFunk lyrics. The study observes the expression of power through violence, ostentation, objectification of women, and the glorification of the maternal figure—themes also identified in other studies (e.g. [Lopes, 2011]). The songs were extracted from the website *Letras.com.br* via scraping, and the division into excerpts corresponds to the separation of verse on the site. The 18 songs in the Reference Corpus were divided into 174 excerpts. The author of [Peres, 2023], an anthropologist expert on the social implications of BRFunk, served as an expert in the qualitative assessment of the results obtained.

4.2 Themes and Topics Identification

The LLM, prompting structure and strategy were determined experimentally. We employed Maritaca's `Sabiá-2-medium`, a Portuguese language model which, according to its documentation[2], demonstrates performance comparable to GPT 3.5[3], though still below that of GPT 4. We adopted Sabia for the theme distribution creation task since, in our experiments, it displayed a slightly better performance than GPT4 at a much lower cost. Regarding PE, the best result was obtained using the zero-shot strategy with the template shown in Figure 3.(a). We observed greater consistency in the results when a specific number of themes was requested. We defined 5 themes per prompt, a threshold low enough to avoid non-representative themes and high enough to ensure a good number of representative themes.

Figure 5 displays the top 10 most frequent themes identified in the final distribution of the seed *corpus*. This result was obtained by repeating this process 20 times with a different set of shuffled groups. As a result, we generate 1,191 different topics from approximately 400 documents. It is possible to observe the redundancy and similarity among themes (e.g., $dance\_and\_body\_movements$ and $music\_and\_dance$).

To extract topics from this theme distribution, we configured BERTopic with SBERT as the pre-trained model, reduced to 5 dimensions (UMAP), and a minimum of 25 themes per cluster (HDBSCAN). As a result, the 1,191 themes were reduced to 13 topics, where 260 themes were considered outliers.

All topics were evaluated by the expert anthropologist. She

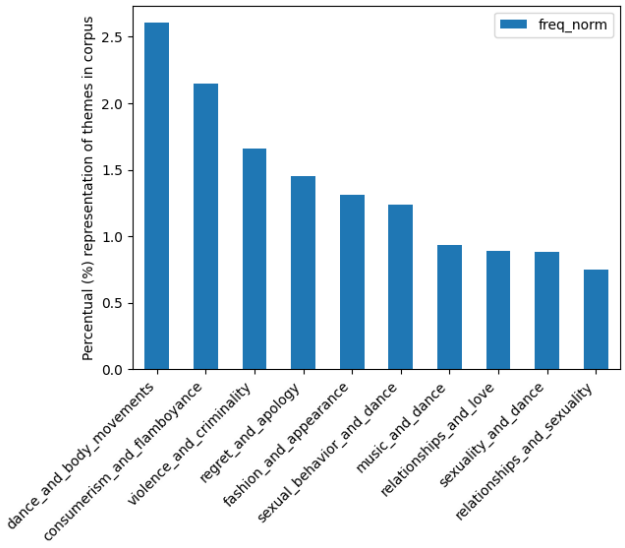

Figure 5. Themes Distribution

[1] `https://github.com/yepez26/FUNK_JIDM`

[2] `https://www.maritaca.ai/`

[3] `https://openai.com/index/chatgpt/`

interpreted the clusters and assigned semantics to the topics. She also recommended discarding two topics that were not meaningful for social discussions. One of them represented praise for the genre (e.g. "Funk is a must"), and the other introduced the DJs who perform songs (e.g. "Here is the great MC Marcinho"). The remainder discussions in this article disregard these two topics.

The result is summarized in Table 1. The table describes each topic using an identifier, the label description suggested by the expert, the most representative themes grouped in the topic, and excerpts from songs that illustrate the topic.

4.3 Topic Classification

We used lyrics extracted from the site *Letras.com.br* to classify the most played BRFunk songs according to the site *Letras.com.br*. We adopted the 11 topics in Table 1. The semantic relationship of each excerpt with each topic is quantified on a scale from 1 to 5 using the prompt template in Figure 3.(b). For this task, the LLM GPT4 performed better compared to Sabiá, which exhibited inconsistent behavior in its responses.

Figure 7 illustrates the topics assigned for two excerpts from the song *Novinha Terrível*, which describes a young girl from the *favela* from a masculine perspective. The narrative of the first excerpt is about clothing brands she wears, and the model related this excerpt to topic 7 (*Consumerism and Ostentation*) with the maximum score (5). The model classifies the second excerpt with a high score in topics 4 (*Appearance and behavior of women*), 5 (*Dynamics of seduction, sex, sexual desire*), and 6 (*Funk parties and dances*), since it criticizes that she prefers going to parties instead of engaging in a romantic relationship. These two excerpts display the potential of the approach to correctly classifying the identified topics despite the nuanced subjectivity of the language.

Figure 6 represents the distribution of topics in the 11 most played BRFunk songs, presented alphabetically. The weights represent the average of the topics considering all the excerpts of each song. The value 5 represents that all the excerpts of the song contain the topic, while the value 1 indicates that none of the excerpts of the song contain the topic. For example, the song *Amor de Verdade* (True Love), where the narrative is the love for a special person, receives a weight of 4.75 in topic 3 (*Relationships*) and 1.12 (i.e., close to nonexistent) in topic 4 (*Appearance and behavior of women*). The prejudice against BRFunk is often centered on the themes associated with topics 4-9, but it is clear that topic 3 is the most recurrent one in the 11 most played songs. Furthermore, note the recurrence of themes such as *Family Relationships* (2) and *Overcoming challenges* (11). Some excerpts from these songs illustrate the topics in Table 1.

The anthropologist assessed the scores assigned to the top-11 BRFunk songs. She examined the topic scores assigned to all excerpts and the aggregated scores per topic per music. According to her, in general, LLMusic produced consistent results in the narratives of the lyrics. For instance, *Angra dos Reis* describes a big party that the composer MC Daleste will throw in Angra dos Reis, a wealthy region of Rio de Janeiro's coast. In the lyrics, MC Daleste points out how unlikely this posh party is for someone who was born in the *favela* (topic 10), as himself. He describes how the party will be an event (topic 6), bragging about luxury brands and their cost (topic 7), and guaranteeing the presence of women (topics 4 and 5). *Fala mal de mim* (*Badmouth me*) and *Novinha Terrivel* (*Bad babe*) are somewhat similar, describing the strength, beauty and charm of the *favela* women (topic 10), from the perspective of a woman (empowerment) and a man (objectification), respectively. The lyrics describe how women wear and look (topic 4), how they are perceived by men (topic 5), and expectations about them in the *bailes funk* (topic 6). From the male perspective, men observe the "bad babe" as they drink expensive alcohol (topic 7), fantasizing about the possibilities, but from a female artist's perspective, the lyric point out that she behaves as she pleases and is ready to fight anyone not respecting her rights to be herself (topic 9).

The songs *Nunca vendi maconha* (*Never sold pot*) and *Fico assim sem você* (*I'm like this without you*) illustrate some of the challenges for classification. The former describes how one may feel good after consuming marijuana (topic 8), describing different hallucinations (topics 3, 4, 5, 6) after inhaling it. This lyrics exemplify the difficulty of perceiving some nuances, given that 100% of this song could be classified in topic 8. Still, only two more explicit excerpts were recognized due to the unique slang referring to the act of inhaling cannabis. *Fico assim sem você* is a very poetic love song that expresses through metaphors how the lovers complete each other (e.g., an airplane without wings). We suspect the model misclassified this song in topic 9 since it repeatedly expresses through these metaphors how painful it is to live one without the other.

4.4 Illustration: Insights from the 100 most played BRFunks

Finally, to illustrate the potential of the LLMusic framework, we analyze the 100 most played BRFunk songs[4] according to the site *Letras.com.br*. This site uses the visualization of *Youtube* for ranking. According to the expert, the number of views on *Youtube* is more representative than the ranking of other streaming platforms (e.g., *Spotify*), too expensive for the *favelas*. These 100 lyrics were segmented into 1,113 excerpts, which, after removing duplicates, resulted in 805 unique excerpts. The public repository details the music and the labels attributed to each excerpt.

We use the prompt in Figure 3.(b), adopting a threshold of $>= 3$ to determine the presence of the topic in the excerpt; otherwise, the topic was considered absent. Figure 8 shows the participation of each of the 11 topics of Table 1 within 100 BRFunk music in percentage terms, using two criteria.

According to the first criterion, a song can be associated with a topic if it contains at least one excerpt related to that topic. Considering this perspective, we observe that many songs primarily focus on themes of sexualization, women, and parties. This trend indicates that topics 6 (sexualization), 4 (women), and 5 (parties) dominate about 80% of the songs analyzed, reinforcing stereotypes commonly associated with

[4]Two songs of the Reference Corpus are among the top-100 songs, namely *"Amor de verdade"* and *"Baile de favela"*. To avoid bias, these songs were replaced with the next two most played songs.

Table 1. Topic Labels, Representative Themes and Illustrative Excerpts.

ID	Label	Representative Themes	Example Excerpts
1	Forgiveness and repentance	Regret and asking for forgiveness from parents Regret and asking for forgiveness	"Forgive me, mother, for not having listened to you ..." "Forgive me, mother, for having dropped out of school ..."
2	Family relationships	Family relationships and financial independence Affection and family relationships	"Oh, how I miss that boy running, smiling ..." "...what a disappointment, my son a drug dealer..."
3	Relationships	Relationships and indecision Relationships and communication	"In life there can be thousands, but none will be like you..." "...the love was real, that made my life happen..."
4	Appearance and behavior of women	Women's behavior and appearance in social situations Stereotypes of women's appearance and behavior	"She smells so good, wearing Morena Rosa ..." "The best one, and when she passes by I say 'wow!' "
5	Dynamics of seduction, sex, sexual desire	Sexual and romantic behavior Sexual behavior and physical attraction	"Leave my friends aside..." "So there's space for more six women..."
6	Parties and BRFunk dances	Funk culture and dance Music and funk culture	"Angra dos Reis, 40 degrees, I want a funk party ..." "From 1100 (motorcycle model), parties goin' on..."
7	Consumption and ostentation	Lifestyle and consumption of famous brand products Lifestyle and consumption of fashion and beauty products	"I'm going on a Hornet, on an Amarok ..." "20k to spend..."
8	Substances trafficking and consumption	Alcohol and drug consumption Drug consumption and violence	"Whoever says money doesn't grow on trees..." "Never sold pot..."
9	Violence and crime	Violence and vulgar language in music Violence and vulgar language in music and society	"Don't look away, it's the gang that's passing by..." "If you get provocative, you'll get beaten up."
10	Living in *favelas*	Geographical location and references to neighborhoods and *favelas* Representation of Brazilian favela culture and behavior	"The slum knows..." "That she's on another level..."
11	Reflections on life and overcoming challenges	Work and overcoming difficulties Reflection on life and overcoming obstacles	"Start getting used to it because I'm here to stay..." "No one can take me away from here..."

Excerpts of the songs "Perdoa mãe" (Forgive me mother) (1), "Mãe de traficante" (Mother of a Drug Dealer) (2), "Amor de verdade" (True love) (3), "Novinha terrível" (Bad babe) (4,10), "Fala mal de mim" (Speak ill of me) (9), "Mais amor, menos recalque" (More love, less jealousy) (7,11), "Angra dos Reis" (5,6) and "Nunca vendeu maconha" (Never sold pot) (8). All excerpts are translated for convenience.

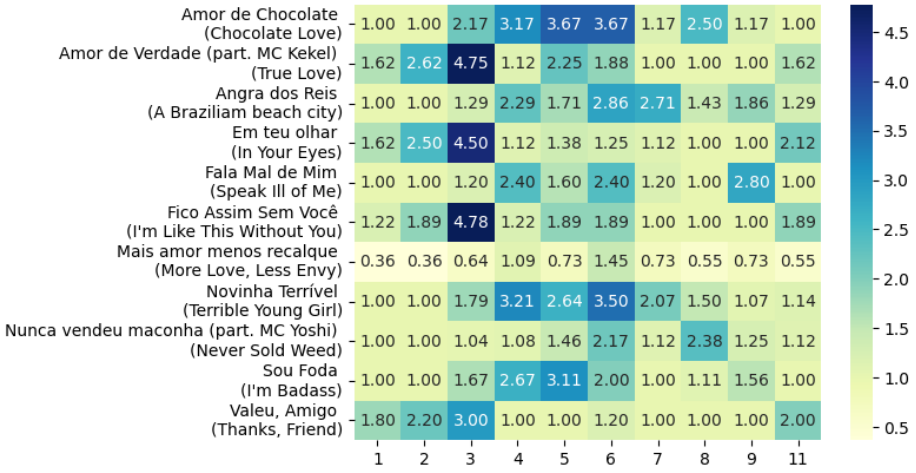

Figure 6. Distribution of topics in the 11 most played funks.

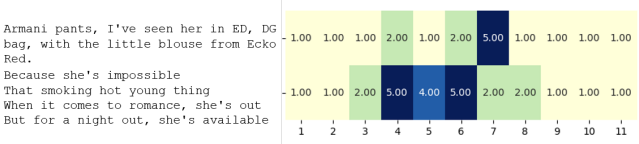

Figure 7. Classification of excerpts from the song *"Novinha Terrível"(Bad Babe)*

this genre. Additionally, around 50% of songs in the top 100 include a broader theme of relationships (topic 3).

The second criterion is stricter. We consider the lyrics address a given topic only if at least 50% of its excerpts are semantically related to it. This approach helps us focus on the song's main narratives rather than superficial mentions. Using this method, we still observe the prevalence of topics 6 and 4, followed closely by topics 5 and 3 in equal proportion.

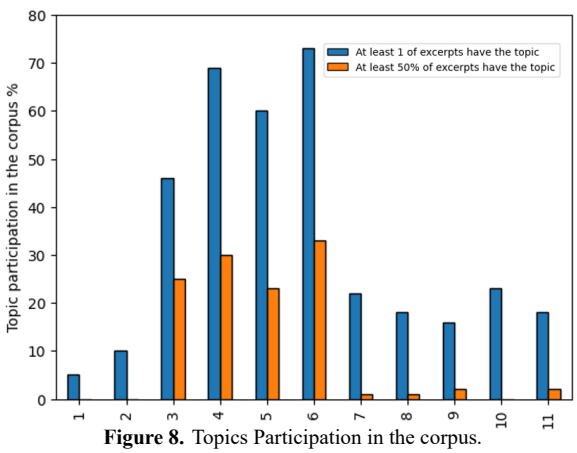

Figure 8. Topics Participation in the corpus.

Figures 6 and 8 illustrate the potential of the framework to provide insights from a larger *corpus* of music. The quantitative data show that the narratives present in BRFunk lyrics go beyond stereotypes, including relationships in general and with family, the daily life in the *favelas*, and stories of overcoming challenges from the perspective of the lack of opportunities in these communities. This variety reveals the multiplicity of experiences lived in the daily life of peripheral communities [Peres, 2023; Lopes, 2011].

5 Evaluation of the Topic Identification Task

To evaluate the quality of the topics generated in the process described in Section 3.1, we made a comparison using BERTopic directly on the Reference Corpus (Section 4.1). In BERTopic, UMAP was used to reduce the embeddings to 5 dimensions, and a minimum of 7 documents per cluster was set in HDBSCAN. These parameters were chosen aiming to maximize the coherence metric c_v [Röder *et al*., 2015]. The maximum value obtained for c_v was 0.54, resulting in 8 topics. The topics obtained with LLMusic resulted in the same c_v value. Both sets of topics were evaluated by an anthropologist specializing in BRFunk.

Table 2 presents the four (4) most representative words of each topic found in each approach. In LLMusic, the topics are identified with the same IDs as in Table 1, with topics 12 and 13 corresponding to those discarded based on the specialist's suggestion. In BERTopic, the topics are identified by the letters *A .. H*. One can observe the huge difference between the topics found considering the respective representative words.

The topics generated by BERTopic are based on the explicitly represented content and, therefore, are limited to the excerpts of the reference corpus, whose songs were carefully selected to study the expression of masculinity in BRFunk music. The topics primarily involve themes identified in [Peres, 2023], and the representative words reveal a certain level of relation with themes of objectification of women (B, E), violence and power (A, C, G), as well as behaviors typical of BRFunk parties (H). Some topics are less clear in their semantics (D, F), and the words do not always allow for a clear identification of the theme (e.g., G, A, C). There are no specific topics for ostentation, the maternal figure, or other topics that reflect BRFunk culture as a social portrait of the periphery [Lopes, 2011].

Regarding the topics identified by LLMusic, the anthropologist classified them as comprehensive and representative of the movement. It is important to emphasize that, despite being extracted from a reduced corpus with a specific focus, the generative capacity of the LLM, combined with the shuffling of excerpts, allowed for the capture of themes only tangentially expressed in these lyrics. Thus, it was possible to capture subtly represented themes, such as drugs, family relationships, life in the favela, etc. Moreover, the topics generated by LLMusic are clearly distinct from one another, not only in relation to the words that represent them, but also regarding the covered content. Themes such as family relationships, ostentation, romantic relationships, and drug consumption are clearly decipherable from the subset of words. In the topics generated by BERTopic, the interpretation requires a greater degree of abstraction from the evaluator, where the repetition of words (e.g., love, woman, boss) across topics and the lack of semantic cohesion in the set of words make it difficult to infer clear meanings from the generated topics.

Table 2. Topics extracted from the reference *Corpus* using LLMusic and BERTopic

ID	Topics
LLMusic	
1	forgiveness, religion, repentance, apology
2	family, relatives, relationships, mother
3	relationships, relationship, love, behavior
4	women, behavior, appearance, stereotypes
5	sexual, behavior, sexuality, relationships
6	behavior, dance, parties, environments
7	fashion, brands, style, appearance
8	violence, consumption, drugs, behavior
9	music, vulgar, violence, language
10	references, favelas, neighborhoods, culture
11	life, success, overcoming, work
12	funk, culture, behavior
13	funk, songs, culture, Brazilian
BERTopic	
A	us, job, who, mine
B	twerk, little booty, sit, big booty
C	go, today, party, mc
D	woman, wants, boss, loves
E	party girl, move, shake, slowly
F	drop, beat, mirror, whistle
G	bullet, hit, love, smelling good
H	favela, party, kid, ready

We conclude that LLMusic is able to identify topics implicitly represented, and that it has a good ability to handle figurative language, poetic expressions, and slang. Furthermore, it is robust with respect to the size, focus, and possible bias of the reference corpus.

6 Evaluation of the Topic Classification Task

6.1 Annotated Dataset Creation

Compared to our previous work [Yepez *et al*., 2024b], in this article, we improved and increased the size of the annotated test dataset used to assess the performance of the unsupervised multi-topic classification approach. The expansion of the dataset was motivated by two reasons. First, we wanted to increase the number of instances per topic to develop a more rigorous and fair performance assessment. Also, the original dataset was imbalanced, and some topics were underrepresented. Second, as we analyzed the topic classification errors in our previous assessment [Yepez *et al*., 2024b], we encountered issues regarding the lack of annotated labels, motivating the confirmation of the misclassifications or the existence of a relationship with the topic.

To streamline the process, we adopted a co-annotation process in which human annotators judge the labels automatically generated using an LLM. By considering the consensus among human annotators, as well as among the LLM and the human annotators, we obtained insights into the challenges of classifying lyrics due to nuanced and subjective language. The remainder of the section describes the annotation process and the insights obtained based on consensus. The resulting anotated dataset is available in the public repository.

6.1.1 The annotation task

The original dataset was compiled to train a supervised topic classification, and the process used to create it is described in [Yepez *et al*., 2024a]. In summary, with the aid of the funk expert, we selected a set of known songs among the top-100 most played BRFunk songs related to one or more topics in Table 1, which were divided into excerpts[5]. First, we clarified the meaning of each topic to three annotators. Then, we instructed each annotator to independently examine the excerpts and label them with one or more topics of Table 1 (or none of them). In total, the annotators examined 644 excerpts. If at least two annotators agreed, we considered an $excerpt_i$ to be labeled by a $topic_j$. The final dataset contained 131 unique excerpts, annotated with at least one topic (82 excerpts were assigned more than one label). Notice this process was labor-intensive, and over half of the initial excerpts were discarded.

We adopted a more streamlined data annotation approach to expand and improve the dataset based on a co-annotation process, in which human annotators validate the output of the LLM. To increase the dataset, we compiled a list of additional songs using suggestive titles and divided them into excerpts. We also included all excerpts among the false positive errors from our previous experiment [Yepez *et al*., 2024b] (419 pairs <*excerpt,topic*>). This resulted in 218 unique excerpts.

Next, we used GPT-4 and the prompt of Figure 3.(b) to classify each excerpt considering all topics of Table 1. We used the threshold $>= 4$ to consider an $excerpt_i$ to be labeled by a $topic_j$, discarding all pairs below this threshold. The automatic labeling produced 716 pairs of <*excerpt,topic*> for the 218 unique excerpts.

Finally, we asked three annotators to individually examine each pair < excerpt_i, topic_j> and to signal when they disagreed with an assigned label. Before the annotation, we clarified the meaning of each topic to the annotators and discussed some examples. A given pair <$excerpt_i$, $topic_j$> was included in the test dataset if at least two annotators agreed that $excerpt_i$ was semantically related to $topic_j$. The final dataset included 492 unique annotated pairs of <*excerpt,topic*> referring to 218 unique excerpts. Table 3 summarizes the number of positive instances for each topic of the final testset. We shall refer to this dataset as the *Annotated Corpus*.

6.1.2 Consensus between Annotators

Table 4 details the agreement rate between annotators, considering the percentage of excerpts in which all three annotators agreed (*Full Agreement*) and the average agreement percentage between pairs of annotators (*Mean Pairwise Agreement*). To compute these rates, we considered when the annotators agreed either on the existence or the nonexistence of a relationship between $excerpt_i$ and $topic_j$, considering any pair <$excerpt_i$, $topic_j$>.

Table 3. Quantity of annotate pairs <*excerpt,topic*> of the test dataset

ID	Topic Label	Qt.
1	Forgiveness and repentance	31
2	Family relationships	38
3	Relationships	36
4	Appearance and behavior of women	50
5	Dynamics of seduction, sex, sexual desire	65
6	Parties and BRFunk dances	40
7	Consumption and ostentation	26
8	Substances trafficking and consumption	35
9	Violence and crime	50
10	Living in *favelas*	47
11	Reflections on life and overcoming challenges	73
Total		**492**

Regarding *Full Consensus*, we observe significant variability across the different topics. While some topics demonstrate high levels of agreement, such as Topic 4 (83%) and Topic 8 (88%), others are the object of significantly lower consensus, notably Topic 10 (44%) and Topic 5 (48%). This variability suggests that some topics may be inherently more ambiguous or open to personal interpretation by the annotators. The overall agreement rate of 61% indicates a moderate consensus across the dataset, and the mean Cohen's Kappa among annotators was 47%. These figures reveal the difficulty of the classification task even for humans.

The *Mean Pairwise Consensus* column in Table 4 was calculated by averaging the consensus rate of the three pairs of annotators. Since only two annotators need to agree, this rate is significantly higher than full agreement. Some topics display a significant difference between the two consensus columns, such as Topic 10, with 27 percentage points (pp) of difference, and Topics 3 and 5, with 20 pp of difference. In other words, the topic is somewhat recognizable to most annotators, yet there is enough ambiguity and subjectivity to prevent a complete consensus. This difference is much smaller in other topics, such as Topic 2, 8, 4 and 7, with differences of 1, 4, 5 and 6 pp, respectively. This means that with a few exceptions, all annotators easily recognize and acknowledge these topics.

The variability in annotator agreement across different topics appears to be related to their specificity level and the topic's expression using explicit, evident cues. High annotation consensus levels were achieved in Topics 2, 4, 7, and 8 because their presence (or absence) in a text is often unquestionable, signaled by explicit and easily identified keywords. In contrast, the other topics exhibited lower agreement. Some of these topics are comparatively broader and more open to subjective interpretation, making it harder for annotators to pinpoint the semantic relationship to a specific topic. In other words, the topic can be expressed by a more implicit discourse and nuanced language, dependent on a context not present in the excerpt, making it more prone to subjective interpretation.

Table 5 presents examples of lyric excerpts related to topics 2, 4, 7 and 8. Overall, these excerpts contain evident, explicit

[5] Excerpts of the two songs in the intersection with the Reference corpus were excluded from the annotation task.

Table 4. Annotation Agreement (%).

ID	Label	Full Consensus (%)	Mean Pairwise Consensus (%)
1	Forgiveness and repentance	59	72
2	Family relationships	78	79
3	Relationships	55	75
4	Appearance and behavior of women	83	88
5	Dynamics of seduction, sex, sexual desire	48	68
6	Parties and BRFunk dances	64	76
7	Consumption and ostentation	83	89
8	Substances trafficking and consumption	88	92
9	Violence and crime	57	71
10	Living in *favelas*	44	71
11	Reflections on life and overcoming challenges	61	74
Total	**Overall Agreement**	61%	74%

cues guiding annotators to identify the themes accurately.

- *Topic 2 (Family Relationships):* The excerpt highlights support actions like "helping a brother", "giving a hand" and "giving advice". These phrases are direct indicators of kinship and assistance, where the narrative of family bonds is explicit.
- *Topic 4 (Appearance and Behavior of Women):* This topic is evident in phrases such as "girls reveal themselves at the party", "they criticize my hair and my makeup", and discussions about style and jealousy. These statements explicitly describe norms and social interactions grounded in female appearance.
- *Topic 7 (Consumption and Ostentation):* The semantic relationship with these excerpts is revealed by mentions such as "counting stacks of hundred-bills", "luxury vehicles" (e.g., Citroen, Hornet, Kawasaki Bandit), and "RR" motorcycle. These references signal material wealth and a desire for status; thus, the consumption narrative can be quickly identified.
- *Topic 8 (Substances Trafficking and Consumption):* This topic is explicitly indicated by expressions like "starting to use drugs" and "skipping class to spend time with the wrong crew" (involved in substance use). Such wording is an explicit marker of the topic's semantic meaning.

Topics 1, 3, 5, 6, 9, 10, and 11 have lower agreement consensus, where the difference between full and mean partial agreement ranges from 12 to 20 pp. By analyzing the annotated excerpts, we observed that while some excerpts can be easily related to the respective topic due to an explicit narrative, others lack context, are implicitly expressed, and are more open to free interpretation. The examples in Table 6 illustrate how the presence or absence of explicit cues to a given topic affects agreement among annotators. Although the excerpts are labeled using the same topic, the one on the left provides a more explicit narrative due to evident cues, and consequently, the semantic relationship with the respective topic was unanimously recognized by all annotators. The example on the right-hand side provides more implicit or subtle references about the topic; therefore, one of the annotators did not identify the relationship. We compare the evident and subtle cues of the excerpts in Table 6 as follows:

- *Topic 1 (Forgiveness and Repentance)*. The excerpt on the left explicitly asks for forgiveness ("I ask you for forgiveness"), unequivocally signaling repentance. By contrast, the excerpt on the right ("If I didn't surrender …") focuses on responsibility for changing the family history and providing a different future for an unborn child, but the narrative of regret is way more subtle.
- *Topic 3 (Relationships)*. The excerpt on the left plainly expresses a romantic desire ("I want to be your love… be more than just your friend"). The one on the right describes how one resisted the temptation of being with a charming person. While it still implies a romantic dilemma ("I remembered us"), the narrative is more nuanced and open for free interpretation (e.g., fleeting desire - Topic 5, reflecting over life - Topic 11).
- *Topic 5 (Sexual/Seductive Dynamics)*. The excerpt on the left highlights sexual attraction using explicit references to belly piercing and temptation. As for the one on the right, there is sexual content, but overshadowed by the complaints of financial exploitation ("extort my money") and ulterior motives ("eager to get pregnant"). This shift toward money and manipulation might prompt some readers to see it under relationships or ostentation rather than seductive/sex dynamics.
- *Topic 6 (Parties and BRFunk Dances)*. The left excerpt explicitly references the funk parties at the Rocinha *favela*, mentioning music styles ("charme, rap, melody, funk"). In contrast, the excerpt on the right focuses on getting a girl out of the party to have sex with her ("I'm crazy about you… I'll get you at this party…"). While the excerpt is definitely about sex (Topic 5), the funk party is the context in which the seduction takes place.
- *Topic 9 (Violence and Crime)*. The excerpt on the left openly discusses criminality ("armed bandits," "money from trafficking," "the boss never showed mercy"). The excerpt on the right suggests violence against a woman ("a peck, two punches, and three hooks!"), but it could be interpreted as a mere figurative language about the lack of interest in that woman.
- *Topic 10 (Living in favelas)*. The left excerpt ("the people of … the favelas of Rio de Janeiro… deserve peace") explicitly mentions favelas and advocates for the rights of their people. The excerpt on the right ("I find you

Table 5. Illustration of Topics with Higher Interannotation Agreement due to Explicit Clues

Topic ID	Topic Label	Excerpt with Evident Topic Cues	Key Points
2	Family relationships	*Family is who helps a brother, gives a hand to someone who is down on the ground, gives advice to get up.*	Emphasizes family support and brotherhood.
4	Appearance and behavior of women	*The girls around here reveal themselves at the party. No matter what I do, it becomes fashion among them. They criticize my hair and my makeup. Oh, how gross, talk all you want.*	Describes social dynamics, beauty standards, and jealousy among women.
7	Consumption and ostentation	*Counting hundred-bill stacks inside a Citroën. Then we invite them, because we know they'll come. Our transportation is good, on a Hornet or a 1100. Kawasaki, there's a Bandit, there's an RR too.*	Mentions wealth, luxury vehicles, and status symbols.
8	Substances trafficking and consumption	*And he started using drugs, skipping class at school to hang out with the crew. That was his life now.*	Describes drug use and its impact on lifestyle choices.

Notes: Excerpts extracted from the following songs: *Família* (MC Tikão); *Fala Mal de Mim* (Ludmilla); *Plaquê 100* (MC Guimê); *Mãe de Traficante* (MC Daleste). All excerpts are translated for convenience.

beautiful with that turban… forgive me if I became a funk singer… we're beating this mess.") indirectly references the local culture (e.g., turban, funk) and its struggles, without explicit references to the *favelas*, making it less obviously to contextualize the excerpt in that environment.

- *Topic 11 (Reflections on Life and Overcoming Challenges)*. The excerpt on the left ("Thank God I'm not just another one leeching off the earth… No one can stop this journey") emphasizes personal growth and resilience. In the excerpt on the right, expressions such as "May God bless the favelas" may lead to a focus on community solidarity (Topic 10), rather than individual resilience ("the ones who fight the revolution"). Although these two topics are valid, the latter is expressed more subtly.

Comparing the examples in Tables 5 and 6, it becomes clear that topic identification hinges significantly on the presence or absence of explicit, keyword-based cues. More specific topics are often expressed using more evident cues (e.g., references to family support, material goods, or drug use), which explain the higher annotator agreement. In contrast, broader themes such as forgiveness or relationships often rely on subtler, more implicit and nuanced language, which results in subjective interpretation, hence explaining the higher levels of disagreement between annotators. We also realize that the excerpt's division may result in a loss of context that affects the comprehension of the narrative. This contrast highlights the critical role of clear cues in guiding consistent labeling and demonstrates why some topics are inherently more challenging to label accurately.

6.1.3 Humans as Judges of LLM Automatic Labeling

To assess the alignment between human annotators and the topics assigned GPT-4, we analyzed agreement rates across the 218 lyric excerpts related to one or more 11 topics from Table 1 (i.e., 716 pairs <excerpt, topic>). Table 7 presents different levels of consensus: (i) full agreement between GPT-4 and all three human annotators, (ii) agreement between GPT-4 and at least two annotators, (iii) agreement between GPT-4 and at least one annotator. The agreement is measured in terms of acknowledgment of the relationship of a given topic_j to excerpt_i.

The results reveal patterns regarding how GPT-4 aligns with human judgment. With a single exception, we observed higher agreement rates in the same topics in the humans achieved higher consensus, i.e., topics 4 (Appearance and Behavior of Women), 7 (Consumption and Ostentation), and Topic 2 (Family Relationships). The agreement levels with GPT-4 considering three annotators range from 83% to 75%, and 98% to 86% considering at least two annotators. The exception is Topic 8 (Substances trafficking and consumption). While this topic is related to the highest agreement level among annotators (88%), complete agreement with GPT-4 annotations was only 58% (67% considering two annotators). This suggests annotators likely share contextual knowledge about substance use and trafficking, such as unique slang or broader cultural or social understanding, which may not be fully represented in the LLM's training data.

Conversely, topics characterized by subjectivity and implicit narrative resulted in lower agreement of humans with GPT-4. The lowest agreement was observed for Topic 9 (Violence and Crime), Topic 10 (Living in favelas), Topic 3 (Relationships), and Topic 5 (Sexual/Seductive Dynamics). Considering three annotators, the agreement with the GPT-4 assigned label is 11%, 26%, 35% and 38%, respectively. Considering agreement by at least annotators, the agreement ranges from 27% (Topic 9) to 55% (topics 3 and 10). These topics involve more context-dependent cues, often requiring an understanding of implicit references, slang, or cultural context that GPT-4 may not fully grasp. Interestingly, the agreement rates increase considerably considering at least one annotator, meaning that GPT-4 often captured aspects of these topics but struggled to meet the stricter agreement thresholds with all annotators, but that at least one of the annotators saw evidence of a relationship to the topic.

A particular case is Topic 6 (Parties and BRFunk Dances). While human annotators had a moderate inter-agreement rate

Table 6. Examples of Evident and Subtle Topic Cues per Topic

Topic ID	Topic Label	Excerpt with Overt Topic Cue	Excerpt with Subtle Topic Cue
1	Forgiveness and repentance	I know I'm not the perfect bro. I own my mistakes. If I act like this, I ask your forgiveness.	If I didn't surrender, who was going to change my family's story — a future for my daughter, not even born yet, but this responsibility I carry within me, ready to get into the minds of these talkers.
3	Relationships	I want to be your love, let's not wait for anyone else. Tonight is so beautiful, let's take the next step. I want to be with you, be more than just a good friend.	The guy who approached me had his own charm. I almost said yes, but I remembered us — I guess I ran away.
5	Dynamics of seduction, sex, sexual desire	Piercing on her sexy belly, well-defined waist, check out this girl. Man, what a temptation. When she passes by, everyone gets worked up, even the DJ goes wild.	You don't love me. You just want to extort my money. You're eager to get pregnant—that's why you fuck me with no protection.
6	Parties and BR-Funk dances	From Friday to Sunday in Rocinha, the hill is full of pretty girls who come to enjoy the party. Listening to charme, rap, melody or montage. It's funk uphill, funk downhill—where should I go.	I'm crazy about you; your eyes don't fool me. I'll get you out of this party and wreck you in bed.
9	Violence and crime	On the corner of the hill, there are only armed bandits. Here, it's just thugs making money from trafficking. Over there in the favela, the boss never showed mercy, and the kids on lookout were born ready.	She wants some sweet affection. A peck, two punches, and three hooks! Feeling sorry? Then take her home, because not even for free I'd want that woman.
10	Living in *favelas*	Hey, watch over the people of the favelas. The favelas of Rio de Janeiro deserve peace, you know? Not only the favelas of Rio, but the favelas of our Brazil.	Oh mama, I find you beautiful with that turban. Forgive me if I became a funk singer. We're beating this mess. I'm very proud.
11	Reflections on life and overcoming challenges	Thank God I am no longer one who is parasitizing on earth. I'm running towards progress, to be able to win. No one can stop us.	A hail to all in the "quebrada" (hood). For all good boys. May God bless the favelas, the ones who fight the revolution.

Notes: Excerpts extracted from the following songs: *Champagne* (Veigh); *Perdoa Mãe* (MC Diogo da VN); *Cinderela* (MC Gui); *Funk de Pelúcia* (Tati Zaqui); *Gata da Favela* (Fagner Pinheiro); *Mete Bala, Te Amo* (MC Delta); *Endereço dos Bailes* (MC's Júnior e Leonardo); *Tráfico* (Dfideliz); *Sou Foda* (Os Avassaladores); *Favela* (MC Cabelinho feat. Filipe Ret); *Obrigado, Mãe (Pt. 2)* (MC Hariel); *Favela* (MC Menor MR); *Dias de Luta* (MC Paulin da Capital). All excerpts are translated for convenience.

Topics	1	2	3	4	5	6	7	8	9	10	11	Total
LLM and all 3 annotators agree	52%	75%	35%	83%	38%	17%	79%	58%	11%	26%	61%	38%
LLM and at least 2 annotators agree	79%	86%	55%	98%	64%	31%	96%	67%	27%	55%	94%	58%
LLM and at least 1 annotators agree	93%	97%	80%	100%	90%	53%	96%	70%	54%	82%	100%	77%

Table 7. Agreement levels between AI and annotators across topics

(64%), GPT-4's full agreement with all annotators is just 17%, a significant mismatch. A possible explanation is that GPT-4 may struggle to differentiate between general social interactions and cultural elements specific to BRFunk, incorrectly recognizing a semantic relationship. Unlike topics such as consumption and ostentation, where material references act as clear clues, party-related content might be more fluid and context-dependent, requiring recognition of subtle textual and cultural signals.

A key takeaway from these observations is that GPT-4, much like human annotators, performs better in topics with evident lexical cues but struggles with implicit and context-dependent narratives. This aligns with the observations on annotation consensus developed in Section 6.1.2.

In summary, the LLM's agreement with annotators appears to be influenced by three key factors:

- Explicit lexical cues: when evident keywords, cues or direct references exist, GPT-4 aligns more closely with human annotators;
- Contextual ambiguity: Topics requiring broader understanding of the discourse within its context (e.g., Violence, Living in favelas) show higher disagreement;
- Overlap with multiple themes – Some excerpts contain elements of multiple topics, making classification more

Table 8. Performance of Topic Classification

Metrics	1	2	3	4	5	6	7	8	9	10	11
Positive Class											
Precision	100.00%	100.00%	100.00%	61.22%	86.96%	85.71%	100.00%	91.67%	86.67%	79.31%	60.00%
Recall	33.33%	23.08%	40.00%	81.08%	58.82%	50.00%	52.63%	55.00%	59.09%	76.67%	12.50%
F1-score	50.00%	37.50%	57.14%	69.77%	70.18%	63.16%	68.97%	68.75%	70.27%	77.97%	20.69%
Negative Class											
Precision	96.26%	95.35%	94.29%	95.86%	92.82%	94.12%	95.67%	95.63%	95.57%	96.30%	90.14%
Recall	100.00%	100.00%	100.00%	89.50%	98.37%	98.97%	100.00%	99.49%	98.98%	96.81%	98.97%
F1-score	98.10%	97.62%	97.06%	92.57%	95.51%	96.48%	97.79%	97.52%	97.24%	96.55%	94.35%
Macro-averaged Metrics											
Macro Precision	98.13%	97.67%	97.14%	78.54%	89.89%	89.92%	97.84%	93.65%	91.12%	87.80%	75.07%
Macro Recall	66.67%	61.54%	70.00%	85.29%	78.60%	74.48%	76.32%	77.25%	79.04%	86.74%	55.73%
Macro F1-score	79.40%	75.51%	81.37%	81.78%	83.86%	81.48%	85.75%	84.66%	84.65%	87.27%	63.97%
Accuracy	96.33%	95.41%	94.50%	88.07%	92.20%	93.58%	95.87%	95.41%	94.95%	94.04%	89.45%

subjective and leading to variation among both humans and the LLM.

The goal of improving the original test dataset was to confirm the topic misclassifications in our previous assessments [Yepez *et al*., 2024b]. For this analysis, we included all the excerpts wrongly associated with a topic (false positives) in that assessment. An interesting result is that, out of 419 pairs previously regarded as false positive errors, only 68,5% were confirmed as such. In other words, at least two annotators agreed with the semantic relationship identified by GPT-4 of an excerpt with a topic (31,5%), which was not recognized in our original test dataset.

6.2 Classification Performance

Using the Annotated Corpus described in Section 6.1.1, we assess the performance of the proposed unsupervised classification approach using precision, recall, and F1. The instances of the positive classes (i.e., relationship to each one of the topics) are the ones in Table 3. All the non-existent relationships between a given topic and the 228 excerpts are regarded as examples of the respective negative class. The overall approach performance is measured using macro-averaged precision, recall, and F1. The results are displayed in Table 8.

The overall model performance is good, considering both the positive and negative classes (i.e., the ability to determine if the excerpt is related or not to a topic). The average accuracy considering all topics is 93.62% and the macro-averaged F1 is 81.23%. The macro-averaged precision varies from 75.07% (Topic 4) to 98.13% (Topic 1), while the macro-averaged recall ranges between 55.73% (Topic 11) and 61.74% (Topic 2). With a single exception (Topic 4), the precision of the model is better than the recall.

Considering only the positive classes (i.e., the relationship to a topic), we observe good precision performance at the expense of recall. While precision ranges from 60% (Topic 11) to 100% (Topics 1, 2, 3 and 7), the recall ranges from 12.5% (topic 11) to 81% (topic 4).

On the other hand, the model's performance on identifying inexistent relationships to a topic (i.e., negative class) is very good. Despite precision being higher than recall, we observe a better trade-off between these metrics. While precision ranges from 90.14% to 98.13%, the recall ranges from 89.5% to 100%.

The poor performance on the recall can be explained by the difficulty of the LLM to grasp the subtle cues and contextual knowledge of the narrative in an excerpt, as discussed in Section 6.1. We conclude that overall, the performance is good, but recall in the topic assignment is undoubtedly an aspect to be improved in our approach; otherwise, topics may be underrepresented in large-scale analyses.

7 Conclusions and future work

In this article, we presented LLMusic, a framework for unsupervised topic identification and classification in song lyrics, which explores the potential of LLMs through prompts to overcome the limitations of traditional topic modeling methods. Despite the potential for generalization, we limited our research to the Brazilian funk genre, in which the lyrics express everyday narratives of the peripheries using slang, figures of speech, and subjectivity. Given a Reference Corpus of the musical genre, we explore the generative capacity of LLMs to interpret randomly combined musical excerpts to identify themes covered in the lyrics. We leverage BERTopic to produce a list of condensed, coherent, and non-redundant topics. The topics derived for the genre can then be assigned to new song excerpts, also following an unsupervised, zero-shot prompting approach.

A case study illustrated the potential of a larger-scale analysis, confirming that the genre goes beyond stereotypes, including narratives about resilience, life in the *favelas*, family, regrets, among others. With an anthropologist expert in funk, we assessed the relevance and representativeness of the topics extracted in our case study. Our results display a superior performance compared to a conventional topic modeling approach.

LLMusic removes the need for annotated data since the classification of new songs is fully unsupervised. Using the improved Annotated Corpus with 218 excerpts, we achieved an average accuracy of 93.62% and a macro F1 of 81.23%, considering all topics. The approach performs better in recognizing an excerpt is not related to a topic and displays good precision in assigning a topic to an excerpt. However, future work must address the low recall in topic assignment, since this limitation results in the underrepresentation of topics in large-scale analysis.

We devised a co-annotation strategy for topic classification

performance assessment to extend and improve an existing annotation dataset. By observing inter-agreement rates between humans and between the LLM and humans, we obtained insights into the challenges of classifying lyrics. While LLM and humans tend to pinpoint the topics in excerpts with explicit cues and references (e.g., brands, material wealth, drug use, family ties), they both struggle in identifying topics with implicit meaning or need of contextualized knowledge, such as violence or romantic relationships, emphasizing the difficulty of fully automating this process. We also realized that this approach was effective in suggesting relationships to topics that were not originally recognized by human annotators.

The framework is characterized by flexibility and adaptability, allowing configuration of corpus size, lyrics separation criteria, and LLM models. It is also robust in terms of the characteristics and size of the Reference Corpus. Additionally, its modular structure allows independent replacements of steps and the use of different prompt structures both in generating themes and assigning topics, in addition to replacing BERTopic with another MT model in grouping themes. The proposed approach can be used as an information retrieval tool, supporting qualitative studies searching for representative songs that address a specific theme. LLMusic also serves to identify songs with certain characteristics, allowing researchers to delve deeper into subtopics (e.g., detail the discourse about women based on the lyrics associated with topic 4), or to search for and characterize songs for qualitative studies.

Future work involves improving the performance of the classification approach, a larger scale and temporal analysis of BRFunk to analyze changes in the genre, generalizations of the method to other musical genres, support for topic evaluation, and improvements to the LLM framework itself.

Funding

Research partially supported by the National Council for Scientific and Technological Development (CNPq) - grants 309334/2022-5 and 131387/2023-5.

Authors' Contributions

According to CRediT the authors participated in the following activities:

JY: Conceptualization, Methodology, Validation, Formal Analysis, Resources, Investigation, Software, Writing – Review & Editing, Visualization.

KB: Conceptualization, Methodology, Validation, Formal Analysis, Writing – Review & Editing, Supervision, Project Administration.

BS and FP: Validation, Formal Analysis.

JY is the main contributor and writer of this manuscript. All authors approved the final manuscript.

Availability of data and materials

The anotated dataset produced in this work is available at github `https://github.com/yepez26/FUNK_JIDM`.

References

Bencke, L., Paula, F., dos Santos, B., and Moreira, V. P. (2024). Can we trust llms as relevance judges? In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 600–612, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2024.243130.

Betti, L., Abrate, C., and Kaltenbrunner, A. (2023). Large scale analysis of gender bias and sexism in song lyrics. *EPJ Data Science*, 12(1):10.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brilhante, A. V. M., Giaxa, R. R. B., Branco, J. G. d. O., and Vieira, L. J. E. d. S. (2019). Cultura do estupro e violência ostentação: uma análise a partir da artefactualidade do funk. *Interface-Comunicação, Saúde, Educação*, 23:e170621.

Calcina, Erik e Novak, E. (2022). Measuring the similarity of song artists using topic modelling. In *Proc. of the 25th Intl. Multiconference Information Society - Data Mining and Data Warehouses (SiKDD)*, page 103–106.

Devi, M. D. and Saharia, N. (2020). Exploiting topic modelling to classify sentiment from lyrics. In *Proc. of the 2nd Intl. Conferemce on Machine Learning, Image Processing, Network Security and Data Sciences (MIND)*, pages 411–423.

Facina, A. (2012). Que batida é essa? In CASTRO, André e HAIAD, J., editor, *Funk, Que batida é essa??*, pages 213–228. Hunter Books.

Grootendorst, M. (2022). BERTopic: Leveraging bert and topic modeling for efficient document clustering. `https://maartengr.github.io/BERTopic`.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., *et al*. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Junior, J. S., Rossi, R., and Lobato, F. (2019). Uma abordagem baseada em letras para a descoberta de conhecimento da música brasileira: o sertanejo como um estudo de caso. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 949–960. DOI: 10.5753/eniac.2019.9348.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023a). Pre-train, prompt, and predict: A systematic survey ofprompting methods in natural language processing. *ACMCom-put.*, 55(9):35.

Liu, P., Yuan, W., Fu, J., Zhengbao, Jiang, Hayashi, H., and Neubig, G. (2023b). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Lopes, A. C. (2011). *Funk-se Quem Quiser: No Batidão Negro Da Cidade Carioca*. Bom Texto FAPERJ.

Lopes, A. C. and Facina, A. (2012). Cidade do funk: expressões da diáspora negra nas favelas cariocas. *Revista do Arquivo Geral da Cidade do Rio de Janeiro*, 6:193–206.

Oramas, S., Espinosa-Anke, L., Gómez, F., and Serra, X. (2018). Natural language processing for music knowledge discovery. *Journal of New Music Research*, 47:365–382.

DOI: 10.1080/09298215.2018.1488878.

Pereira, A. B. (2014). Funk ostentação em são paulo: imaginação, consumo e novas tecnologia da informação e da comunicação. *Revista Estudos Culturais*, (1):1–18.

Peres, F. C. (2023). Puta ou santa: as relações com mulheres enquanto elemento constituinte das masculinidades do funk brasileiro? In *Anais do IV Encontro Anual de Antropologia do Mercosul*.

Pham, C. M., Hoyle, A., Sun, S., Resnik, P., and Iyyer, M. (2024). Topicgpt: A prompt-based topic modeling framework. `https://doi.org/10.48550/arXiv.2311.01449`.

Qin, C., andZhuosheng Zhang, A. Z., Chen, J., Yasunaga, M., and Yang, D. (2023). Is chatgpt a general-purpose natural language processing task solver? In *Proc. of the 2023 EMNLP*, pages 1339–1384.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez Gutiérrez, E., Gouyon, F., Boyer, H., Jordà Puig, S., *et al*. (2013). Roadmap for music information research.

Spencer-Espinosa, C. (2022). Music and social change. reflections on the relationship between sound and society. *Review of the Aesthetics and Sociology of Music*, 53:57–76. DOI: 10.2307/48689101.

Watanabe, K. and Goto, M. (2020). Lyrics information processing: Analysis, generation, and applications. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pages 6–12.

Yepez, J., Tavares, B., Peres, F., and Becker, K. (2024a). Llms got the funk: leveraging llm, prompt engineering and fine-tuning for topic modeling on brazilian funk lyrics. In *Proc. of the 2024 Conf. on Web Intelligence Conference (WI-IAT)*.

Yepez, J., Tavares, B., Peres, F., and Becker, K. (2024b). Na batida do funk: modelagem de tópicos combinando llm, engenharia de prompt e bertopic. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 613–625, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2024.243148.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. (2024). Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2023). A survey of large language models.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2024). Can large language models transform computational social science? *Comput. Linguistics*, 50.