Eurographics Workshop on Visual Computing for Biology and Medicine (2024)
L. Garrison and D. Jönsson (Editors)

Short Paper

Leaving the Lab Setting: What We Can Learn About the Perception of Narrative Medical Visualizations from YouTube Comments

Sarah Mittenentzwei[1], Danish Murad[1], Bernhard Preim[1], Monique Meuschke[1]

[1]Department of Simulation and Graphics, University of Magdeburg, Germany

Abstract

The general public is highly interested in medical information, particularly educational media about diseases, healthy biological processes such as pregnancy, and surgical procedures. Efforts to develop educational materials using data-driven approaches like narrative visualization exist, but studies are often performed in lab settings. Since there are few public sources for visualizations of medical image data, YouTube videos, which often contain 3D medical visualizations, are an important reference. We aim to better understand the user base of these videos. Therefore, we curated a dataset of 76 videos featuring medical 3D visualizations. We analyzed 14,550 comments across all videos using manual review and machine learning techniques, including natural language processing for sentiment and emotion analysis of user comments. While few comments directly link visual attributes or design choices to user sentiment, insights into users' motivation and opinions of specific design choices have emerged.

CCS Concepts

• ***Applied computing*** → ***Life and medical sciences;*** • ***Information systems*** → *Web mining;*

1. Introduction

For narrative visualization targeted at explaining medical concepts to a lay audience, analysis of existing visualizations provides insights into the design space and the frequency of certain design decisions. Although medical image data typically depict anatomical structures in 3D space and 3D models are increasingly used in anatomy education [AGAZG23], explanatory 3D medical visualizations for laypersons are rare. We are investigating YouTube videos that contain medical 3D visualizations since they can be found in larger quantities. Analyzing YouTube videos allows us to analyze a large amount of user feedback in the form of views, replay rates, and comments outside of artificial lab settings. Particularly in light of the COVID-19 pandemic, there has been a transformation in the educational landscape. More and more people have increasingly turned to the internet, social media, and YouTube for health-related content [WJJ17]. Our analysis strives to investigate how health education materials, similar to the narrative medical visualizations produced by the research community, are perceived by users.

We analyze 76 YouTube videos featuring 3D visualizations across diverse medical topics, spanning diseases, procedures, and prevention from five medical YouTube channels. Leveraging Natural Language Processing (NLP) techniques, we conduct a comprehensive analysis of user comments, both qualitatively and quantitatively. Our research integrates NLP with the examination of video content. By analyzing the comments using state-of-the-art NLP techniques to analyze user sentiment and emotion towards each video, we aim to derive insights about user motivation and preferences.

2. Related Work

We review research on 3D visualization in education, specifically medical content on YouTube and video analysis. We also discuss foundational NLP work for analyzing social media comments.

2.1. Analysis of YouTube Videos in Health Education

3D medical visualizations enhance understanding and interactive learning [AGAZG23] but necessitate expert and animator involvement. Narrative medical visualization focuses on making medical data engaging and accessible [GMPB23], with YouTube serving as a key platform for enhancing health literacy among diverse audiences [LTX*22]. However, due to the risk of misinformation from non-expert creators [LL19], our analysis will exclusively focus on videos from health companies, excluding private individuals. Park & Goering [PG16] found that non-health professionals use YouTube for health reasons due to its convenience and empowering effects, which enhance users' healthcare competence and learning.

Burgess & Green [JJ20] analyzed YouTube's transformation into a public communication tool, highlighting its democratizing impact, creativity, accessibility, and community due to open access. Welbourne & Grant [WG16] examined factors affecting the popularity of science and health channels, noting user-generated content often outperforms professional content in views, with shorter videos being preferred for science communication [YBSX22]. Additionally, Amini et al. [AHRL*15] identified design patterns for data videos, and Shi et al. [SLL*21] defined a design space for such videos.

DOI: 10.2312/vcbm.20241193

2.2. Natural Language Processing

We use NLP techniques to analyze user-generated content on YouTube, deriving insights from the extensive data. Sentiment analysis evaluates emotions in text, with Pang and Lee [PL08] demonstrating its effectiveness in extracting opinions. The *transformer model* by Vaswani et al. [VSP*17] marked a significant advancement, as its attention mechanism captures long-term dependencies and enables parallel processing. Radford et al.'s [RNSS18] GPT model, focusing on sequential text generation, and Devlin et al.'s [DCLT19] BERT model, excelling in context understanding, further pushed the boundaries of NLP. Alaparthi & Mishra [AM21] compared BERT to previous models, highlighting its superiority in sentiment classification. Chiorrini et al. [CDMP21] confirmed BERT's accuracy in emotion and sentiment analysis. Liu et al. [LOG*19] improved upon BERT with the RoBERTa model, achieving better performance across various tasks. These recent developments highlight RoBERTa's potential for analyzing large text datasets, like our collection of YouTube comments.

Other work proves how state-of-the-art NLP techniques can be used to analyze YouTube comments. Porreca et al.'s study [PSN20] highlighted a sentiment shift in comments on Italian vaccination videos post-campaign, while Bozkurt and Aras [BA20] found mixed sentiments in comments on videos about cleft lip and palate disease, with negative comments often detailing surgical pain and public embarrassment.

3. Video Selection

We included four aspects in our analysis: (1) number of views, (2) number of comments, (3) content of comments, and (4) replay rates. Our search was conducted using the keywords "3D medical animation" and "3D medical visualization". YouTube allows filtering based on relevance (default), view count, upload date, and rating. Upload date and rating filters were found to produce less useful results, mainly showing irrelevant latest videos due to the lack of user engagement. In 2022, YouTube introduced a new feature called *replay graph* (see Fig. 1) which highlights the segments of a video that viewers replay most frequently. It is depicted as a semi-transparent area chart right above the video progress bar. However, not every video is accompanied by a replay graph. In addition, few details about the replay graph are public, so it is not clear what criteria a video must meet to be equipped with the replay graph. The collection of videos was narrowed down based on these criteria:

- The videos must show 3D visualizations of human anatomy.
- The replay graph is present in the video.
- The channels must be credible.
- The videos must be in the English language.
- The videos need to also have comments in the English language.

Efforts were made to select a diverse range of topics, anatomic structures, length of video, and range of upload period. Credibility was checked by determining whether the channels were managed by companies with expertise in health communication. Finally, a total of 76 videos from the following five channels were selected: *Nucleus Medical Media* (47 videos), *Scientific Animations* (9 videos), *Dandelion Medical Animation* (15 videos), *BioDigital, Inc.* (4 videos), and *Infuse Medical* (1 video).

4. Data Collection and Preprocessing

We used YouTube's API to extract video titles, channel names, view counts, upload dates, durations, and top-level comments (excluding replies). Since the number of likes lacks context without dislikes, which were unavailable, we did not analyze them. We ensured privacy by not collecting usernames or any personally identifiable information. Replay graphs were extracted as images from the site's HTML. These graphs do not show exact replay counts but indicate relative replay frequencies across different parts of a video, with peaks denoting high replay numbers. All videos were manually reviewed to correlate user responses with video content.

Preprocessing of Comments. Once collected, the textual parts of the comments underwent two essential preprocessing steps for any NLP system, as the characters, words, and sentences identified are crucial for all subsequent processing stages. First, informalities like spelling errors, slang, new words, URLs, special characters, and emoticons were filtered out using the Python libraries *demoji* and *regular expression (RegEx)*. Second, we removed non-English comments using a fine-tuned RoBERTa-based model for language detection that has an average accuracy of 99.6 % over 20 languages [Pap23]. Following these preprocessing steps, the final dataset consists of 14,550 comments across 76 videos.

Sentiment Analysis. For sentiment analysis, we use the "Twitter-RoBERTa-base for Sentiment Analysis" model [LBN*22], the latest model trained on the concise, informal nature of social media posts. This model is relevant as Twitter's limit of 280 characters per post encourages unstructured informal content similar to YouTube comments [SOA*23]. The chosen model is pre-trained on 124 million tweets from January 2018 to December 2021 and fine-tuned for sentiment analysis on social media. The model classified each comment as positive, neutral, or negative, along with the respective sentiment score that encodes the probability for a comment to contain the chosen sentiment.

Emotion Analysis. Emotion detection offers a more granular insight into the feelings that users might be experiencing. A pre-trained RoBERTa-based model fine-tuned on a dataset containing emotions was chosen [Low23]. The dataset contained 28 different emotions, which resulted in a multi-label classification for the model output. This way we are able to identify a range of emotions from YouTube comments, such as joy, approval, disapproval, anger, and disappointment. As output, each comment was assigned an emotion and a respective emotion score.

5. Video Analysis

We discuss our analysis and insights gained. The comments were analyzed regarding sentiment and emotion. Additionally, individual comments directly referring to the video's content are discussed. The number of views and comments per video has a moderate positive correlation (0.65), except for one video that had an exceptionally high number of views. This outlier video about COVID-19 had about 339 million views but only 836 comments, compared to other videos with fewer views that had more than 1000 comments. However, the number of comments is not a reliable metric because it can be influenced by the channel owner by filtering comments before they are posted and deleting comments that have already been posted.

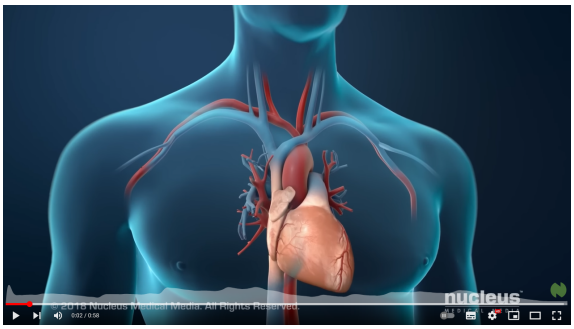

Figure 1: *Screenshot from the video:* What is a Coronary Angioplasty? *(Nucleus Medical Media) showing the replay graph.*

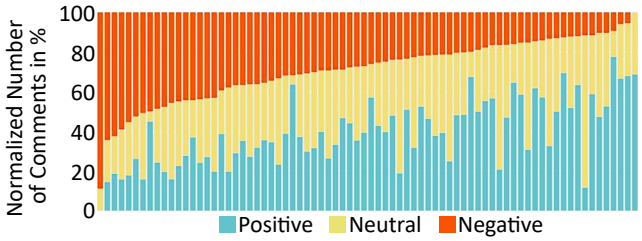

Figure 2: *Normalized distribution among all the videos, in ascending order regarding the number of negative comments.*

The replay graphs vary greatly not only in terms of where peaks are visible in relation to the video content but also in their overall shape. While some show few clear peaks, others show an almost smooth curve, while still others show a zigzag pattern of multiple small high-frequency peaks. It is possible that replay graphs act as a self-fulfilling prophecy; that is, as peaks become visible, more users tend to jump to those time stamps, making the peaks larger and more prominent to other users. As a result, the shape of the replay graph may not be fully based on the content of the videos and, therefore, does not provide insight into the content of the videos. Furthermore, we did not find any connection between the video content with the highest replay rates and the content of the comments.

Sentiment Analysis. Overall, the sentiment distribution was fairly even, with positive comments slightly edging the numbers. The overall sentiment distribution shows that most comments were positive (35.92 %), followed by neutral comments (30.46 %) and negative comments (33.62 %). The following observations were made through a manual investigation:

- **Positive Sentiment (35.92 %):** These viewers had a favorable or supportive response to the videos. Appreciation for the content, agreement with the information presented, or positive experiences related to the video topics.
- **Neutral Sentiment (30.46 %):** Comments with a neutral sentiment were more likely informational, seeking clarity, or sharing personal experiences. They also included queries about the video topic, requests for more information or videos, or discussions that do not express strong positive or negative feelings.
- **Negative Sentiment (33.62 %):** These comments usually expressed concerns, disagreements, or negative personal experiences. Negative sentiments could arise from viewers who have faced challenges, side effects, or unfavorable outcomes related to the video topics.

While the majority of the videos had less than 200 comments, it was important to get a normalized view to find patterns or videos that needed a more in-depth analysis, see Figure 2. To cover the extremes of the distribution, we look at the ten videos with the most negative and the ten videos with the most positive comments.

Ten Videos with Highest Proportion of Negative Comments. It was not feasible to manually examine thousands of comments. A keyword-based method was employed to sift through comments, focusing on those that mentioned words or phrases related to feedback about the video's presentation, style, quality, clarity, and similar aspects. These included "video," "visualization," "animation," "graphics," "quality," "audio," "voice," "presentation," "background," and "sound." After filtering for these keywords, a combination of emotions presented and a manual reading approach was undertaken to gain insights into viewers' perceptions of the videos:

- Viewers generally praised the videos for being helpful, improving their knowledge, and being informative, simple, and easy to understand. Here are a few example comments:
  - "*Your videos are really helpful. Well explained in a simple and uncomplicated way. You made learning enjoyable.*"
  - "*This really broke it down & made it so easy to understand! Thank you. The visual is great.*"
- Viewers generally felt negative about robotic, monotone, or computer voiceovers. Some viewers also expressed concerns about emotional or health triggers that might be present in the video. Some comments mentioning them are below:
  - "*Pleaseee use a different voice for these videos. People who are concerned about their health do not want a cold robotic voice adding to their anxiety.*"
  - "*just an FYI, you should include a trigger warning before showing visuals of an aura. It can cause migraines for those who get them.*"

Ten Videos with Highest Proportion of Positive Comments. The majority of these videos had less than 50 comments, one with 116 and another video called "From Fertilization to Childbirth" with 1113 comments as a complete outlier. The user reactions were naturally skewed by the sheer number of comments in this video.

Similar themes as the videos with predominantly negative sentiment were found here as well. Viewers generally praised the videos and animations for being helpful, improving their knowledge, being informative, and thanking the channel. The negative comments generally revolved around viewers' personal experiences or general comments regarding the topic of the videos. While there were some comments that talked about an individual video not showing symptoms (*"What are the symptoms for this"*) or the choice of having music in one of the videos (*"WHOLLY unnecessary music....."*), these comments were few and far between. It is difficult to say with certainty if it directly led to low or high user engagement or users liking or disliking the video.

Rest of the Videos. After analyzing the extremes, the other 56 videos had a fairly even distribution of sentiments. The presence of neutral comments suggested that some viewers might have questions, suggestions, or general observations that are neither strongly positive nor negative. These videos also showed similar content in terms of user comments after a combination of sentiment and emotion as well as keyword analysis. Positive comments praised them for being informative, increasing their understanding, and thanking the channels. Some negative and neutral comments were observed:

- Viewers' comments on the use of music varied, often within the same video, highlighting the subjective nature of preferences and the complexity of decision-making in design.
  - "*umm the music is really annoying..... can't concentrate*"
  - "*Thanks for the video. Very well done and explained. Music is very relaxing also.*"
- The use of plain or dark background was questioned: "*Why are medical 3D animations always scary like there is something about the plain dark background, or the lack of sound apart from the voice of the narrator, idk*". While this gives an insight into how backgrounds might induce certain emotions in the viewer, the lack of other comments about backgrounds in other videos made it hard to substantiate it as a general design goal.
- Again, the use of natural, and relaxing human voice was praised, as opposed to computer-generated or monotonous voiceovers.
  - "*Does anyone else find these videos very relaxing? Aside from being fascinating, I just feel at home amongst these animations and calm voice and neutral colours.*"
- Some viewers questioned the pacing of videos:
  - "*I set the speed to 1.25x. This video is a bit slow otherwise...*"
  - "*Would've been great if the narrator explained a bit slow for non-medical people like me.*"

Emotion Analysis. The comments for each emotion were manually inspected, with observations focused on the top five emotions, as others appeared in less than 5 % of all comments:

- **Neutral (34.37 %):** These comments were mostly discussing the topic, seeking answers, or sharing experiences without a strong emotional expression. This reaffirms our earlier observation from sentiment analysis that many comments that are informational, inquisitive, or descriptive in nature do not show strong emotions.
- **Admiration (14.54 %):** Comments expressing appreciation, respect, and positive regard for the content, or the people going through experiences described in the videos.
- **Gratitude (9.13 %):** Many viewers expressed gratitude, thankful for the information provided, the clarity of the presentation, or the insights gained from watching the videos.
- **Sadness (7.99 %):** Comments with sadness were sharing challenges, personal losses, or empathizing with others.
- **Curiosity (7.57 %):** Comments expressing curiosity suggested that viewers are keen to learn more, have questions, or are intrigued by the content.

Sentiment and emotion were used to narrow down interesting videos and comments. For example, emotions such as confusion, annoyance, anger, or disapproval were more likely to provide insight into what people did not like about the videos.

Other Correlating Factors. The normalized sentiment distribution was fairly consistent across various channels and videos, with comments in preventive/informative content generally slightly positive. This is expected in topics that do not evoke strong emotions, unlike content about severe conditions, which could distress some viewers. Despite using sentiment and emotion analysis tools, clear patterns were elusive, likely due to the subjective nature of user preferences—what engages one user may seem lacking or overly complex to another.

6. Insights and Limitations

We reflect on limitations and insights, focusing on why users watch the videos, what they like or dislike about them, and possible correlations of sentiment with the visual content of the videos.

6.1. Limitations

Our video set is dominated by the *Nucleus Medical Media* channel, which makes up more than half of the videos since this channel is the most active channel producing videos featuring medical 3D visualizations. Our results might be influenced by the unequal sample size and including additional channels might uncover new insights. When searching for videos, their sorting on YouTube is not static and is tweaked to the account's preference. Thus, the same search methodology might show different videos to another person. The comments posted per video can also be influenced by channel owners, e.g., by filtering and deleting comments. As a result, certain content may have been mentioned by viewers, but the comments are not publicly available and were not included in our analysis. Also, the users who comment on YouTube may not be representative of the general user who only watches the video. Furthermore, the influence of external factors like the popularity of the channel, the time of posting, the video topic, and concurrent trends might confound the analysis. There is a possibility of bias or error in automated sentiment analysis tools, especially in interpreting sarcasm, irony, or cultural nuances, which we counteracted by manually investigating the comments.

6.2. Insights

User Motivation. As many users share personal experiences in their comments, a major reason for watching the videos seems to be personal dismay. Either the user themselves is or was affected by the topic or a relative. Interestingly, some users also write about personal experiences that lie in the past, thus, the personal dismay is not always acute. Other contexts to watch these videos could be for personal interest or academic purposes. However, this was not specifically mentioned in the comments.

Positive/Negative Feedback. Users appreciated the videos for simplifying medical topics with clear visuals, highlighting a preference for simplicity and aesthetics. Though specific visual styles were not singled out, the overall satisfaction with the visuals was high. Human voiceovers were favored over AI voices, although advancements in AI voice quality could change this perception. Background music received mixed reviews; it could be irritating if too dominant but added liveliness in the absence of voiceover. Gender-specific

medical videos typically featured narrators of the same gender. The narration speed also drew comments: It was deemed too slow in the video "How do carbohydrates affect your weight?" and too fast in "COVID-19 Animation: What Happens If You Get Coronavirus?", suggesting that the ideal pace may depend on the topic's complexity. For example, the biomedical processes of COVID-19 are likely perceived as more complex than basic information on healthy eating.

Correlations with Visual Content. Darker-themed videos generally received more negative comments, whereas videos with mixed colors received more positive comments. Additionally, the percentage of comments showing "sadness" was slightly larger in dark-themed videos. The results are not pronounced enough to conclude if color alone influences emotions in these cases because the content of the video and how it is presented might play a major role as well. If the video's topic is more somber or deals with serious topics like cancer, and it also has a dark theme, the combination could amplify feelings of sadness. While we observed a diverse range of design decisions in the videos, user comments were only rarely commenting on these. Instead, it became clear that viewers mainly wanted to share their personal experiences with the medical topic covered in the video.

7. Conclusion and Future Work

We analyzed the user responses of 76 YouTube videos showcasing 3D medical visualizations. While the user comments rarely target concrete design decisions, we are able to provide an overview and reflect on the motivation for watching the videos, aspects praised or criticized in the videos, and correlations with the videos' visual content. We only included top-level comments, excluding replies. However, analyzing the hierarchical structure of comments would allow us to also investigate discussions between several viewers. Channel owners can be asked to contribute to the analysis of the videos by providing additional information not accessible to viewers, such as demographics of the viewers and dropout rates. Since the comments may not be representative of the general audience, it is important to gain further insight into the behavior of users who watch medical videos but do not comment. Future studies can combine the analysis of the comments with an in-depth analysis of the video content, e.g., analyzing the design of transitions, anatomical structures, body functions, and the logical structure of the videos, potentially discovering further correlations.

References

[AGAZG23] ARDILA C. M., GONZÁLEZ-ARROYAVE D., ZULUAGA-GÓMEZ M.: Efficacy of three-dimensional models for medical education: A systematic scoping review of randomized clinical trials. *Heliyon 9*, 2 (2023), e13395. 1

[AHRL*15] AMINI F., HENRY RICHE N., LEE B., HURTER C., IRANI P.: Understanding data videos: Looking at narrative visualization through the cinematography lens. In *Proc. of ACM Conference on Human Factors in Computing Systems* (2015), p. 1459–1468. 1

[AM21] ALAPARTHI S., MISHRA M.: BERT: A sentiment analysis odyssey. *Journal of Marketing Analytics 9*, 2 (2021), 118–126. 2

[BA20] BOZKURT A. P., ARAS I.: Cleft lip and palate YouTube videos: Content usefulness and sentiment analysis. *The Cleft Palate-Craniofacial Journal 58*, 3 (2020), 362–368. 2

[CDMP21] CHIORRINI A., DIAMANTINI C., MIRCOLI A., POTENA D.: Emotion and sentiment analysis of tweets using bert. In *Prof. of EDBT/ICDT Workshops* (2021). 2

[DCLT19] DEVLIN J., CHANG M., LEE K., TOUTANOVA K.: BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2019), Association for Computational Linguistics, pp. 4171–4186. 2

[GMPB23] GARRISON L. A., MEUSCHKE M., PREIM B., BRUCKNER S.: *Current Approaches in Narrative Medical Visualization*. Springer Nature Switzerland, Cham, 2023, pp. 95–116. 1

[JJ20] JEAN B., JOSHUA G.: Youtube: Online video and participatory culture. *European Journal of Communication 35*, 4 (2020), 419–423. 1

[LBN*22] LOUREIRO D., BARBIERI F., NEVES L., ANKE L. E., CAMACHO-COLLADOS J.: Timelms: Diachronic language models from twitter. arXiv, 2022. 2

[LL19] LANGFORD A., LOEB S.: Perceived patient-provider communication quality and sociodemographic factors associated with watching health-related videos on YouTube: A cross-sectional analysis. *Journal of Medical Internet Research 21*, 5 (2019), e13512. 1

[LOG*19] LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L., STOYANOV V.: RoBERTa: A robustly optimized bert pretraining approach. *ArXiv abs/1907.11692* (2019). 2

[Low23] LOWE S.: roberta-base-go_emotions. https://huggingface.co/SamLowe/roberta-base-go_emotions, 2023. 2

[LTX*22] LEE J., TURNER K., XIE Z., KADHIM B., HONG Y.-R.: Association between health information–seeking behavior on YouTube and physical activity among U.S. adults: Results from health information trends survey 2020. *AJPM Focus 1*, 2 (2022), 100035. 1

[Pap23] PAPARIELLO L.: xlm-roberta-base-language-detection. https://huggingface.co/papluca/xlm-roberta-base-language-detection, 2023. 2

[PG16] PARK D. Y., GOERING E. M.: The health-related uses and gratifications of YouTube: Motive, cognitive involvement, online activity, and sense of empowerment. *Journal of Consumer Health on the Internet 20*, 1-2 (2016), 52–70. 1

[PL08] PANG B., LEE L.: *Opinion mining and sentiment analysis*, vol. 2. Now Publishers, Inc., 2008. 2

[PSN20] PORRECA A., SCOZZARI F., NICOLA M. D.: Using text mining and sentiment analysis to analyse YouTube italian videos concerning vaccination. *BMC Public Health 20*, 1 (2020). 2

[RNSS18] RADFORD A., NARASIMHAN K., SALIMANS T., SUTSKEVER I.: Improving language understanding by generative pre-training, 2018. 2

[SLL*21] SHI Y., LAN X., LI J., LI Z., CAO N.: Communicating with motion: A design space for animated visual narratives in data videos. In *Proc. of ACM Conference on Human Factors in Computing Systems* (2021), CHI '21. 1

[SOA*23] SHEVTSOV A., OIKONOMIDOU M., ANTONAKAKI D., PRATIKAKIS P., IOANNIDIS S.: What tweets and YouTube comments have in common? Sentiment and graph analysis on data related to US elections 2020. *PLOS ONE 18*, 1 (2023), e0270542. 2

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U., POLOSUKHIN I.: Attention is all you need. In *Proc. of Advances in Neural Information Processing Systems* (2017), vol. 30, Curran Associates, Inc. 2

[WG16] WELBOURNE D. J., GRANT W. J.: Science communication on youtube: Factors that affect channel and video popularity. *Public Understanding of Science 25*, 6 (2016), 706–718. 1

[WJJ17] WURA JACOBS A. O. A., JEON K. C.: Health information seeking in the digital age: An analysis of health information seeking behavior among us adults. *Cogent Social Sciences 3*, 1 (2017), 1302785. 1

[YBSX22] YANG S., BROSSARD D., SCHEUFELE D. A., XENOS M. A.: The science of YouTube: What factors influence user engagement with online science videos? *PLOS ONE 17*, 5 (2022), e0267697. 1