EduQuick: A Dataset Toward Evaluating Summarization of Informal Educational Content for Social Media

Zahra Kolagar[*]
Fraunhofer IIS
Erlangen, Germany
zahra.kolagar@iis.fraunhofer.de

Sebastian Steindl[*]
Ostbayerische Technische Hochschule
Amberg-Weiden, Germany
s.steindl@oth-aw.de

Alessandra Zarcone
Fraunhofer IIS, Erlangen, Germany
Technische Hochschule Augsburg, Germany
alessandra.zarcone@tha.de

Abstract

This study explores the capacity of large language models (LLMs) to efficiently generate summaries of informal educational content tailored for platforms like TikTok. It also investigates how both humans and LLMs assess the quality of these summaries, based on a series of experiments, exploring the potential replacement of human evaluation with LLMs. Furthermore, the study delves into how experienced content creators perceive the utility of automatic summaries for TikTok videos. We employ strategic prompt selection techniques to guide LLMs in producing engaging summaries based on the characteristics of viral TikTok content, including hashtags, captivating hooks, storytelling, and user engagement. The study leverages OpenAI's GPT-4 model to generate TikTok content summaries, aiming to align them with the essential features identified. By employing this model and incorporating human evaluation and expert assessment, this research endeavors to shed light on the intricate dynamics of modern content creation, where AI and human ingenuity converge. Ultimately, it seeks to enhance strategies for disseminating and evaluating educational information effectively in the realm of social media.

1 Introduction and Motivation

The omnipresence of social media in recent years is well-known (Ortiz-Ospina, 2019). The short video platform TikTok is a notable example due to its advanced content recommendation algorithm that is related to the users' flow experience (Qin et al., 2022). The algorithm uses multiple features, such as user interaction or watch time, to tailor the presented content to each individual. The average usage of the TikTok app is roughly 45-60 minutes per day, mostly prevalent among school students (Goetzen et al., 2023; Lebow, 2023). Therefore, motivating students to allocate some of their TikTok screen time to educational content can substantially boost the amount of time they dedicate to educational activities. It seems natural then to leverage the same mechanisms that lead to high usage time in a positive way, e.g. to increase engagement with and consumption of educational videos (Shaafi et al., 2023). However, educators are already heavily burdened in their day-to-day work, as shown by the high reported burnout rates (Marken and Agrawal, 2022), and should not have the additional task of summarizing classroom content to fit a short video format. Therefore, we see a need for effective summarization as a strategy to distil intricate information and to adapt educational content to social media short videos, especially TikTok.

We focus on text summarization, which serves to condense comprehensive information into concise yet coherent forms while preserving fundamental essence and enabling effective knowledge transmission in various domains, ranging from news articles to research papers (Allahyari et al., 2017; El-Kassas et al., 2021; Nenkova and McKeown, 2012). Particularly within the educational context, content summarization emerges as a potential solution to bridge the disparity between information overflow and the necessity for accessible and comprehensible insights. In social media-driven day-to-day life, the educational landscape has not wholly adapted to this. Content summarization would thus align seamlessly with the evolving demands of pedagogical practices on the one hand, and creating engaging content on the other hand.

The experiments in this paper led to EduQuick, a textual dataset for educational content summarization for TikTok video content creation. EduQuick is a multidomain textual dataset containing 500

[*] These authors contributed equally to this work.

Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, pages 32–48
November 1, 2023 ©2023 Association for Computational Linguistics

items, topics, source text, summarized articles, and metadata regarding the source text. We sourced the educational material from HowStuffWorks and constructed a template with features we identified for creating successful TikTok content (Section 4). We effectively instructed the GPT-4 model (OpenAI, 2023) to generate educational content adhering to the defined template. Next, for 150 of these summaries, we assessed the generated summaries through both human evaluation and an instruction-based evaluation using GPT-4 model (Section 5). Additionally, we elaborate on our efforts to gauge the quality and suitability of the educational content summaries for TikTok by seeking the insights of experienced TikTok content creators.

To sum up, we see a lot of unused potential in the use of short videos on social media to drive education, but are also aware that the production of these videos is too time-consuming to be adopted universally. This is further aggravated because the short format requires precise planning of the content and its deliverance, which in turn also makes them an effective tool for learning (Guo et al., 2014). Hence, we propose to leverage the summarization capabilities of modern LLMs to automatically create scripts for short educational videos suitable for social media platforms from input documents of educational content. This would shorten the video production process and make it a more feasible approach to modern, blended learning. In order to make progress on these summaries and their special requirements to be successful at TikTok, an automatic evaluation procedure is necessary. We will investigate if LLMs can fill this gap. Our endeavors culminated in the creation of a novel summarization dataset, along with the formulation of a comprehensive set of experimental designs. These designs were meticulously crafted to assess the proficiency of the GPT-4 model in both summarization and evaluation tasks. Furthermore, our work has generated a series of insights highlighting the existing deficiencies within the summarization domain and offering valuable guidance on its future trajectory. With this study we hope to inspire more efforts in this research direction and offer an approach to the summarization of data tailored for social media.

The following sections address the research questions listed below.

- RQ 1: Can LLMs efficiently generate summarizations of informal educational content for social media?
- RQ 2: How do humans judge the quality of these summarizations?
- RQ 3: Can the human evaluation be replaced by LLMs?
- RQ 4: Do experienced content creators rate the automatic summarizations regarding their usefulness for TikTok videos in the same way as crowd-sourced workers (RQ 2)?

2 Background and Related Work

2.1 Short videos for education

The effectiveness of short videos as educational tools has been part of multiple studies. Guo et al. (2014) investigate a large-scale dataset of video engagement data from an online course platform. They found that videos should be short, and informal but enthusiastic to increase engagement.

Brame (2016) compiled a survey on how to make educational videos more effective. The three core principles they identified are

- to manage cognitive load, e.g. highlighting keywords and chunking topics into multiple videos,
- increase engagement, e.g. by the same mechanisms identified by Guo et al. (2014),
- and invoke active learning by using interactive or guiding questions.

We deem TikTok as a possible target platform for the videos since campaigns like #learnontiktok already exist. Shaafi et al. (2023) found it to be a useful teaching tool, as it is widely used, easy to use and leads to a more engaging learning experience. Once the video has been produced, it can be distributed via various social media channels beyond TikTok.

2.2 Summarization and Education

Summarization is a fundamental Natural Language Processing (NLP) task that involves distilling large volumes of information into concise and coherent summaries. It serves as a valuable tool in various domains, including news articles, scientific papers, and legal documents (Altmami and Menai, 2022; El-Kassas et al., 2021; Kanapala et al., 2019). Automated summarization techniques have gained significant attention since the 1950s due to the exponential growth of digital content, which necessitates efficient information retrieval and consumption (Luhn, 1958; Allahyari et al., 2017).

The summarization of educational content has been a topic of research from multiple points of view. Yang et al. (2013) identified the trend of learning on mobile devices, and the inconvenience this creates, if the texts are too long. Their study found that apt summarizations can help the users' learning, especially if this aligns the content with the device used. Miller (2019) used a BERT model to develop a lecture summarization service to be used by students, paving the way for the use of modern Deep Learning-based solutions.

2.3 Large Language Models

Newer advancements such as pre-trained GPT-3/4 (Koubaa, 2023; OpenAI, 2023), BLOOM (Scao et al., 2022; Science, 2023), Llama models (Touvron et al., 2023), or dialogue-optimized models like InstructGPT (Ouyang et al., 2022), ChatGPT (OpenAI, 2022), and Falcon-40B-instruct (Almazrouei et al., 2023; Penedo et al., 2023; Xu et al., 2023), have gained attention also in summarization research. These LLM-based approaches have shown promising results in generating high-quality summaries. These models can capture long-range dependencies, handle complex sentence structures, and produce coherent and contextually appropriate summaries while being dependent on a small set of annotated datasets (few-shot approach also known as in-context learning) or without task-specific training (zero-shot approach) among others (Bražinskas et al., 2020; Fabbri et al., 2020; Adams et al., 2022). Both few-shot and zero-shot approaches make use of prompt-based instructions to tailor the model to generate a desired output. As such, prompt engineering has emerged as a crucial discipline for optimizing LLMs by tailoring their output through prompt-based instructions. It involves developing effective prompting techniques and leveraging LLMs for various tasks including summarization. More recent examples of prompting techniques include chain-of-thought prompting (Wei et al., 2022), self-consistency prompting for creating diverse reasoning paths (Wang et al., 2022), tree-of-thought prompting (Yao et al., 2023), graph integration (Liu et al., 2023b), active prompting (Diao et al., 2023), and multimodal chain-of-thought prompting through image integration (Zhang et al., 2023b).

2.4 Summarization Datasets

While a variety of datasets have been instrumental in advancing content summarization techniques, it is crucial to note that none of these datasets directly cater to the specific needs of summarizing educational content for social media platforms, which demand a more informal and engaging style. Prominent datasets like CNN/Daily Mail (Nallapati et al., 2016) and Gigaword (Rush et al., 2015; Graff et al., 2003) primarily focus on news articles, while the Wikihow (Koupaee and Wang, 2018) focuses on Wikipedia articles – just to name a few. These existing datasets have undoubtedly contributed to the evolution of summarization models. However, they do not align with the unique characteristics of educational content designed for platforms like TikTok. The informal and conversational language style, as well as the succinct yet attention-grabbing nature of educational content on social media require a new approach capturing these distinctive qualities. This recognition has led to the creation of EduQuick, a summarization dataset created specifically for educational content summarization for social media, filling a gap that currently exists in the response to the popularity of bite-sized educational videos, and a need for summarization techniques that can distill complex topics into captivating and digestible narratives using the capabilities of LLMs.

3 Data Collection and Preprocessing

To create a dataset containing educational yet entertaining content for TikTok videos, the data was extracted using web scraping techniques from "HowStuffWorks"[1] (Brain, 2023), a website known for its diverse educational content on subjects including science, history, animals, entertainment, culture, technology, and lifestyle. This choice was based on the website's abundance of interesting and informative material, aligning perfectly with our aim to produce engaging and educational TikTok content. We have extracted 100 articles per topic, resulting in a dataset comprising 500 articles across 5 diverse topics (*health*, *entertainment*, *animals*, *science* and *auto*).

Throughout the collection process, we applied minimal preprocessing, ensuring that the entirety of each article's content was retained to maintain its integrity and authenticity. In addition to the articles'

[1] https://www.howstuffworks.com/

content, we collected valuable metadata, including citation information, such as article links, authors' names, publication dates, and extracting dates. The metadata offers crucial contextual information and simplifies the process of citing the TikTok content accurately.

4 Enhancing TikTok Content Creation through Strategic Prompt Selection

Prompt Design plays a pivotal role in guiding LLM models to create engaging TikTok content summaries based on the collected articles (as described in section 3). We decided to adopt a template prompt that incorporates essential features identified for viral educational TikTok video content, as described in section 4.1. These features were curated based on insights from a qualitative analysis of renowned educational TikTok content creators (i.e. @Veritasium, @renegadescienceteacher, @distilledscience, @ChemTeacherPhil) and the research cited above. The selected prompt approaches were chosen for their ability to enhance relevance, captivate viewers' attention, and ensure an appealing learning experience.

4.1 Characteristics of Viral TikTok Content

Successful educational TikTok content exhibits a combination of key features that captivate viewers and foster a positive learning experience. In this section, we will focus on some of the aspects that are relevant to the textual content of viral TikTok videos. First, incorporating trending hashtags into TikTok textual content provides enhanced visibility and reach, drawing more attention to the content (Ling et al., 2022; Rauschnabel et al., 2019; Zappavigna, 2015; Daer et al., 2014). To further seize viewers' interest, a compelling *hook* is crucial – beginning the video with an attention-grabbing introduction, such as a surprising fact, a thought-provoking question, or a fascinating statistic related to the educational topic. By employing storytelling techniques, creators can establish a connection with the audience, presenting the content in the form of a short narrative or engaging anecdote related to the subject matter. Moreover, making use of storytelling features in creating educational content enhances emotional engagement, making it relatable and fostering a deeper connection with viewers, as evident in popular TikTok content.

Educational creators are encouraged to cover a range of topics, ensuring that the content caters to various interests and preferences. Additionally, simplifying complex concepts is key, especially when targeting viewers who may not possess in-depth knowledge of the subject. Through the use of clear and concise language, along with relatable examples or analogies, content creators can make their content more accessible. To further promote engagement, concluding each video with a strong call-to-action encourages viewers to like, comment, share, and follow the content creator's account for more educational content (Le Compte and Klug, 2021). By inviting viewers to participate by asking questions or suggesting future topics, creators can establish an interactive and collaborative environment. Finally, teasing upcoming content; e.g., using hashtags like "#StayTuned" or "#ComingSoon" as well as dividing content into more parts, fosters anticipation and cultivates a loyal following (Lin, 2023; Oktopi, 2022; Radulescu, 2022).

4.2 Crafting Effective Prompt for Engaging Content Creation

In pursuit of creating engaging and consistent content summaries, we adopted a template prompt approach to streamline the content creation process. By designing a comprehensive prompt template (cf. Fig. 2) based on the selected TikTok features, we aimed to enhance viewer engagement and align with our objective of producing educational yet entertaining content. This template encompasses key elements described in section 4.1 to ensure that the LLM generates content summaries that incorporate the features. Leveraging this prompt design, we empowered the model to effectively distill the essence of the collected articles and deliver compelling TikTok content.

4.3 Zero-Shot Template Utilization for TikTok Content Generation

We used OpenAI's GPT4-8k (OpenAI, 2023) model to generate TikTok content by adopting a systematic process. To instruct the model, we used a template which consists of an instruction that guides the model on the key features to include in the generated TikTok content. The dataset of articles from HowStuffWorks was used as an input, paired with the instruction. Upon generating the TikTok content summaries, the output from the model was saved alongside the original dataset of articles (cf. Appendix A.2 for an example summary). These combined datasets formed the basis for the empirical study described in section 5.

5 Evaluating GPT-4 Generated Content

5.1 Comparing Human and GPT-4 as Evaluators

To ensure the validity and effectiveness of the generated TikTok content, an empirical study was conducted following the methodology proposed by Liu et al. (2023a). For the evaluation process, five participants were recruited from Amazon Mechanical Turk (AMT). We set the workers approval rate to greater than 98% and provided detailed annotation instructions. Each participant was presented with both the original text and the content generated by GPT-4. They were asked to rate the generated content on three essential criteria using a 1 to 5 scale (1 being the worst, and 5 being the best), namely:

- **Cohesiveness**: Assessing how well the sentences in the story fragment fit together to form a coherent narrative.
- **Likability**: Gauging the level of enjoyment and enjoyment experienced by the participants while reading the story fragment.
- **Relevance**: Determining how closely the output aligns with the instruction given to GPT-4 through the template.

See Appendix A.3 for details on the annotation instructions and a sample of the task presented to the workers. We also included an optional comment section for workers. We collected five different annotations for each combination of the educational article, assignment (prompt), and summaries. In the interest of practicality, the evaluation was conducted on a subset of the dataset, consisting of 150 randomly selected samples (30 samples per topic). Given the high cost of human evaluation, we opted to assess the summaries using an evaluative template prompt created for GPT-4, following the same instructions as provided to human participants, described in Figure Number. We evaluated the same 150 samples with only the GPT-4 model following Liu et al. (2023a), and focused on this model as earlier versions did not demonstrate the level of performance achieved by this one.

Additionally, to ensure the reliability and consistency of the human evaluations, we calculated the inter-annotator agreement among the five recruited participants. Cases, where at least three annotators provided identical ratings for the enlisted questions, were considered instances of agreement. Overall, the annotation process yielded a high inter-annotator agreement, with an overall Krippendorff's α 84,57 % (Hayes and Krippendorff, 2007; Artstein and Poesio, 2008). To answer RQ 2, this table shows that the humans give the summaries good ratings on all criteria with a high inter-annotator agreement. This indicates that the model successfully created summaries that are suitable for short educational videos on social media. We therefore answer RQ 1 positively. The human evaluation results are compiled in Table 1.

Criteria	Avg. Rating (150 samples)	Inter-annotator Agreement
Cohesiveness	3.73	85.06 %
Likeability	3.72	82.26 %
Relevance	3.71	86.40 %

Table 1: Comparison of Average Rating Scores on 150 samples and Inter-Annotator Agreement.

Initially, our intention was to assess not only the randomly selected 150 samples, which were also rated by humans but to evaluate the entire dataset using GPT-4. However, upon reviewing the results of GPT-4's evaluation for the 150 samples, we observed a consistent pattern where the model consistently assigned a perfect score of 5 across all criteria, including cohesiveness, relevance, and likeability. This pattern indicated a low level of agreement with human annotators. Therefore, we could not reproduce the same results as Liu et al. (2023a).

One plausible explanation for this pattern could be that the summaries were also generated by the same model. As previously demonstrated, GPT-based models tend to favor their own generated text more than text generated by other models (Chiang and Lee, 2023; Liu et al., 2023a). Conversely, these models exhibit a high degree of sensitivity to the provided prompt and input data. Consequently, we hypothesize that this sensitivity may also contribute to the observed disparity in results. This phenomenon underscores the inherent challenge of consistently reproducing the same outcomes using this approach. Other explanations might be that the criteria are inherently subjective, especially *Likeability*, which might be a problem for the used model. Also, the pretraining on large amounts of internet text, where positivity is often rewarded, might play a part in this. In contrast, humans often opt for moderate scores, such as 4, as they may hesitate to select perfect scores or lower scores like 2 or 1 (Kiritchenko and Mohammad, 2017; Asaadi et al., 2019). Therefore, the shortcomings of the model demonstrated during the evaluation experiments, lead us to answer RQ 3 negatively: The model is not a suitable replacement for human evaluation.

5.2 Experiments on the Evaluation Prompt

To ensure the robustness of our experiment's design, we experimented with several precautionary measures as outlined below. However, they did not lead to any changes in the evaluation results as we still obtained a perfect score of 5 across all criteria.

First, we tried requesting the model to provide reasoning for its chosen ratings.

Sample system response:

- Cohesiveness (5/5): The summarization flows smoothly, linking ideas effectively and painting a vivid picture of Ledger art and its evolution.
- Relevance (5/5): The summary perfectly aligns with the instructions, providing an engaging, informal and enthusiastic script for a short video. It also incorporates a call for engagement.
- Likeability (5/5): The summary is very enjoyable and compelling. It manages to convey the historical and cultural importance of Ledger art in a casual, conversational tone.

Second, we included additional criteria for the model to assess the summaries. While we acknowledge that the resulting evaluations may not directly align with human assessments due to the novel criteria, our objective was to examine whether a more comprehensive evaluation framework would impact the behavior of the GPT-4 model. The additional criteria included:

- **Clarity**: How clear and easily understandable is the summary?
- **Conciseness**: Is the summary free from unnecessary or redundant information?
- **Utility**: How useful is the summary for the purpose of creating content for TikTok videos?
- **Novelty**: Does the summary offer a fresh perspective or new insights on the source text, or does it merely restate existing information?

Third, we presented the model with a sample summary that had been independently evaluated by two human annotators. This served a dual purpose: firstly, it demonstrated to the model that human evaluations could still exhibit traces of subjectivity in their ratings. Secondly, we assumed it would educate the model on the nuances of human evaluation, highlighting the disparities in assessment between humans and models for this specific task. However, we observed that the model copies human annotations across the given criteria.

Lastly, given the inclination of each LLM to favor their own generated content over text generated by other models or humans, we opted for a systematic approach. We handpicked 20 educational articles from our dataset and enlisted a single AMT participant per article. These participants were tasked with summarizing the articles, utilizing the exact same prompt employed with the model. In a subsequent phase, we once again employed GPT-4 to assess the summaries created by humans, taking into account the source article, the assignment (prompt), and the three criteria outlined in section 5.1. The results of the GPT-4 evaluation revealed consistently low scores of 1 across all criteria for all human-generated summaries.

Finally, we initiated a second round of annotation experiments. In this phase, we recruited 5 participants and requested them to select the summary that best conformed to the assignment (prompt) in order to determine human preference. Remarkably, in all instances, all 5 annotators unanimously favored the text generated by GPT-4 over that generated by humans for the same article.

5.3 Recommendations for Enhancing GPT-4's Evaluation Competence

Based on our observations, we offer the following suggestions to fellow researchers who rely on GPT-4 or other LLMs for evaluation tasks. Due to the necessity for thorough analysis and experiment design for each point, we only provide our insights and potential suggestions.

Fine-Tuning for Summarization: When feasible, consider fine-tuning your LLM on a dataset specifically tailored for summarization tasks.

Iterative Feedback Loop: Implement an iterative feedback mechanism that fosters collaboration between the LLM and human evaluators e.g., using a reward mechanism. See Stiennon et al. (2022)

Objective Evaluation Metrics: Explore the possibility of introducing objective evaluation metrics where the model provides scores based on mathematical formulas rather than relying solely on subjective criteria.

Comparative Evaluations: If you have access to multiple LLMs with similar capabilities, consider conducting comparative evaluations. Pair one model's generated output with another model's evaluation and vice versa.

The empirical study serves as a vital step in validating the quality and adherence of the GPT-4 generated TikTok content to the designated prompt design and example context.

5.4 Evaluation Involving Content Creators

To further assess the quality and suitability of the generated educational content summaries for TikTok, we sought the expert opinions of three experienced TikTok content creators. Their deep understanding of the platform's dynamics and audience preferences makes their insights invaluable in evaluating the generated content's efficacy.

We provided the content creators with a sample set of 10 of the generated summaries and requested their evaluation. They were asked to assess the suitability of the summaries as educational content for TikTok, considering factors such as engagement potential, alignment with TikTok's informal style, and the ability to convey information concisely. To facilitate this evaluation, we devised a simple questionnaire comprising 6 questions, tailored to capture their impressions and observations. The questionnaire, responses, and observations provided by these experts are summarized in Figure 1, and the questionnaire is presented in the Appendix 8. The participants provided a unanimous response to questions 3 to 5, showing a high level of agreement in those areas. Their responses to the other questions exhibited only slight variations. Overall, their ratings consistently exceeded 3, speaking for the experiment's validity and the quality of the generated summaries. Thus, RQ 4 is also answered positively.

6 The Dataset

The presented dataset is a curated collection of model-generated text for educational TikTok content, abbreviated as EduQuick. This dataset is the result of evaluating and selecting high-quality content generated by the GPT-4 model following an empirical study. It aims to provide engaging and informative summaries suitable for TikTok's educational audience. While our evaluation of GPT-4's assessment capability, as discussed in the preceding section, did not meet our expectations, it is worth noting that LLMs have already demonstrated their capacity to generate high-quality summaries (Zhang et al., 2023a).

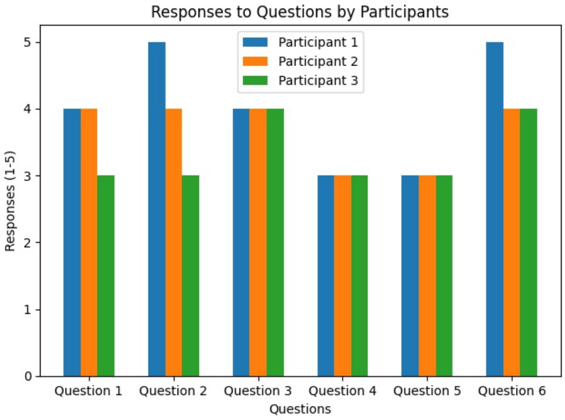


Figure 1: Evaluation of the summaries by experienced TikTok content creators. The questions are provided in the Appendix in Fig. 8.

Table 2 presents descriptive statistics of the dataset. We present statistics that include the average length of educational articles and their corresponding summaries per topic, the token count per topic, and the distinct count of lemmatized word forms. Tokenization was performed by splitting text based on whitespace. For lemmatization, which involves obtaining the base form of words found in a dictionary, we utilized the English SpaCy model en_core_web_sm version 3.6.0 (Honnibal et al., 2020).[2] We evaluate lexical richness across topics by reporting root type-token ratio (RTTR; Guiraud, 1958) as well as the measure of textual lexical diversity (MTLD; McCarthy and Jarvis, 2010) computed with the threshold of 0.72 using the Lexical-Richness library (Shen, 2022)[3], as MTLD is less affected by the length of the text. The educational articles as well as the summaries exhibit high measures for both RTTR and MTLD, indicating a noteworthy level of lexical diversity within the EduQuick dataset.

[2] https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.3.0

[3] https://github.com/LSYS/LexicalRichness

	Average Length		Tokens		Lemma		RTTR		MTLD	
Topics	Articles	Summaries	Articles	Summaries	Articles	Summaries	Articles	Summaries	Articles	Summaries
animals	6185	1156	78395	14526	91604	18178	34.61	25.74	118.23	135.34
auto	7777	1085	132564	18357	156079	22990	33.70	26.65	95.60	124.49
entertainment	8025	1125	1372299	18821	161142	23573	38.29	30.35	95.54	125.72
health	6960	1157	115683	18976	134448	23834	33.39	26.92	103.74	132.87
science	7014	1182	114830	19224	132645	23864	37.81	29.99	95.70	130.06

Table 2: Descriptive statistics of the EduQuick dataset containing a total of 500 samples (RTTR = root type-token ratio; MTLD = measure of textual lexical diversity).

7 Conclusion and Future Work

In this work we focused on generating engaging educational content for TikTok. We extracted materials from HowStuffWorks and created a template based on successful TikTok features. Using GPT-4, we instructed the model to generate educational content based on the template we crafted. We evaluated the generated content through human assessment and GPT4 evaluation, resulting in two sets of evaluation scores, which we term silver[4] standard dataset. The released dataset and evaluation scores offer valuable resources for future research and development in natural language generation for TikTok's educational content creation.

In future work, exploring advanced techniques for fine-tuning LLM models specifically for TikTok content generation could lead to higher-quality and more engaging educational content. We argue that the automatic evaluation of this specific content is still a challenging task since GPT-4 was not able to fill this gap. While human evaluation through crowdsourcing is possible, we argue that due to its high cost, it is impracticable for the development cycle of summarization systems. We therefore call on the scientific community to devise an automatic evaluation procedure, that will in turn facilitate research into the automatic summarization for educational short videos.

Moreover, integrating summaries with AI-generated talking-head videos and audio presents an intriguing niche for enhancing the educational impact and viewer engagement of the generated TikTok content as well as providing a complete automatic pipeline for social media video generation. Finally, conducting user studies and collecting feedback directly from TikTok users can provide valuable insights into their preferences and interests in educational content, guiding the refinement of the content generation process and creating TikTok videos that resonate more effectively with the platform's diverse audience.

Limitations

The research presented here has notable strengths in generating engaging educational content for TikTok and conducting comprehensive evaluations. However, certain limitations should be acknowledged. The dataset was limited to specific topics and sources, and a more diverse range of educational content could provide broader insights. Additionally, while the automatic evaluation metrics were effective, they might not capture all content quality aspects. employing AMT for human evaluation presented a challenge concerning the utilization of emojis, as they were not allowed on this platform.

Furthermore, our evaluation involving TikTok content creators, while informative, is subject to certain limitations. The use of limited sample size was due to challenges in accessing a broader range of participants, limiting the representation of diverse content creator perspectives. Moreover, individual variations in content creation styles and preferences may have influenced evaluations despite efforts to elicit general impressions. While this study focused on content creators, the insights might not fully extend to the broader TikTok audience. To address these limitations, future research could consider broader participation and a larger, more diverse content creator sample.

Despite these limitations, this research serves as a solid foundation for future explorations in educational content generation for TikTok and other social media platforms.

[4] The term "Silver Standard Dataset" is employed in this paper instead of "Gold Standard Dataset" to reflect the approach used for evaluation. While traditional gold standard datasets are typically assessed by human evaluators, our evaluation process involves employing GPT models and humans. This distinction underscores the unique evaluation methodology applied in this research, where an AI model contributed to the assessment process, leading to the adoption of the term "Silver Standard Dataset."

Ethics Statement

Social Media Platforms

While multiple social media platforms have been a global success, many have raised concerns about their negative impacts, with research focusing for example on social media addiction (Pellegrino et al., 2022). The same mechanisms that lead to the flow experience, also increase the risk of addiction (Qin et al., 2022). Consequently, the utilization of any social media platform for educational purposes should be subject to vigilant oversight and thorough planning to prevent any potential harm, especially among younger students.

Experiments Involving Human Participants

The workers we recruited on AMT platform maintain their anonymity, a practice aligned with ethical norms within the community. They were recruited voluntarily and provided a written consent form to participate in the study and were allowed to opt-out at any point in time. Moreover, the AMT workers were compensated in accordance with the norms and regulations of the AMT platform for their time and effort spent on our tasks. We encouraged feedback from AMT workers and offered to promptly address any concerns or issues that might arise during the research process. However, we did not record any issues and we received positive feedback regarding the experiments.

Furthermore, the content creators assessing our summaries also opted for anonymity. They were contacted through the TikTok platform and were recruited voluntarily for this research project. Prior to involving TikTok content creators in our study, we provided a transparent information regarding the research's purpose, methodology, and potential implications. Content creators provided informed consent, demonstrating their voluntary participation.

Acknowledgements

We extend our heartfelt gratitude to the anonymous participants whose invaluable contributions were instrumental in the completion of this research. Additionally, we wish to express our appreciation to our reviewers for their valuable insights.

References

David Adams, Gandharv Suri, and Yllias Chali. 2022. Combining state-of-the-art models with maximal marginal relevance for few-shot and zero-shot multi-document summarization. *arXiv preprint arXiv:2211.10808*.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2022. Automatic summarization of scientific articles: A survey. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1011–1028.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Shima Asaadi, Saif Mohammad, and Svetlana Kiritchenko. 2019. Big BiRD: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 505–516, Minneapolis, Minnesota. Association for Computational Linguistics.

Marshall Brain. 2023. Howstuffworks.

Cynthia J. Brame. 2016. Effective Educational Videos: Principles and Guidelines for Maximizing Student Learning from Video Content. *CBE—Life Sciences Education*, 15(4):es6.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. *arXiv preprint arXiv:2004.14884*.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations?

Alice R Daer, Rebecca Hoffman, and Seth Goodman. 2014. Rhetorical functions of hashtag forms across social media applications. In *Proceedings of the 32nd ACM International Conference on the Design of Communication CD-ROM*, pages 1–3.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Alexander R Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2020. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. *arXiv preprint arXiv:2010.12836*.

Angelica Goetzen, Ruizhe Wang, Elissa M. Redmiles, Savvas Zannettou, and Oshrat Ayalon. 2023. Likes and Fragments: Examining Perceptions of Time Spent on TikTok.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

P. Guiraud. 1958. *Problèmes et Méthodes de La Statistique Linguistique*. Dodrecht: D. Reidel.

Philip J. Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, pages 41–50, Atlanta Georgia USA. ACM.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. 2019. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51:371–402.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Anis Koubaa. 2023. Gpt-4 vs. gpt-3.5: A concise showdown.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Daniel Le Compte and Daniel Klug. 2021. "it's viral!"-a study of the behaviors, practices, and motivations of tiktok users and social activism. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pages 108–111.

Sara Lebow. 2023. 5 charts on video marketing's momentum. https://www.insiderintelligence.com/content/5-charts-on-video-marketing-momentum.

Ying Lin. 2023. How to go viral on tiktok: 15 ideas for 2023.

Chen Ling, Jeremy Blackburn, Emiliano De Cristofaro, and Gianluca Stringhini. 2022. Slapping cats, bopping heads, and oreo shakes: Understanding indicators of virality in tiktok short videos. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 164–173.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Zemin Liu, Xingtong Yu, Yuan Fang, and Xinming Zhang. 2023b. Graphprompt: Unifying pre-training and downstream tasks for graph neural networks. In *Proceedings of the ACM Web Conference 2023*, pages 417–428.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Stefanie Marken and Sangeeta Agrawal. 2022. K-12 Workers Have Highest Burnout Rate in U.S. https://news.gallup.com/poll/393500/workers-highest-burnout-rate.aspx.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data*, pages 43–76.

Oktopi. 2022. How to increase the visibility of educational content on tiktok.

OpenAI. 2022. OpenAI: Introducing ChatGPT. [Online; posted 30-November-2022].

OpenAI. 2023. Gpt-4 technical report.

Esteban Ortiz-Ospina. 2019. The rise of social media. *Our World in Data*. Https://ourworldindata.org/rise-of-social-media.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Alfonso Pellegrino, Alessandro Stasi, and Veera Bhatiasevi. 2022. Research trends in social media addiction and problematic social media use: A bibliometric analysis. *Frontiers in Psychiatry*, 13.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Yao Qin, Bahiyah Omar, and Alessandro Musetti. 2022. The addiction behavior of short-form video app TikTok: The information quality and system quality perspective. *Frontiers in Psychology*, 13.

Anda Radulescu. 2022. How to create viral content on tiktok in 10 easy steps.

Philipp A Rauschnabel, Pavica Sheldon, and Erna Herzfeldt. 2019. What motivates users to hashtag on social media? *Psychology & Marketing*, 36(5):473–488.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Big Science. 2023. Introducing the world's largest open multilingual language model: Bloom.

Nur Farha Shaafi, Mohammad Mubarrak Mohd Yusof, Nurul Nabilla Mohammad Khalipah, and Norhazly Mohd Hanif. 2023. Investigating TikTok as A Learning Tool for Learning Chemistry: A Study Among Secondary School Students in Malaysia. 11(1).

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Guangbing Yang, Nian-Shing Chen, Kinshuk, Erkki Sutinen, Terry Anderson, and Dunwei Wen. 2013. The effectiveness of automatic text summarization in mobile learning contexts. *Computers & Education*, 68:233–243.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

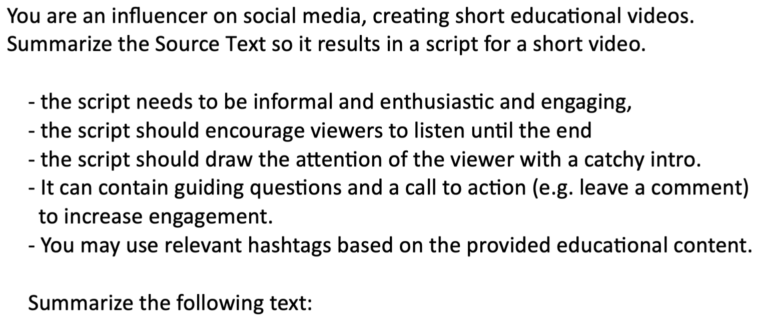
Michele Zappavigna. 2015. Searchable talk: The linguistic functions of hashtags. *Social semiotics*, 25(3):274–291.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023a. Benchmarking large language models for news summarization.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

A Appendix

A.1 Prompt Presented to the GPT-4 Model for the Summarization Task



You are an influencer on social media, creating short educational videos.
Summarize the Source Text so it results in a script for a short video.

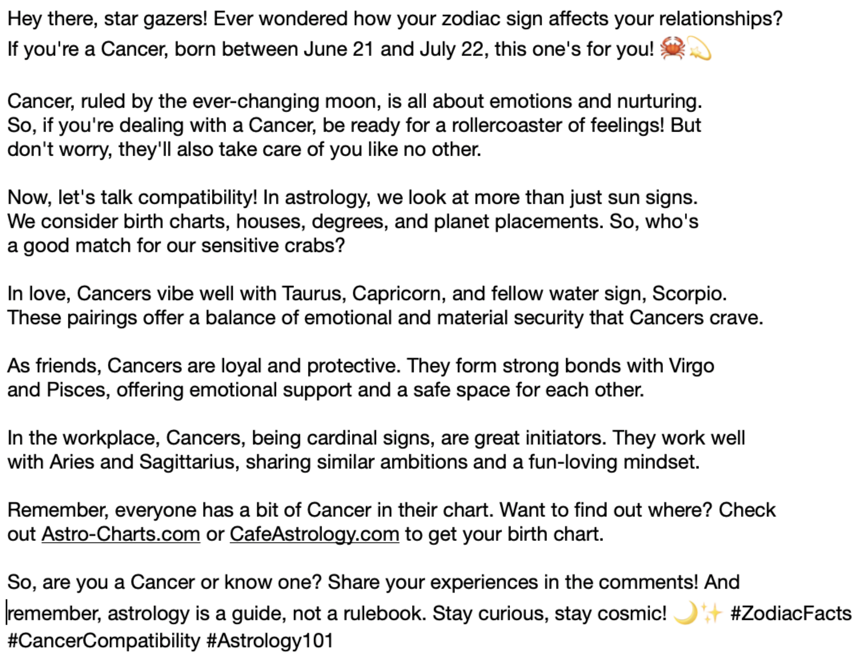
- the script needs to be informal and enthusiastic and engaging,
- the script should encourage viewers to listen until the end
- the script should draw the attention of the viewer with a catchy intro.
- It can contain guiding questions and a call to action (e.g. leave a comment) to increase engagement.
- You may use relevant hashtags based on the provided educational content.

Summarize the following text:

Figure 2: The prompt presented to the GPT-4 model for the summarization task.

A.2 Example GPT-4 Generated Summary on the Topic of Entertainment

Link to the source article: https://entertainment.howstuffworks.com/horoscopes-astrology/scorpio.htm



Hey there, star gazers! Ever wondered how your zodiac sign affects your relationships?
If you're a Cancer, born between June 21 and July 22, this one's for you! 🦀💫

Cancer, ruled by the ever-changing moon, is all about emotions and nurturing. So, if you're dealing with a Cancer, be ready for a rollercoaster of feelings! But don't worry, they'll also take care of you like no other.

Now, let's talk compatibility! In astrology, we look at more than just sun signs. We consider birth charts, houses, degrees, and planet placements. So, who's a good match for our sensitive crabs?

In love, Cancers vibe well with Taurus, Capricorn, and fellow water sign, Scorpio. These pairings offer a balance of emotional and material security that Cancers crave.

As friends, Cancers are loyal and protective. They form strong bonds with Virgo and Pisces, offering emotional support and a safe space for each other.

In the workplace, Cancers, being cardinal signs, are great initiators. They work well with Aries and Sagittarius, sharing similar ambitions and a fun-loving mindset.

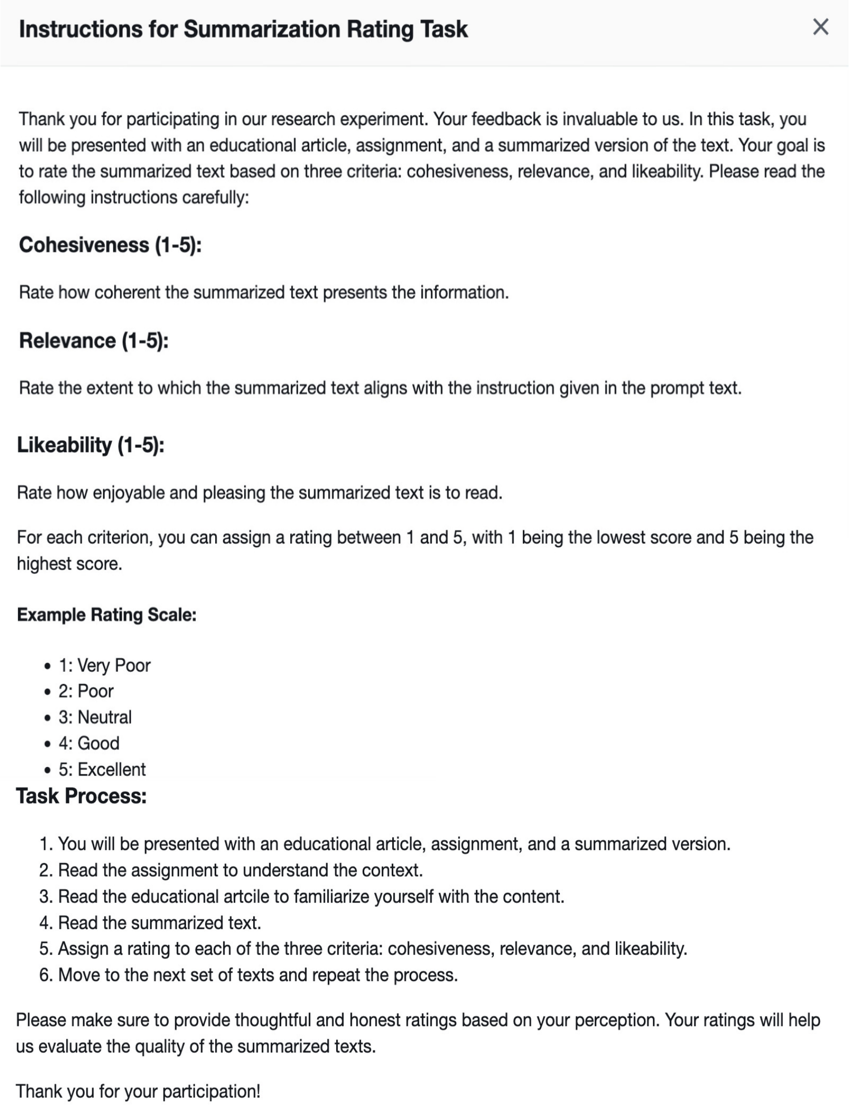
Remember, everyone has a bit of Cancer in their chart. Want to find out where? Check out Astro-Charts.com or CafeAstrology.com to get your birth chart.

So, are you a Cancer or know one? Share your experiences in the comments! And remember, astrology is a guide, not a rulebook. Stay curious, stay cosmic! 🌙✨ #ZodiacFacts #CancerCompatibility #Astrology101

Figure 3: Sample GPT-4 generated summary

A.3 Annotation Task Described in Section 5.1

Upon accepting the task, AMT workers were directed to a dedicated page containing concise yet comprehensive instructions detailing the task's execution. The provided figures illustrate these instructions, offering both a visual guide and an exemplar showcasing the task's format. The AMT workers received the source article, the prompt used for generating the summaries, and the summarized text. We provided straightforward and minimal instructions, asking them to evaluate the summarized texts in relation to the source articles and the prompt; which was presented as guidance for crafting the summaries.



Instructions for Summarization Rating Task

Thank you for participating in our research experiment. Your feedback is invaluable to us. In this task, you will be presented with an educational article, assignment, and a summarized version of the text. Your goal is to rate the summarized text based on three criteria: cohesiveness, relevance, and likeability. Please read the following instructions carefully:

Cohesiveness (1-5):

Rate how coherent the summarized text presents the information.

Relevance (1-5):

Rate the extent to which the summarized text aligns with the instruction given in the prompt text.

Likeability (1-5):

Rate how enjoyable and pleasing the summarized text is to read.

For each criterion, you can assign a rating between 1 and 5, with 1 being the lowest score and 5 being the highest score.

Example Rating Scale:

- 1: Very Poor
- 2: Poor
- 3: Neutral
- 4: Good
- 5: Excellent

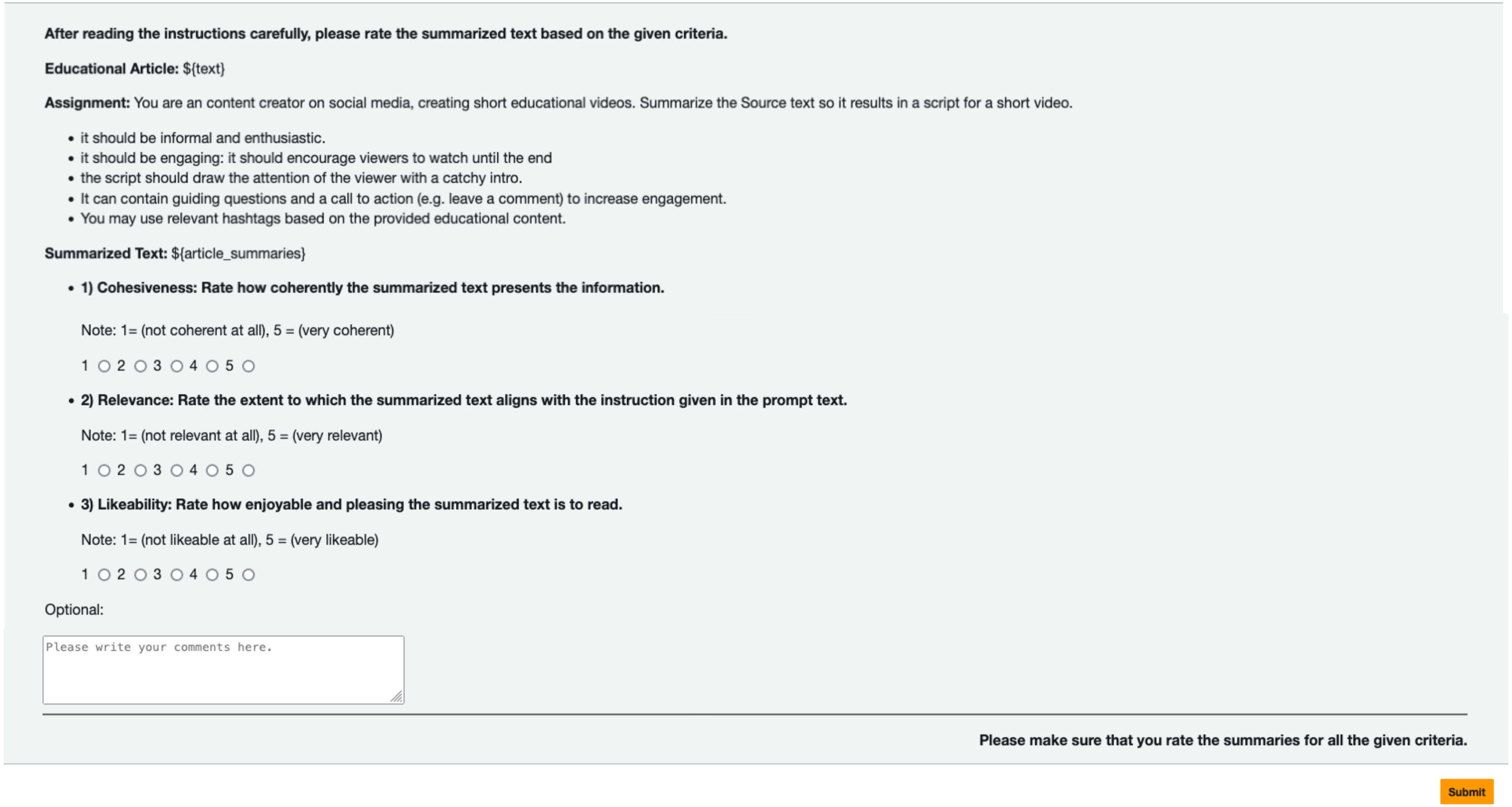
Task Process:

1. You will be presented with an educational article, assignment, and a summarized version.
2. Read the assignment to understand the context.
3. Read the educational artcile to familiarize yourself with the content.
4. Read the summarized text.
5. Assign a rating to each of the three criteria: cohesiveness, relevance, and likeability.
6. Move to the next set of texts and repeat the process.

Please make sure to provide thoughtful and honest ratings based on your perception. Your ratings will help us evaluate the quality of the summarized texts.

Thank you for your participation!

Figure 4: The instruction presented on Amazon Mechanical Turk.



After reading the instructions carefully, please rate the summarized text based on the given criteria.

Educational Article: ${text}

Assignment: You are an content creator on social media, creating short educational videos. Summarize the Source text so it results in a script for a short video.

- it should be informal and enthusiastic.
- it should be engaging: it should encourage viewers to watch until the end
- the script should draw the attention of the viewer with a catchy intro.
- It can contain guiding questions and a call to action (e.g. leave a comment) to increase engagement.
- You may use relevant hashtags based on the provided educational content.

Summarized Text: ${article_summaries}

- **1) Cohesiveness: Rate how coherently the summarized text presents the information.**

  Note: 1= (not coherent at all), 5 = (very coherent)

  1 ○ 2 ○ 3 ○ 4 ○ 5 ○

- **2) Relevance: Rate the extent to which the summarized text aligns with the instruction given in the prompt text.**

  Note: 1= (not relevant at all), 5 = (very relevant)

  1 ○ 2 ○ 3 ○ 4 ○ 5 ○

- **3) Likeability: Rate how enjoyable and pleasing the summarized text is to read.**

  Note: 1= (not likeable at all), 5 = (very likeable)

  1 ○ 2 ○ 3 ○ 4 ○ 5 ○

Optional:

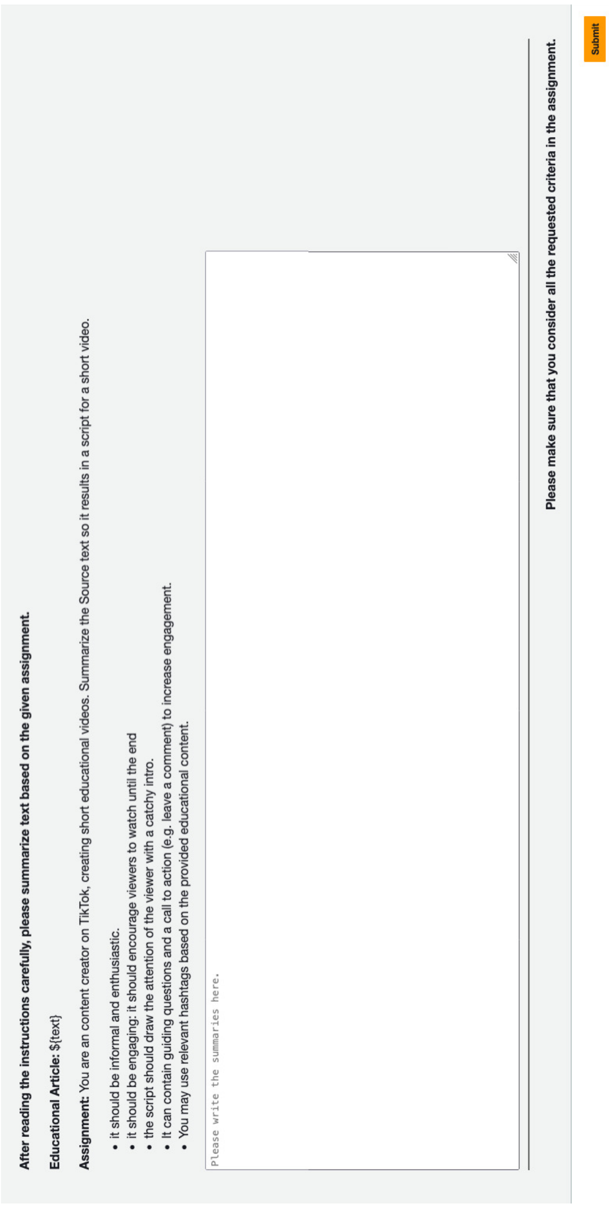
Please write your comments here.

Please make sure that you rate the summaries for all the given criteria.

Figure 5: The annotation task presented on Amazon Mechanical Turk.

A.4 Summarization Task Described in Section 5.2

The summarization task involved one per source text whose task was to generate a summary based on the educational article and the requested assignment.



After reading the instructions carefully, please summarize text based on the given assignment.

Educational Article: ${text}

Assignment: You are an content creator on TikTok, creating short educational videos. Summarize the Source text so it results in a script for a short video.

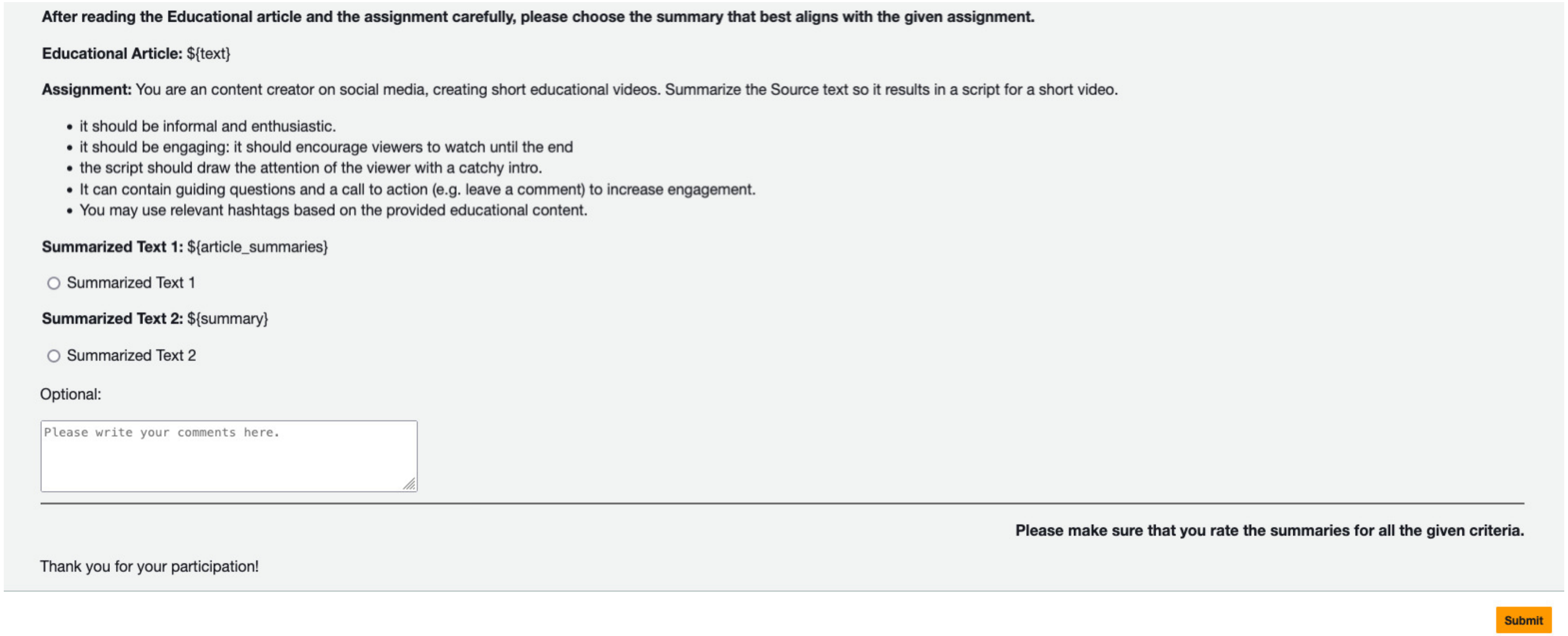
- it should be informal and enthusiastic.
- it should be engaging: it should encourage viewers to watch until the end
- the script should draw the attention of the viewer with a catchy intro.
- It can contain guiding questions and a call to action (e.g. leave a comment) to increase engagement.
- You may use relevant hashtags based on the provided educational content.

Please make sure that you consider all the requested criteria in the assignment.

Figure 6: The summarization task presented on Amazon Mechanical Turk.

A.5 Summarization Preference Task Described in Section 5.2

The summarization preference task required participants to make a single choice between the summary generated by GPT-4 and the one produced by humans for each of the 20 selected articles, along with the corresponding assignment (prompt). We enlisted the assistance of 5 participants from AMT and provided them with the task instructions displayed in the image below.



After reading the Educational article and the assignment carefully, please choose the summary that best aligns with the given assignment.

Educational Article: ${text}

Assignment: You are an content creator on social media, creating short educational videos. Summarize the Source text so it results in a script for a short video.

- it should be informal and enthusiastic.
- it should be engaging: it should encourage viewers to watch until the end
- the script should draw the attention of the viewer with a catchy intro.
- It can contain guiding questions and a call to action (e.g. leave a comment) to increase engagement.
- You may use relevant hashtags based on the provided educational content.

Summarized Text 1: ${article_summaries}

○ Summarized Text 1

Summarized Text 2: ${summary}

○ Summarized Text 2

Optional:

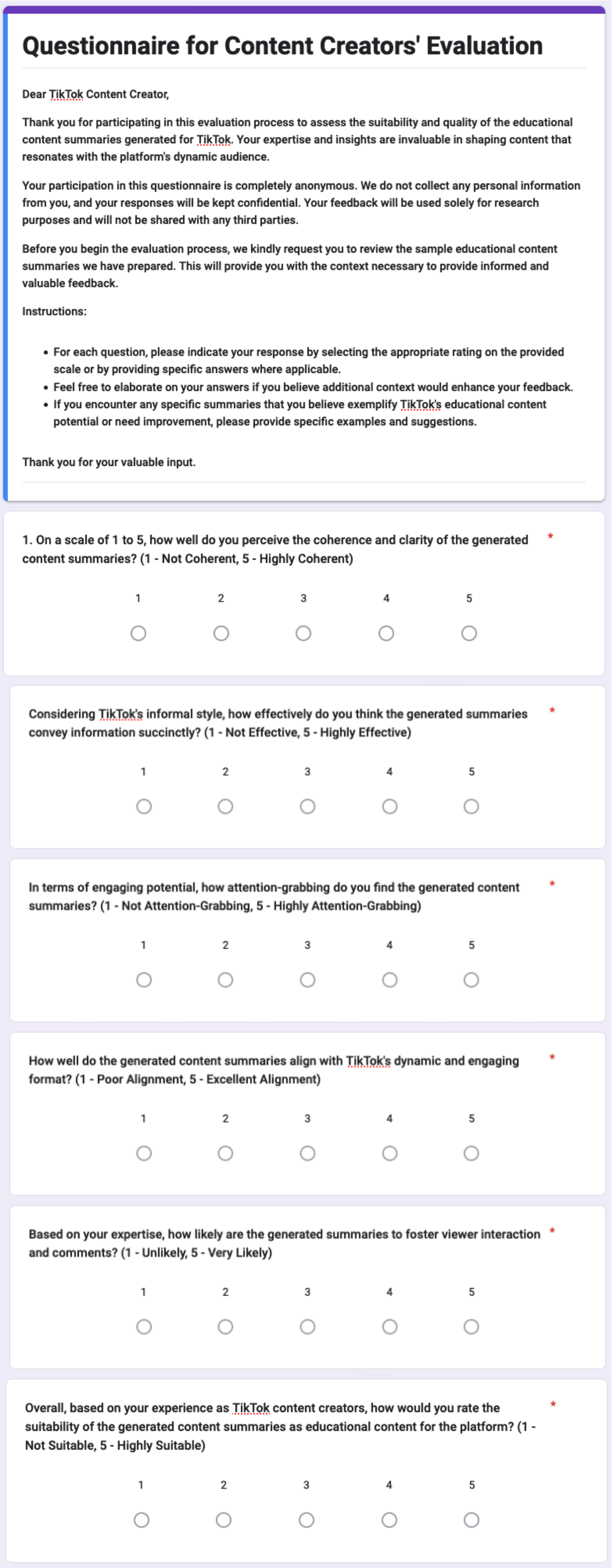
Please write your comments here.

Please make sure that you rate the summaries for all the given criteria.

Thank you for your participation!

Figure 7: The summarization Preference Task

A.6 Questionnaire Evaluation by TikTok Experts Described in Section 5.4



Questionnaire for Content Creators' Evaluation

Dear TikTok Content Creator,

Thank you for participating in this evaluation process to assess the suitability and quality of the educational content summaries generated for TikTok. Your expertise and insights are invaluable in shaping content that resonates with the platform's dynamic audience.

Your participation in this questionnaire is completely anonymous. We do not collect any personal information from you, and your responses will be kept confidential. Your feedback will be used solely for research purposes and will not be shared with any third parties.

Before you begin the evaluation process, we kindly request you to review the sample educational content summaries we have prepared. This will provide you with the context necessary to provide informed and valuable feedback.

Instructions:

- For each question, please indicate your response by selecting the appropriate rating on the provided scale or by providing specific answers where applicable.
- Feel free to elaborate on your answers if you believe additional context would enhance your feedback.
- If you encounter any specific summaries that you believe exemplify TikTok's educational content potential or need improvement, please provide specific examples and suggestions.

Thank you for your valuable input.

1. On a scale of 1 to 5, how well do you perceive the coherence and clarity of the generated content summaries? (1 - Not Coherent, 5 - Highly Coherent) *

1	2	3	4	5
○	○	○	○	○

Considering TikTok's informal style, how effectively do you think the generated summaries convey information succinctly? (1 - Not Effective, 5 - Highly Effective) *

1	2	3	4	5
○	○	○	○	○

In terms of engaging potential, how attention-grabbing do you find the generated content summaries? (1 - Not Attention-Grabbing, 5 - Highly Attention-Grabbing) *

1	2	3	4	5
○	○	○	○	○

How well do the generated content summaries align with TikTok's dynamic and engaging format? (1 - Poor Alignment, 5 - Excellent Alignment) *

1	2	3	4	5
○	○	○	○	○

Based on your expertise, how likely are the generated summaries to foster viewer interaction and comments? (1 - Unlikely, 5 - Very Likely) *

1	2	3	4	5
○	○	○	○	○

Overall, based on your experience as TikTok content creators, how would you rate the suitability of the generated content summaries as educational content for the platform? (1 - Not Suitable, 5 - Highly Suitable) *

1	2	3	4	5
○	○	○	○	○

Figure 8: The questionnaire instructions for content creators evaluation