Leveraging RAG for a Low-Resource Audio-Aware Diachronic Analysis of Gendered Toy Marketing

Luca Marinelli[1] **Iacopo Ghinassi**[2] **Charalampos Saitis**[1]
[1]Centre for Digital Music, Queen Mary University of London, UK
[2]College of Computing and Data Science, Nanyang Technological University, Singapore
{l.marinelli, c.saitis}@qmul.ac.uk, iacopo.ghinassi@ntu.edu.sg

Abstract

We performed a diachronic analysis of sound and language in toy commercials, leveraging retrieval-augmented generation (RAG) and open-weight language models in low-resource settings. A pool of 2508 UK toy advertisements spanning 14 years was semi-automatically annotated, integrating thematic coding of transcripts with audio annotation. With our RAG pipeline, we thematically coded and classified commercials by gender-target audience (feminine, masculine, or mixed) achieving *substantial* inter-coder reliability. In parallel, a music-focused multitask model was applied to annotate affective and mid-level musical perceptual attributes, enabling multimodal discourse analysis. Our findings reveal significant diachronic shifts and enduring patterns. Soundtracks classified as energizing registered an overall increase across distinct themes and audiences, but such increase was steeper for masculine-adjacent commercials. Moreover, themes stereotypically associated with masculinity paired more frequently with louder, distorted, and aggressive music, while stereotypically feminine themes with softer, calmer, and more harmonious soundtracks. Code and data to reproduce the results are available on *github.com/marinelliluca/low-resource-RAG*.

1 Introduction

Toy advertisements are a rich site for investigating gendered multimodal discourse, with five decades of research showing persistent and marked gender polarization (Verna, 1975; Johnson and Young, 2002; Marinelli et al., 2024). However, prior studies have relied on manual annotation of relatively small corpora, limiting the ability to conduct large-scale, longitudinal analyses, which are necessary to track the societal impact of evolving stereotypes.

Large language models (LLMs) offer a promising outlook for analyses of large corpora (Xiao et al., 2023; Alonso del Barrio et al., 2024; Gao and Feng, 2025). We extended earlier research on gendered toys marketing by integrating transcript-based thematic coding via RAG and audio-based automatic affective and music tagging. Notably, we computed inter-coder reliability scores to assess the quality of the results.

This study was conducted under self-imposed compute constraints to highlight the benefits of using small open-weight models, rather than relying on pay-walled APIs. There are two key reasons for this approach: first, to ensure reproducibility; and second, to keep computational costs low, making it feasible to replicate this study on consumer hardware. A key issue with pay-walled technology is model deprecation: when models endpoints are retired and become inaccessible, any research based on them becomes immediately non-reproducible. For these reasons, we used smaller open-weight LLMs (4–9 billion parameters) deliberately steering clear of resource-heavy alternatives.

The contributions to the fields of Digital Humanities and Computational Linguistics are manifold. First, this study shows that small open-weight language models combined with RAG can achieve substantial inter-coder reliability, making large-scale discourse analysis feasible on consumer-grade hardware without relying on commercial APIs. Second, it provides a reproducible pipeline that integrates linguistic thematic coding with audio-based affective and musical analysis, enabling large-scale multimodal analysis. Moreover, our empirical findings revealed a constantly evolving *multimodal alignment of gender stereotypes*.

In the following, we briefly introduced the domain of the study, then we surveyed recent work on applications of LLMs for discourse analysis, we then provided details on our data collection, annotation, and RAG pipeline, and finally we presented and discussed the results of the diachronic analysis.

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 102–111, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-009

1.1 Toy commercials as gender-based multimodal genres

Across five decades of research, TV advertisements targeted at children have consistently shown marked gender polarization (Verna, 1975; Johnson and Young, 2002; Kahlenberg and Hein, 2010; Marinelli et al., 2024). Differences between feminine-targeted, masculine, and mixed-audience commercials have been registered in sound (voices, background music and sound effects), language, transitions, and camera work, setting, interactions and activities, and colors. Stark gender polarization was observed in both multimodal emotion ratings and perceptual ratings of music in toy adverts (Marinelli et al., 2024). Specifically, masculine-targeted commercials were found to be significantly more aggressive and auditorily abrasive than feminine-targeted adverts.

Music can be imbued with distinct identity dimensions *upon which* ideological discourses are promulgated. Gender is one of these identity dimensions (Dibben, 2002) while androcentrism and heteronormativity are its hegemonic ideological discourses. *Multimodal genres*—which underlie this phenomenon in media portrayals—describe "regular patterns of semiotic choices in multimodal communicative objects and events that are particular to specific communities and cultures" (MODE, 2012). Toy commercials are organized in distinct gender-based multimodal genres. In this work, we explore how multimodal genres change over time, at the intersection of music and language.

2 Recent work on LLMs for discourse analysis

Historically, discourse analysis, including thematic coding, and other forms of qualitative analysis have been seen as domains reserved exclusively for human interpretation (DeJeu, 2025). The introduction of LLMs marks a significant development in qualitative research methodologies. LLMs are well-positioned to assist in qualitative coding, potentially augmenting early analysis, reducing the workload, and expanding the breadth of research corpora through semi-automated coding. Xiao et al. (2023) implemented in-context learning of a pre-compiled codebook with LLMs and achieved fair to substantial agreement with human coders; Gamieldien et al. (2023) employed LLMs to thematically code responses to a physics exam without an initial codebook; Bryda and Sadowski (2024) applied them on podcast interviews to semi-automate the creation of their codebook structure; and Yu et al. (2024) investigated the use of LLMs to automate pragma-discursive corpus annotation of apologies, reporting near to human-level accuracy, although stressing the importance of human oversight.

Curry et al. (2024) have reported negative results for the use of ChatGPT in replication studies that exemplified important tasks in corpus analysis, citing issues of repeatability and replicability tied to its non-deterministic nature (i.e., using its web-based interface). Garg et al. (2024) evaluated the use of LLMs for automated discourse coding in learning analytics on a dataset of questions and responses from secondary school teachers. Even though they obtained promising outcomes through fine-tuning, none met the reliability standards required in their field. Which highlights *the importance of reporting inter-coder reliability measurements*, as opposed to traditional classification metrics that do not account for chance agreement and can lead to the overestimation of performance.

Building from earlier observations of the application and limitations of the use of LLMs in analyzing discourse in corpora, Li and Wang (2024) proposed to improve the prompting method for LLM-based discourse analysis via contextual learning, output formatting, careful task description, and step by step procedure. They reported good results, showing that better prompting can overcome a number of previously described shortcomings.

Successful attempts have also been reported within critical discourse studies. Gao and Feng (2025) employed LLMs to analyze a Hong Kong news corpus and track media attitudes towards China, reporting performance on a par with trained coders. While Alonso del Barrio et al. (2024) showed promising results towards a semi-automated analysis of media content in the complex task of analyzing the framing of TV shows.

3 Methods

3.1 Data collection and annotation

On the 23rd of May 2025 we downloaded 4968 videos from the official YouTube channel of Smyths Toys Superstores, a major UK toy retailer. We paired this sample with another 5614 ads originally downloaded for (Marinelli et al., 2024). Merging these two datasets resulted in many duplicates, which were discarded. Then, only high-quality videos were selected, where ads without audio,

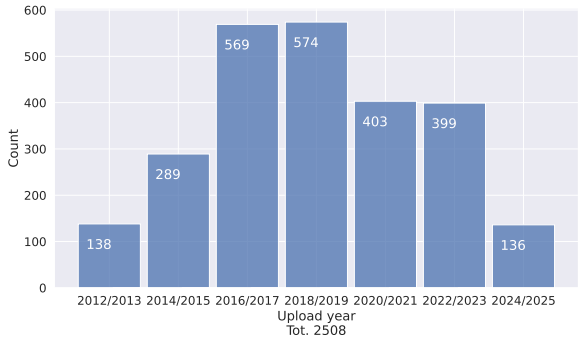


Figure 1: Distribution of the ads by year of upload.

formatted for mobile phones, or with substantial on-screen text were excluded. We then selected commercials where actors' faces were visible, to ensure the internal validity of the gender-targeting construct. Finally, we only kept those commercials that—once transcribed with Whisper large-v3 (Radford et al., 2023)—had transcripts longer than 40 characters. This resulted in **2508**, by and large unique, toy ads spanning a 14-year time frame that were analyzed in this study. Their distribution per *upload* year is shown in Figure 1. In all of the following analyses, the upload years were grouped together two-by-two, as some years had disproportionately less commercials than others.

Ground truth: The present analysis builds on a previous study on the role of music in gendered toy marketing (Marinelli et al., 2024). In the initial ground truth, 606 commercials were manually annotated in terms of their *gender orientation*—i.e., their intended target audience. This was determined by the gender of the presenters and accounted for token representations (Johnson and Young, 2002). In order to determine the inter-coder reliability for this variable, 15% of the commercials was double-coded by two coders independently, and a Krippendorff's alpha of .91 was obtained. Of the 606 commercials, 163 were coded as being targeted at a feminine audience, 149 as targeted at a masculine audience, 200 as targeted at a mixed audience, and 94 did not cast any actors. A pool of 152 musically trained participants rated the soundtracks of the commercials on music-focused scales: Electric/Acoustic, Distorted/Clear, Loud/Soft, Many/Few instruments, Heavy/Light, High/Low pitch, Punchy/Smooth, Wide/Narrow pitch variation, Harmonious/Disharmonious, Clear melody/No melody, Complex/Simple rhythm, Repetitive/Non-repetitive, Dense/Sparse, Fast/Slow tempo, and Strong/Weak beat. A different pool of 151 participants rated the commercials on seven emotion scales: Happy or Delightful, Amusing or Funny, Beauty or Liking, Calm or Relaxing, Energising or Invigorating, Angry or Aggressive, and Triumphant or Awe-inspiring.

In this study, based on insights from (Let Toys be Toys, 2021; de Iulio and Jarrin, 2004), we manually annotated a subset of the masculine, feminine and mixed ads in the ground truth that, once transcribed, contained more than 40 characters. This subset of 467 commercials (which constituted the RAG pool) was manually annotated in terms of the following themes: Domesticity and Nurturing, Fashion and Beauty, Nature and Animals, Love and Tenderness, Magic and Fantasy, Action and Adventure, Fight and Combat, Horror and Monsters, Speed and Racing, and Arts and Crafts.[1] Inter-coder reliability scores (Cohen's kappa) across 467 ads between a human coder (first author) and the RAG pipeline are provided at the end of section 3.3; where due to subpar reliability, three of the ten themes were excluded from the final diachronic analysis.

3.2 Automatic annotation of the soundtracks

We based our multitask transfer-learning model on a previous study performed on the same ground truth (Marinelli et al., 2023). However, differently from that study, we employed CLAP (Contrastive Language-Audio Pretraining) embeddings as audio representation (Elizalde et al., 2024). In addition, we turned the mid-level music-focused descriptors and the emotion scales into binary classifications by binning them on the 50% percentile, and we only kept those scales that performed well above chance in a 10-fold cross validation (i.e., scoring an average F1 above .60). The remaining emotion scales were Happy or Delightful, Beauty or Liking, Calm or Relaxing, and Angry or Aggressive; whilst the mid-level descriptors were Strong beat/Weak beat, Electric/Acoustic, Distorted/Clear, Loud/Soft, Heavy/Light, High pitch/Low pitch, Punchy/Smooth, Harmonious/Disharmonious, and Dense/Sparse. The resulting average F1 for the emotion scales was $.67 \pm .03$, whilst for the mid-level descriptors it was $.75 \pm .05$.

[1] Themes related to consumerism, competitions and sports, science and technology, and fun and play, were initially coded, but were later excluded due to their too broad applicability.

3.3 RAG pipeline

It is well-known that language models are few-shot learners (Brown et al., 2020; Schick and Schütze, 2021) as these models are able to achieve strong performance on many downstream NLP tasks without updating any parameters (i.e., without fine-tuning) by pairing their existing knowledge base and understanding of natural language with few examples or context directly within the prompt. Notably, Liu et al. (2022) found that retrieval-augmented generation (RAG) helps to considerably reduce hallucinations and stabilise performance, by coupling LLMs with an external pre-trained retriever model (Reimers and Gurevych, 2019). With this approach, the examples provided to the LLM are dynamically retrieved so that they only reflect the most relevant labels to the current data point, thereby making up for issues related to the short context windows. In particular, Milios et al. (2023) demonstrated that RAG with off-the-shelf retriever models can deliver robust performance in text classification tasks that involve many labels. In addition, with only 16 examples per class, the clever prompting proposed by Sun et al. (2023), Clue And Reasoning Prompting (CARP), achieved comparable performances, in text classification, to supervised models with 1,024 examples per class.

Themes detection: To perform the detection of the themes, we implemented what could be called a "reverse CARP" technique (see Listing 1) where the model is presented with negative and positive examples of transcripts that are either unrelated or related to the theme under analysis, in order to give it access to patterns and cues associated with the theme. Both negative and positive examples are retrieved via the retriever model (cosine similarity) and each of the positive examples is presented with a list of cues related to the theme under analysis. The model is then given a specific theme definition, the current datapoint transcript, and is asked to determine if cues related to the theme are present. The model is then asked to provide a brief reasoning paragraph, and to return a list of cues related to the theme (or an empty list otherwise). Then, any hallucinated cue not present in the transcript is automatically removed. Finally, a theme is deemed present when the system returned a non-empty list.

Target classification: The classification between feminine, mixed audiences, and masculine-targeted commercials was broken down in five sub-tasks:

Model	Average F1 across themes (1, 5, 10 examples)		
CohereForAI/c4ai-command-r7b	82 ± 05	**82 ± 04**	81 ± 06
microsoft/Phi-3-small-8k-instruct	80 ± 07	79 ± 07	79 ± 07
microsoft/Phi-3.5-mini-instruct	81 ± 05	81 ± 06	**82 ± 04**
allenai/Llama-3.1-Tulu-3.1-8B	77 ± 07	77 ± 08	75 ± 08
google/gemma-2-9b-it	81 ± 05	81 ± 06	80 ± 06

Table 1: Preliminary evaluation of the themes detection. The reported deviations are computed across themes.

one vote from each binary subset of the classes (i.e., feminine/masculine, feminine/mixed, mixed/masculine), one vote from the full classification (where the LLMs were presented with all three classes), and one tie-breaking vote from the music-focused model. This is justified under the assumption that the models would be more reliable when comparing only two classes at a time. Once collected all votes, the hard-coded decision logic would either choose the most frequent class, or in case of a tie, choose the final class from the corresponding sub-task.

The prompting structure employed for each sub-task is shown in Listing 2. First the model is presented with a set of examples that illustrate the different possible classes, each with their corresponding themes and transcripts. The model is then instructed to analyze the themes (automatically collected at the previous step) and transcript of the current datapoint, taking into account potential gender stereotypes to determine the target audience. The model is then asked to provide a reasoning paragraph that explains its decision, following a given structure that highlights the relevant themes and their relationship to the target audience. Finally, the model is prompted to return the inferred class, choosing depending on the sub-task, from a predefined set of possible values.

Preliminary evaluation: All experiments were run on a single NVIDIA A5000 GPU. Different language models between 4 and 9 billion parameters were evaluated on the RAG pool consisting of 467 documents. Considering that the music model was also being trained on the 606 ground truth commercials (thus including data from the RAG pool), we needed to implement a cross-validation algorithm within which we positioned the RAG pipeline. This consisted in excluding the data points of the test set (i.e, fold) from the available RAG pool at each iteration. Which means, that whilst we performed a 10-fold cross-validation for the music model—which on target classification achieved an F1 of .70 ± .05—the RAG pipeline was instead effectively evaluated as a leave-one-out cross-validation,

Model	Full logic (1, 5, 10 examples)			Without music model (1, 5, 10 examples)			3-classes sub-task (1, 5, 10 examples)		
CohereForAI/ c4ai-command-r7b	71 [62, 79]	71 [62, 79]	71 [63, 79]	64 [55, 73]	66 [57, 74]	68 [59, 76]	67 [58, 75]	64 [55, 73]	64 [55, 73]
microsoft/ Phi-3-small-8k-instruct	73 [65, 81]	68 [58, 76]	71 [63, 79]	65 [55, 73]	61 [51, 69]	62 [53, 71]	73 [64, 80]	63 [54, 72]	66 [56, 74]
microsoft/ Phi-3.5-mini-instruct	76 [68, 83]	77 [69, 84]	73 [65, 81]	71 [63, 79]	71 [62, 79]	68 [59, 76]	70 [62, 78]	72 [63, 80]	68 [60, 76]
allenai/ Llama-3.1-Tulu-3.1-8B	77 [69, 84]	78 [70, 85]	77 [69, 85]	75 [67, 83]	74 [66, 81]	74 [65, 82]	75 [66, 82]	75 [66, 82]	77 [69, 84]
google/ gemma-2-9b-it	77 [69, 84]	79 [71, 86]	78 [70, 85]	75 [67, 82]	78 [70, 85]	76 [68, 84]	74 [66, 82]	75 [67, 82]	73 [65, 80]
c4ai (theme) + gemma-2 (target)	//	**80 [72, 87]**	//	//	78 [70, 85]	//	//	75 [66, 82]	//

Table 2: Preliminary evaluation of the target classification, F1 scores with corresponding 95% CI (bootstrapped).

with a slightly reduced RAG pool at each fold.

For the sake of brevity, the results of the theme evaluation are reported as averages across themes in Table 1. The best-performing models were, on a par, Microsoft's Phi-3.5-mini-instruct with 10 examples (effectively, 10 negative and 10 positive examples), and Cohere's c4ai-command-r7b with 5 negative and 5 positive examples. Generally, most models performed similarly well on this task.

The results of the target classification are instead reported in Table 2, where given the higher level of abstraction of this task, we also performed bootstrapping (with replacement, 10k iterations) to provide the corresponding 95% confidence intervals. The observed performance peak at 5 examples *per class* may be due to the limited size of the RAG pool, with more examples leading to an increased proportion of non-relevant examples being presented to the model. Microsoft's Phi-3.5-mini, AllenAI's Llama-Tulu, and Google's Gemma performed similarly well, with Gemma coming out slightly ahead. As our last evaluation, we used Cohere's model to detect the themes, which then were provided as context to Gemma. This resulted in the best-performance, with an F1 score of .80 [.72, .87]. It is with this final combination that we proceeded to analyze the larger corpus of toys commercials.

Listing 1: Themes detection prompt (shortened).

```
# Negative examples: unrelated to {current_theme}
{negative_examples}

# Positive examples: related to {current_theme}
{positive_examples}

# Theme definition
Examples of {current_theme} contain cues referring
to {current_theme_definition}.

# Current datapoint
transcript: {current_transcript}

# INSTRUCTIONS
Based on the theme definition and on the examples
determine if the current datapoint contains cues
about {current_theme}. First, provide a reasoning
paragraph, then return the list of cues. If no
relevant cues are found return an empty list.
```

Listing 2: Target classification prompt (shortened).

```
# EXAMPLES (grouped by class)
{examples}

# Current datapoint
transcript: {current_transcript}
themes: {current_themes}

# Definitions of the collected themes
{current_themes_definitions}

# INSTRUCTIONS
Based on the previous examples, determine the target
audience for the current datapoint. First, choose
only one of the following reasoning structures.

In the current transcript, the themes <theme1>,
<theme2> .. are mainly associated with femininity
so the the target of the toy ad is 'Girls/women'.

or

In the current transcript, the themes <theme1>,
<theme2> .. are mainly associated with masculinity
so the the target of the toy ad is 'Boys/men'.

or ...

Finally, return the above determined value.
Choose only one from: {current_sub-task_classes}.
```

Inter-coder reliability: Before proceeding any further we provided the Cohen's kappa and related 95% confidence intervals (with replacement, 10k repetitions) for each of the themes and for the classification of the gender-based target audience. Due to space constraints, we only focus on the best-performing combination highlighted in Table 1. The theme Domesticity and Nurturing achieved a κ of .72 [.52, .89]; Fashion and Beauty obtained a κ of .76 [.57, .91]; Nature and Animals .67 [.50, .81]; Love and Tenderness .65 [.46, .81]; Magic and Fantasy .77 [.59, .91]; Action and Adventure .45 [.30, .60]; Fight and Combat .71 [.49, .89]; Speed and Racing .78 [.62, .91]; Horror and Monsters .54 [0, 1] (due to numerical errors during bootstrapping); Arts and Crafts .57 [.36, .75]. Considering the traditional threshold of .61—where a score between .61 and .8 indicates *substantial* agreement (Landis and Koch, 1977, p. 165)—we excluded the following themes from the analysis of the results:

Action and Adventure, Horror and Monsters, and Arts and Crafts. Finally, the main logic of the best-performing combination of models achieved a κ of .69 [.57, .79] on the gender target classification.

4 Results

Besides the 467 commercials from the ground truth, 2041 unseen ads were automatically annotated with the previously described pipeline. In the following we reported the results of analyses performed on the hybrid dataset of 2508 commercials.

Diachronic analysis: Spearman's rank correlation coefficients were computed between upload years and the *ratio of positive predictions*. For the themes this number is simply the ratio of commercials where a theme is predicted as present in the transcript. Concerning unipolar affective scales, positive predictions corresponded to commercials with a soundtrack that was predicted as belonging to the upper 50% percentile, while for bipolar mid-level descriptors (e.g. Electric/Acoustic) positive predictions corresponded to the rightmost polarities of the scales, which were binned in the ground truth as the upper 50% percentile.

A minimum count of 20 commercials per bin (at the intersection of control variable and upload year) was imposed. In one case, at the intersection of the theme Love and Tenderness and the years 2024/2025, we dropped said bin from the analysis, as it contained less than 20 commercials, and therefore representativeness could not be guaranteed.

First, no relevant trends were detected across gender targets—that is, across the entire corpus. However, once grouped by predicted gender target, two relevant trends emerged for the predicted themes. Within commercials that were predicted as targeted at a mixed audience, the theme Nature and Animals has been almost steadily decreasing over the last 14 years (Spearman's $\rho = -.86$, $p = .014$), as reported in Figure 2a. Similarly, although to a lesser extent, as reported in Figure 2b, the theme Fight and Combat has been steadily decreasing within commercials predicted as targeted at a masculine audience ($\rho = -.93$, $p = .003$). No trends were detected for any theme within commercials targeted at a feminine audience.

No relevant trends were also detected for emotions within gender targets. However, as reported in Figure 3a, commercials classified as targeted to a mixed audience show a negative trend in the ratio of soundtracks with a weak beat ($\rho = -.82$, $p = .023$). Similarly, in Figure 3b, for commercials classified as targeted to a masculine audience, the ratio of those with a weak beat or with a soft soundtrack has been steadily decreasing (both with $\rho = -.93$, $p = .003$).

Theme	ρ	p	Scale
fight_combat	-.93	.003	Strong beat/Weak beat
	-.89	.007	Loud/Soft
	-.79	.036	Punchy/Smooth
love_tenderness	-.83	.042	Strong beat/Weak beat
	-.89	.019	Electric/Acoustic
magic_fantasy	-.79	.036	Punchy/Smooth
speed_racing	-.86	.014	Loud/Soft
	-.93	.003	Punchy/Smooth
	-.86	.014	Electric/Acoustic
	-.86	.014	Distorted/Clear
	-.79	.036	Strong beat/Weak beat

Table 3: Spearman's coefficients and p-values of themes-wise time trends of the mid-level music descriptors.

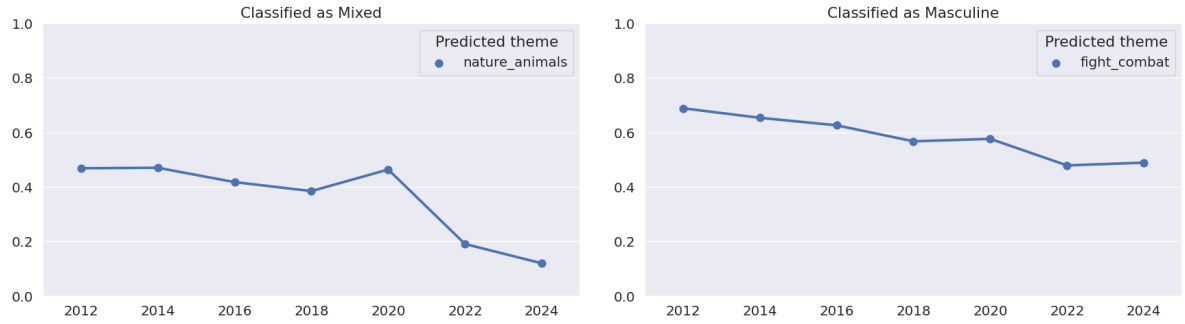
When grouped by theme, no relevant trend was detected in any emotion scale. Instead, several significant correlations emerged from the analysis of mid-level music descriptors within commercials grouped by theme. Which were reported in Figures 4a to 4d and in Table 3, with the corresponding correlation coefficients and p-values.

Interactions between scales and themes: Finally, we explore the interaction between emotions and mid-level descriptors predicted from the soundtracks, and the themes found in spoken language that were annotated with the RAG pipeline. In Figure 5 we reported the ratio of soundtracks—grouped by theme—that were predicted in the upper 50% percentile of each scale.

A clear pattern emerged, where themes stereotypically associated with femininity (de Iulio and Jarrin, 2004) co-occurred with soundtracks that were softer, happier, calmer, more harmonious, lighter, and—with the exception of the theme Fashion and Beauty—with weaker beats, smoother and more acoustic than electric; where such exception is likely a reference to the electronic dance music that accompanies fashion shows. Conversely, themes associated with masculinity were mostly paired with angrier, louder, heavier, more distorted, [...], and more disharmonious soundtracks.

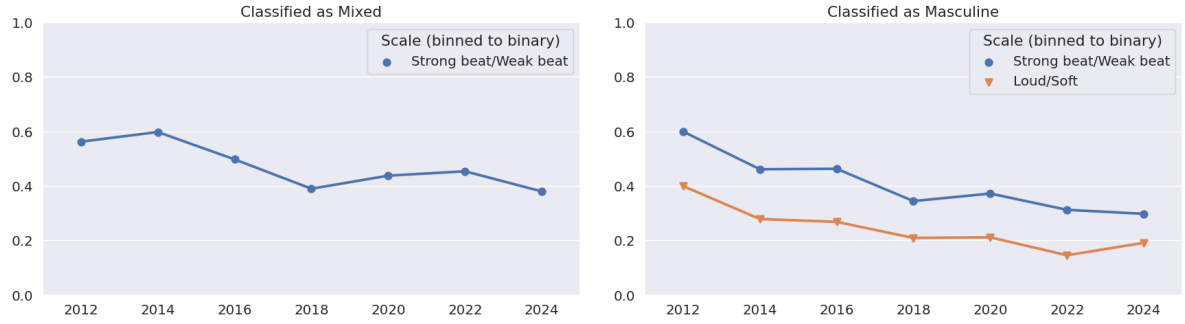
5 Discussion

As evidenced in the results, and as discussed in a previous study (Marinelli et al., 2024), the mid-level music descriptors in the ground truth are collinear. Therefore, trends related to those scales are better interpreted along two latent axes: the



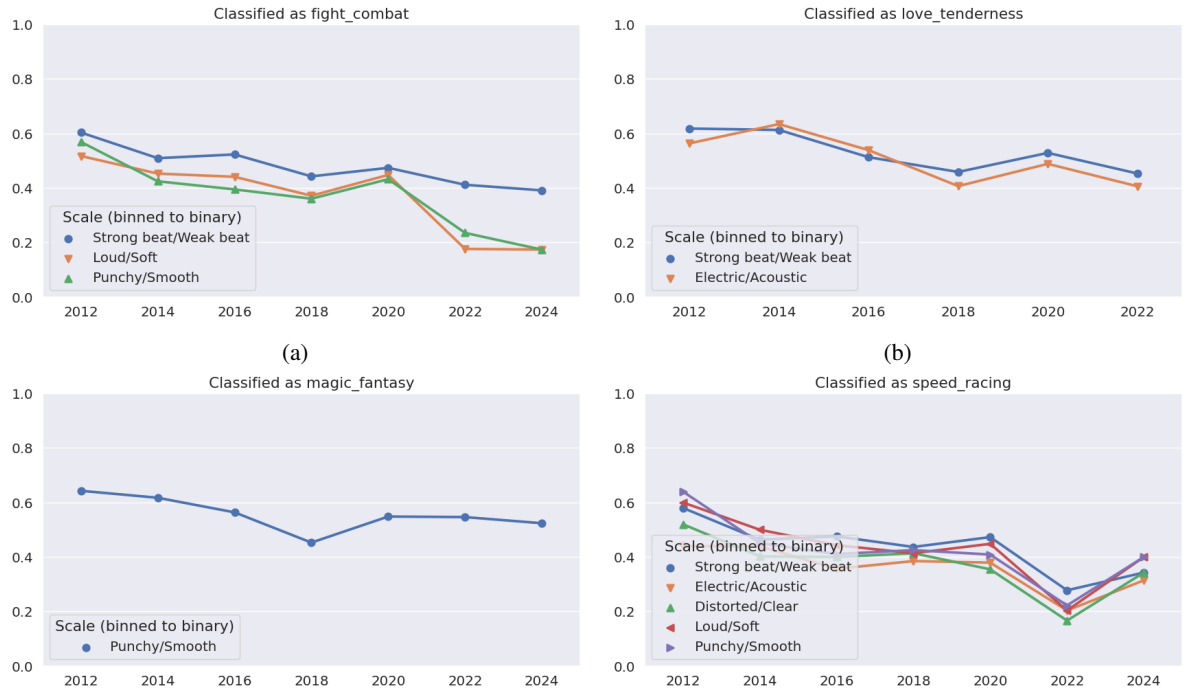
(a) (b)

Figure 2: Ratio of commercials grouped by gender target that contain the predicted theme.



(a) (b)

Figure 3: Ratio of commercials grouped by gender target in the upper 50% percentile of the scales.



(a) (b)

(c) (d)

Figure 4: Ratio of commercials grouped by theme in the upper 50% percentile of the scales.

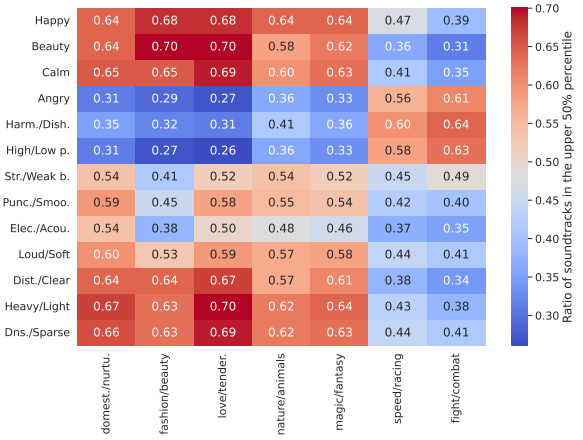


Figure 5: Heatmap showing the ratio of soundtracks grouped by theme (columns) predicted in the upper 50% percentile of each scale (rows).

"Energizing/Soothing" axis and the "Harmonious and clear/Dissonant and distorted" axis.

Concerning the reported diachronic changes, a counter-stereotypical trend was found in the decreasing ratio of masculine-targeted commercials referring to Fight and Combat (Figure 2b). However, this change appeared to be offset by an overall increase of loud soundtracks with a strong beat in masculine-targeted ads (Figure 3b), also reported for mixed-audience commercials (Figure 3a). In particular, ads referring to Fight and Combat (Figure 4a) show a substantial increase in energizing soundtracks. A similar trend appeared also for the theme Love and Tenderness (4b) although to a lesser degree. Moreover, ads referring to Speed and Racing (Figure 4d) show the most pronounced increase in both energizing and dissonant/distorted soundtracks. Therefore, it appears that masculine-adjacent commercials have become more abrasive over the last 14 years, with little to no change to report for feminine-targeted ones.

Tracking these patterns showed that, despite widespread public backlash (Fine and Rush, 2018; Marinelli et al., 2024), toy commercials continue to reinforce traditional gender stereotypes through increasingly polarized multimodal strategies. Specifically, the finding that masculine-adjacent commercials have become more auditorily aggressive, while feminine ones remain largely unchanged, suggests that gender polarization may be widening rather than narrowing. Moreover, diachronic analysis can reveal counter-intuitive trends, such as the decrease of Fight and Combat being offset by increasingly aggressive soundtracks, showing how stereotyped media portrayals can evolve in form while maintaining their underlying polarization.

6 Conclusion

Our best-performing configuration achieved substantial agreement with a human annotator. Specifically, it achieved an average Cohen's κ of .66 across ten themes[2] (average F1 of .82) and a κ of .69 for gender-target classification (F1 of .80) in a leave-one-out cross-validation on 467 ads. The pipeline was then applied on 2041 unseen commercials, for a total of 2508 toy ads spanning 14 years. Evolving patterns emerged in the interactions between language and music in gendered toy advertisements. Our findings reveal a *multimodal alignment of gender stereotypes*, where stereotypically feminine themes in the transcripts co-occur with soft, calm, harmonious soundtracks, while stereotypically masculine themes consistently align with loud, aggressive, and distorted soundtracks. An overall increase of energizing soundtracks was reported for masculine-targeted, mixed-audience commercials, their themes, and even for one traditionally feminine theme (Love and Tenderness). However, such increase was steeper for masculine-targeted commercials and associated themes.

This study had a few limitations. In the ground truth, the gender orientation of the commercials was based on the gender of the actors in each video. However, the pipeline based its decisions on patterns in text and audio which can change over time, undermining the internal validity of this particular inference. Another limitation, which we partially addressed by providing confidence intervals for the performance metrics, is that the RAG pipeline was built using the same set on which it was evaluated.

Future studies might consider testing the pipeline across different cultures, as portrayals may vary between countries. Moreover, adding visual analysis such as colors, character positions, and facial expressions would provide a complete picture of how text, audio, and image cooperate in gendered advertising. Finally, preregistration and other strategies should be considered to ensure that a portion of unseen data is also manually coded by experts and only disclosed to system developers at evaluation time. This is especially relevant when automated annotation is employed on corpora that are orders of magnitude larger than the initial ground truth.

[2]However, three subpar themes were excluded from the final analysis. The remaining themes averaged at a κ of .72.

Acknowledgments

This work was supported by UK Research and Innovation [grant number EP/S022694/1].

Iacopo Ghinassi worked on this study whilst still affiliated to Queen Mary University of London.

References

David Alonso del Barrio, Max Tiel, and Daniel Gatica-Perez. 2024. Human interest or conflict? leveraging llms for automated framing analysis in tv shows. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*, IMX '24, page 157–167, New York, NY, USA. Association for Computing Machinery.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Grzegorz Bryda and Damian Sadowski. 2024. From words to themes: Ai-powered qualitative data coding and analysis. In *Computer Supported Qualitative Research*, pages 309–345, Cham. Springer Nature Switzerland.

Niall Curry, Paul Baker, and Gavin Brookes. 2024. Generative ai for corpus approaches to discourse studies: A critical evaluation of chatgpt. *Applied Corpus Linguistics*, 4(1):100082.

Emily Barrow DeJeu. 2025. Can (and should) llms perform critical discourse analysis? *Journal of Multicultural Discourses*, pages 1–8.

Nicola Dibben. 2002. Gender identity and music. In *Musical Identities*. Oxford University Press.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340.

Cordelia Fine and Emma Rush. 2018. "Why does all the girls have to buy pink stuff?" The ethics and science of the gendered toy marketing debate. *Journal of Business Ethics*, 149(4):769–784.

Yasir Gamieldien, Jennifer M Case, and Andrew Katz. 2023. Advancing qualitative analysis: An exploration of the potential of generative ai and nlp in thematic coding. *Available at SSRN 4487768*.

Qingyu Gao and Dezheng (William) Feng. 2025. Deploying large language models for discourse studies: An exploration of automated analysis of media attitudes. *PLOS ONE*, 20(1):1–17.

Ryan Garg, Jaeyoung Han, Yixin Cheng, Zheng Fang, and Zachari Swiecki. 2024. Automated discourse analysis via generative artificial intelligence. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, page 814–820, New York, NY, USA. Association for Computing Machinery.

Simona de Iulio and Zouha Jarrin. 2004. Toy commercials across Europe. *Young Consumers*, 5(4):39–45.

Fern Johnson and Karren Young. 2002. Gendered voices in children's television advertising. *Critical Studies in Media Communication*.

Susan G Kahlenberg and Michelle M Hein. 2010. Progression on nickelodeon? gender-role stereotypes in toy commercials. *Sex roles*, 62(11):830–847.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

Let Toys be Toys. 2021. Who gets to play now? – New research on TV toy ads. Accessed: 2025-19-06.

Bingru Li and Han Wang. 2024. Tacomore: Leveraging the potential of llms in corpus-based discourse analysis with prompt engineering.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Luca Marinelli, György Fazekas, and Charis Saitis. 2023. Gender-Coded Sound: Analysing the Gendering of Music in Toy Commercials via Multi-Task Learning. *24th International Society for Music Information Retrieval Conference (ISMIR 2023)*.

Luca Marinelli, Petra Lucht, and Charalampos Saitis. 2024. A Multimodal Understanding of the Role of Sound and Music in Gendered Toy Marketing. *PLOS ONE*, 19(11):1–32.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, Singapore. Association for Computational Linguistics.

MODE. 2012. Glossary of multimodal terms. Edited by Jewitt, Carey and Bateman, John. Accessed: 2025-19-06 at `https://multimodalityglossary.wordpress.com/genre/`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Mary Ellen Verna. 1975. The female image in children's tv commercials. *Journal of Broadcasting & Electronic Media*, 19(3):301–309.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, page 75–78, New York, NY, USA. Association for Computing Machinery.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, 29(4):534–561.