Memes at Work

How real is synthetic data?

Team Members

Yunyi Zhang (Echo), Elia Meregalli, Zhiyao Luo (Leo)

Contents

Team Members 1
Summary of Key Findings 2
1. Introduction 2
2. Initial Data Sets **3**
3. Research Questions **4**
4. Methodology **4**
5. Findings **9**
6. Discussion **10**
7. Conclusion **1**12
8. References **13**

Summary of Key Findings

Synthetic data for social media research seems to be characterised by an interesting potential, but it is still tarnished by a lack of references to real-world data.

Synthetic data produced by ChatGPT, when verifying the existence of posts containing the hashtags suggested by the LLM, shows that this correlation is extremely inconsistent and that the vast majority of Instagram posts provided by Chat-GPT either do not resolve or do not contain any of the related hashtags it produces. Synthetic data, thus, is in this sense not representative of real-world data, which highlights how ChatGPT seems to focus on semantic similarities when suggesting related hashtags and does not have a grip on the communitarian practices of sharing content on the platform, may them content or market-oriented (positioning).

Nevertheless, synthetic data can still prove analytic usefulness. When comparing automated and manual cluster labelling (after individuating the communities and posts using these hashtags on the platform, or having them generated by ChatGPT), in fact, there is a significant affinity in the labels produced by the chatbot and by researchers (4 out of 6 when compared with the manually coded clusters).

1. Introduction

In recent years, the rise of large language models (LLMs) has sparked growing interest in their ability to generate human-like text, including synthetic, which refers to artificially generated content designed to mimic real-world data. This phenomenon reflects the growing demand for data to train artificial intelligence models. In this sense, synthetic data is particularly relevant as scholars have underlined its potential and possibility to overshadow real data in AI models (Linden 2022).

While LLMs can produce seemingly coherent and contextually relevant outputs, their use raises critical questions about the authenticity and reliability of synthetic data. This issue is particularly relevant in the context of social media research, where platforms like Instagram influence and mirror public discourse, user behaviour, and evolving cultural trends.

On social media platforms such as Instagram, memes have become a crucial medium for expressing and disseminating viewpoints in contemporary society through digital culture (Shifman 2013). Work-related memes, in particular, serve as cultural symbols that users leverage to express their views on labour. In recent years, anti-work sentiments have surged among young people, fuelled by disillusionment with neoliberal work values and a reconsideration of work-life balance. As a result, hashtags like #quietquitting, #antiwork, and #workreform have gained significant traction, becoming focal points in online discussions about labour issues.

Against this background, this project examines the realness of synthetic data by evaluating the extent to which ChatGPT-generated Instagram-related content aligns with real-world social media data. Specifically, we examine whether ChatGPT-generated Instagram-related hashtags, URLs, and meme structures correspond to real-world social media data or are entirely fabricated. The significance of this research lies in the following aspects:

1) The rise of synthetic data in research – LLMs are increasingly used to simulate human behaviour, but their validity as research tools remains an open question.
2) Social media as a cultural archive – Instagram memes provide crucial insights into shifting societal attitudes, particularly in discussions surrounding work and labour. Evaluating whether LLMs can accurately reflect these trends is essential for the future of digital media research.
3) Potential epistemic limitations – If synthetic data significantly diverges from real-world data, its applicability in digital humanities, social sciences, and automated media analysis becomes questionable.

2. Initial Data Sets

This study aims to investigate how real is synthetic data, and how analytically productive it can be when compared to manually produced analysis. To achieve this, we collected two datasets: a real-world dataset gathered manually, and a synthetic dataset generated by ChatGPT. Data collection was conducted on the Instagram platform.

Dataset 1: Synthetic Dataset Generated by ChatGPT:

Construction Method:

An initial set of five hashtags (e.g., #work, #quietquitting, etc.) was supplied to ChatGPT. The model was then prompted to generate additional related hashtags along with corresponding post information.

Data Characteristics:

The resulting synthetic dataset includes supplementary hashtags and post links generated based on semantic associations. It is important to note that these post links do not necessarily correspond to actual content; some links may lead to non-existent posts. This limitation highlights challenges regarding the authenticity and verifiability of synthetic data.

Dataset 2: Real-World Dataset

Construction Method:

The Digital Methods' Zeeschuimer tool (4CAT: Capture and Analysis Toolkit) was employed to collect data from Instagram. By utilizing a dedicated plugin, targeted

hashtag searches were conducted on Instagram, and post data was automatically captured as the page was scrolled downward.

Data Characteristics:

This dataset comprises real Instagram posts, including various metadata such as posts' links, captions, hashtags, and other relevant information. It provides a robust empirical foundation for analysing user behaviour and content dynamics within a real social media environment.

3. Research Questions

Overall Project Focus:
How real is synthetic data?

Research Question:
How accurately do synthetic-related hashtags reflect their presence on Instagram

Sub-questions:
1. How to discuss how synthetic data is real?
2. How does synthetic data relate to real-world data?
3. How analytically productive is synthetic data (when compared to manually produced analyses)?

4. Methodology

Data acquisition

a) GPT Dataset

To obtain the synthetic dataset, we first provided ChatGPT with an initial set of hashtags, including:

- #work
- #quietquitting
- #quietquittinglife
- #workwork
- #antiworkmemes
- #antiworkaholic

We prompted ChatGPT with the question: "can you give me more related hashtags?" To ensure the relevance and validity of the generated data, we employed an iterative prompting approach, asking follow-up questions such as:

- "Where did you get them from? Why do you think they are related?"

- “I need more related hashtags to find memes for analysing social imaginaries at work”
- You are a media scholar studying social imaginaries through memes about work on Instagram. We have a list of hashtags that are about these imaginaries: #work, #quietquitting, #quietquittinglife, #workworkworkworkwork, #antiworkmemes and #antiworkaholic, #workmemes, #officememes, #corporatememes, #workplacehumor, #millennialwork, #genzwork, #remoteworklife, #coworkingmemes, #jobhumor, #workingclassmemes, #grindculture, #overworkculture, #teamnomotivation, #workfromhomememes, #bosslife, #toxicworkculture, #officebanter, #meetingmemes. For each of these hashtags, could you provide 5 hashtags associated with each of them in the same posts containing memes from March 2022 to January 2025 that are relevant to the study of social imaginaries created through memes about work on Instagram? Could you provide one Instagram post that the original hashtag and the related hashtag appeared in? Can you also provide the number of posts, likes, comments, and views per hashtag? Please provide a table with four columns where the first column is the hashtag we provided, the second column is the related hashtag, the third column is the number of posts, likes, comments and views per hashtag in parenthesis after the hashtag and the fourth column is an Instagram post that the original hashtag and the related hashtag appeared in

This iterative method allowed for the systematic expansion of the synthetic dataset.

b) Real-world Dataset

This study utilizes a curated dataset of work-related memes from Instagram. The dataset is constructed using GPT-generated related hashtags, starting with a selection of 17 relevant hashtags.

- #workfromhome
- #anticapitalism
- #workmemes
- #worklifebalance
- #strike
- #wfh
- #mentalhealth
- #antiworkaholic
- #fairwages
- #9to5
- #unionmemes
- #toxicworkplace
- #burnout
- #thegreatresignation
- #worksucks
- #labormemes
- #workermemes

We then used these hashtags to search for related posts on Instagram and collected 100 posts along with their metadata for each hashtag.

Data Visualization & Clustering Analysis

We used the “Table 2 Net” tool to convert the collected CSV dataset into GEXF format, which is compatible with Gephi for network visualization and analysis.
For the hashtags generated by ChatGPT, we mapped the Related Hashtags to the Provided Hashtags and then applied the Modularity Class algorithm in Gephi, dividing the data into 6 distinct communities. These modularity classes help identify different groups or themes within the dataset.

For the manually collected dataset, we also applied the Modularity Class algorithm to categorize the hashtags in 6 distinct communities based on their thematic relevance.

Finally, we ask ChatGPT to cluster the hashtags into 6 communities based on their shared modularity class numbers and interpret the common themes within each cluster using the following prompt:

- I want you to put hashtags in this table into 6 communities based on the common modularity class number. Then label the communities by interpreting the hashtags in the community in an expert way based on their common meanings and shared themes – and produce these labels of the 6 communities. Each label should contain more than 1 word but less than 10. Provide also an explanation on why you came up with certain labels.

Fig.1- ChatGPT Data

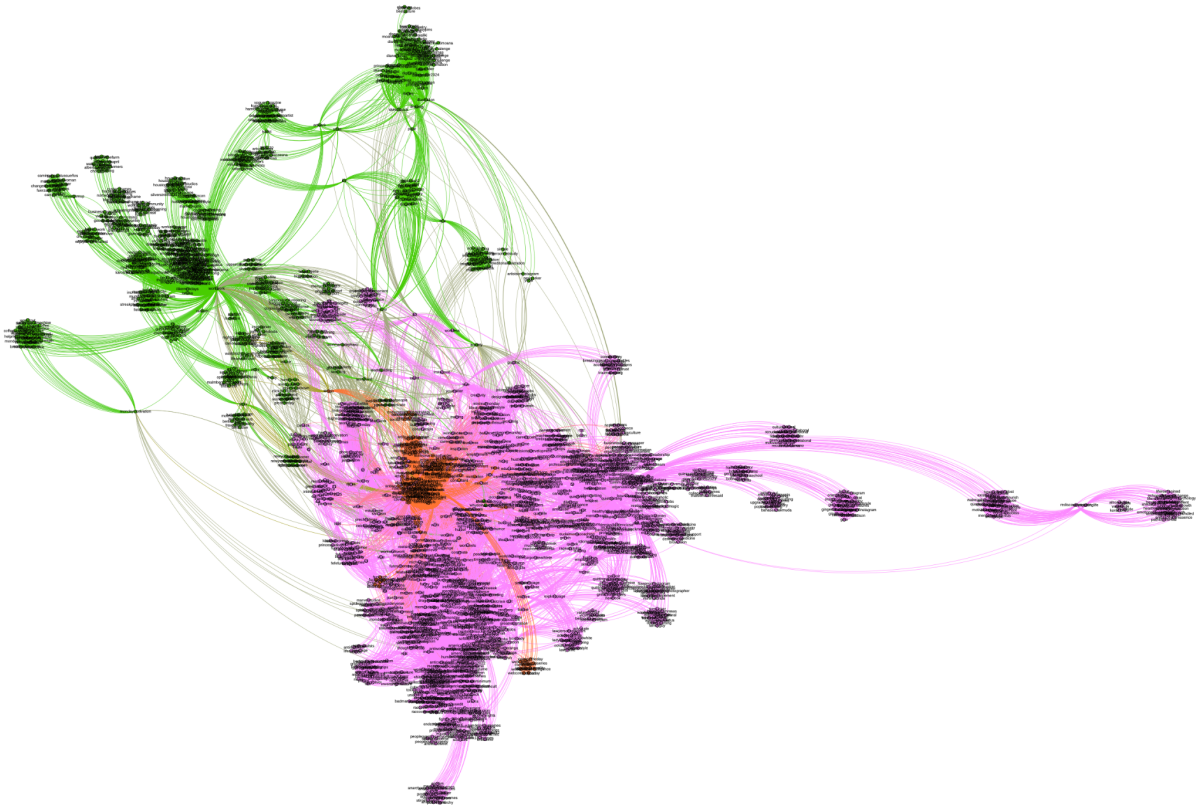

Fig.2- Manual Data

Data Comparison

We manually aligned the themes of the Manual Data and GPT Data clusters as closely as possible. The highlighted sections represent areas that could not be matched. 4 out of 6 clusters were successfully aligned.

Manual Data	**GPT Data**
Corporate and Workplace Humour #work, #memes, #antiwork, #officehumor, #workmemes, #quietquitting	Office Humour and Workplace Relatability #antiworkmemes, #corporatememes, #coworkingmemes, #jobhumor, #officebanter, #officememes, #teamnomotivation, #work, #workhumor, #workmemes, #jobmemes, #coworkinghumor, #coworkers, #workingclassmemes, #laborhumor, #officehumor, #workproblems, #workplacehumor
Freelance Work and Marketing #freelance, #digitalmarketing, #socialmedia,	Remote Work and Digital Nomadism #remoteworklife, #digitalnomad, #remotework,

#workremote	#remoteworker, #wfh, #workfromhome
Motivation and Self-Improvement #motivation, #goals, #selfemployed, #worksmarternotharder	Remote Work and Virtual Office Life #meetingmemes, #workfromhomememes, #remoteworkmemes, #wfhmemes, #quarantinememes, #workfromhomehumor, #zoommemes
Anti-Work Sentiment #fuckwork, #antiwork, #nojob, #workreform	Work Culture Critique and Self-Care #antiworkaholic, #grindculture, #overworkculture, #quietquitting, #quietquittinglife, #toxicworkculture, #burnout, #burnoutprevention, #worklifebalance, #stressatwork, #workhard, #selfcareatwork, #hustleculture, #slowliving, #mindfulwork, #unhealthyworkenvironment, #workstress, #workplacebullying, #badmanagement
-	Millennial and Gen Z Work Perspectives #genzwork, #millennialwork, #genzhumor, #genzlife, #genzproblems, #millennialhumor, #millenniallife, #millennialproblems, #worklife
Regional and Local Events #群馬県, #伊勢崎市, #観光特使, #農業まつり	-
Art and Creative Expression #art, #drawing, #digitalart, #procreate	-
-	Entrepreneurial Mindset and Hustle Culture #bosslife, #ceolife, #businessowner, #hustle, #entrepreneurlife, #grindculture, #workworkworkworkwork, #workhardplayhard, #leadership

Table.1- Comparison of 6 Communities in Manual Data vs GPT Data

5. Findings

Our research suggests the presence of relevant differences when confronting the manually collected data on the platform and the synthetic data generated by the model. These findings underline how ChatGPT incurs limitations when trying to reproduce social-media activity on Instagram.

The vast majority of Instagram posts provided by ChatGPT either do not resolve or does not contain any of the related hashtags it produces. In this sense, synthetic data does not match its real-world referent. This element suggests that as the model does not possess access to the metadata of the platform, it seems to base its suggestions on semantic components or previously recorded patterns in its training dataset. As the use of hashtags on social media is not only based on these mechanics but incurs into a series of contextually specific influences, reflected in the platform's affordances and logics, the result is a sensible lack of efficacy in reflecting real-world data.

There is, then, a second consideration to be discussed. This report seeks to not only reflect on the 'real' or 'non-real' evaluation posed on synthetic data; it additionally inquires about the alignment between analyses of synthetic data with real-world data. In essence, it reflects on whether synthetic data can prove to be analytically useful despite its lack of adherence to real-world data. Undertaking both automated and manual analyses of the hashtags and related hashtags provided by ChatGPT and manually retrieved on Instagram, we produced co-hashtag graphs. The clusters highlighted by the graphs were labelled automatically for the synthetic dataset and manually for real-world data. Results show that, on average, 4 out of the 6 thematic clusters generated by the model had strong connections and relatedness to the coded clusters which were manually identified, based on fifteen separate analyses. Thus, our reflections show that synthetic data, although quite imprecise when matching its real-world referent, does show strong affinities in terms of labelling.

6. Discussion

Significance

This study conducts a real-time analysis of work meme culture on social media, revealing deficiencies in the authenticity and timeliness of synthetic data. When generating related hashtags, ChatGPT primarily relies on semantic similarity and fails to accurately predict co-occurrence patterns on social media, lacking an understanding of real social media interactions. This research direction not only helps evaluate the applicability of LLMs in social media studies but also provides optimization ideas for future AI development.

On the other hand, synthetic data still holds exploratory value in certain research contexts. Particularly in the early stages of research, synthetic data can be used to construct preliminary topic classifications, assisting researchers in identifying potential research directions. For example, when analysing social media discussions related to work culture, ChatGPT-generated hashtags can provide researchers with an initial conceptual framework, allowing them to quickly identify potentially important topics. At the same time, ChatGPT has demonstrated some potential in hashtag classification tasks. In this study, its generated hashtag classifications were highly consistent with manually annotated results in four out of six major categories. This suggests that LLMs

can provide researchers with some assistance in topic classification and conceptual mapping, but their conclusions still need to be validated with real data.

Limitations

Despite following the characteristics and evolution of the internet by leveraging its native and dynamic data, this study still has certain limitations. First, as Richard (2013) noted regarding digital methods, data on digital platforms may be influenced by algorithms, user behaviour, and platform policies, leading to data that does not fully represent real-world conditions. S. Banerjee and G. A. Veltri (2024) also emphasized that while LLMs can fill data gaps, they cannot capture spontaneous behaviours or unique social dynamics on social media.

This study only examines data generated by ChatGPT and does not compare the performance of other LLMs on the same task. Different LLMs may vary in terms of data generation accuracy, co-occurrence pattern prediction, and contextual understanding. Furthermore, excessive reliance on digital data may overlook the deeper background and complexity of social phenomena. Due to the limitations of data collection tools, we were unable to obtain large-scale, systematic social media data. Therefore, the real dataset in this study was collected through manual and partially automated tools. This may introduce certain biases into the dataset, affecting the comparison between ChatGPT-generated data and real data due to differences in data scale and sampling strategies. Additionally, since popular hashtags on social media may change over time, the findings of this study may not be applicable to all time periods but rather to specific time windows.

Insights

Future research can further explore multiple directions to optimize the application of LLMs in social media studies. First, integrating LLMs with fine-tuning based on social media data could enable dynamic adjustments according to real social media trends. Another promising research direction is examining the ability of different LLMs to simulate social media data, such as their accuracy in generating social media hashtags (i.e., whether they match real co-occurrence patterns), their ability to simulate social dynamics (i.e., whether they can capture trending topics, community interaction patterns), and their potential applications in social sciences and digital humanities (i.e., whether they can assist in public opinion research, text analysis, and network communication studies).

Finally, future research should also focus on potential biases in synthetic data generation by LLMs. Since ChatGPT's training data may contain historical biases, its generation of social media-related data may amplify dominant narratives while ignoring the voices of minority groups. This phenomenon is similar to what B. M. Guțu and N. Popescu (2024) pointed out in their study on AI bias—LLM-generated social media data may exaggerate trends, overlook marginalized groups, and lead to biases. Therefore, in future social media research, researchers need to explore how to introduce fairness

mechanisms in the data generation process of LLMs to reduce bias and improve data representativeness.

7. Conclusion

This project reflected on the nature of the relationship between synthetic and real data, a phenomenon becoming the object of focus for many researchers with the development of efficient and powerful AI models such as ChatGPT. Specifically, it inquired about the accuracy with which synthetic related hashtags suggested by the model relate to the real-world data gathered on Instagram.

Our findings suggest that, when referring to the representation of work cultures through memes on Instagram, synthetic data is still a contested object. In particular, it contemporarily showed validity for thematic analysis whilst demonstrating serious limitations when representing real-world social media dynamics. The model lacks access to the metadata contained in the Platform, and it subsequently has to leverage connections based on semantic similarities rather than real-world social media dynamics, thus ignoring the context-specific implications of Instagram posting. At the same time, the significant affinity demonstrated between thematic clusters labelled automatically (by the model) and manually (by researchers) suggests that the generated synthetic data has analytical validity.

Further research would need to expand on the possibilities offered by synthetic data. The lack of access to metadata remains a fundamental challenge, but there can be productive avenues in mixed approaches where researchers are able to leverage LLMs' capacity for data production and analysis. Additionally, research could focus on the distance between real-world and synthetic data itself, to understand how and why AI models understand and reproduce determinate online social dynamics.

8. References

Banerjee, S. and Veltri, G.A. (2024). Harnessing pluralism in behavioral public policy requires insights from computational social science. *Frontiers in Behavioral Economics*, *3*, p.1503793.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Guțu, B. M., & Popescu, N. (2024). Exploring Data Analysis Methods in Generative Models: From Fine-Tuning to RAG Implementation. *Computers*, *13*(12), 327.

Linden, A. (2022). Is synthetic data the future of AI?. *Gartner*, June 22. https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic data-the-future-of-ai.

Rogers, R. (2013). *Digital methods*. The MIT Press.

Shifman, L. (2013). *Memes in Digital Culture*. The MIT Press.