Aarhus2025 Synthetic Data Representation Workshop Submission

Reverse construction of the synthetic datasets

Large Language Models (LLMs) are designed to combine textual analysis and neural networks as two fundamental sources for their processing. They are trained on datasets of texts from both digital archives and "internet data." The latter is not a well-defined set of sources, consisting of texts posted online by individuals and organizations as well as their (our) digital traces. Together with digital archives, these are taken (capta) from the services where people and organizations act and/or perform themselves online. This vagueness of the sources enables the “question of legitimacy and situatedness of knowledge”, and inspires scholars to develop the approaches based on segmentability and representativeness, as well as curation of the data (Rossi, Harrison, Shklovski, 2024). In this submission I try to develop these approaches and add theoretical resources and methodological experiments in order to suggest the *reverse construction of datasets* as a way of understanding LLM-based instruments.

The datasets based on online sources are presumed to represent the "real" way of argumentation, action and representation of knowledge. However, as we know from the field of Internet studies (boyd, 2014 Marwick, 2013, Abidin, 2018), these actions are part of self-performance that occur in particular situations. We can describe them as framed (in the terms of Erving Goffman, 1958) or situated (if we follow Lucy Suchman, 1987) actions. None of these conceptualizations implies that they are "raw data" about people's actions and words; they need to be understood within the specific context of their genealogy, structure and situatedness.

However, the LLM approach treats data from sources such as social networking sites, blog texts, or books equally as "text." The methodological approach to these texts is based on linguistics, behavioral sciences, and pattern recognition, meaning that the texts become an equal source of "information." This information is recognized as text. Therefore, the frame, situation, and other properties are assumed to be part of genre, rhetorical structures, reasoning via analogy or argumentative strategy (Webb et al 2024). These features are intended to be part of the dataset for further "information" retrieval and production.

I follow the idea (introduced primarily by Gilbert Simondon, 1958 but also developed by other theoreticians) that "information" is a verb, a continuous sense-making process. If so, we can argue that the synthesis of internet-based data is continuously including the contexts of its production. These contexts are not just metadata. They need to be unfolded as the frame, the situation, the interface, and the communicative space where they were produced. For instance, an internet forum is different from a social networking site, a blog, or an online shop review. People act differently in these online spaces. Moreover, we are not sure the metaphor of space works equally well for all the above-listed environments. We might consider it as an instrument or a way of being (Markham, 2003).

This critical reconstruction suggests that the synthetic structure of LLM datasets and their development needs a reversive methodological revision. We propose to reconstruct the rules of synthetic data based on experiments with several LLMs as well as a qualitative research of reviews that are thought to be the basis for their training.

The empirical part of the research is inspired by Digital Humanities projects on classical Russian literature. These projects (Slovo Tolstogo, Pushkin Digital) are based on academic archives. They include

not only the texts by these authors but also critical commentaries and literary studies. There are also numerous other data sources about these classical authors, notably readers' reviews. We include two main sources for additional understanding: Reddit.com and book reviews on bookshop websites (Amazon and the Russian shop Ozon). These online venues are platforms for sharing readers' emotions and thoughts about books. We focus specifically on "Eugene Onegin" by Alexander Pushkin and "Anna Karenina" by Leo Tolstoy.
There are two methodological parts of the research. The aim of the first one, the internet studies research, is to systematize the rules and situations of readers' reviews. We identify their frames as well as provide rich descriptions of the interfaces. The same methodology is applied to Digital Humanities websites. Considering all these findings, we conduct a series of experimental conversations with LLM-based services (Perplexity and DeepSeek). This second part of the research is a series of conversations aimed at revealing the rules of data systematization and determining whether we can identify patterns that help us distinguish particular data sources. We initiate conversations about the content and different perspectives on understanding the books (from professionals, scholars, ordinary readers in various situations, etc.).

The aim of the research is to confirm (or refute) the suggestion that the data sources of LLM databases are distinguishable through conversations. This is still ongoing research. Whether the experimental conversations/prompting will be successful or not, preliminary fieldwork indicates that the "representativeness" of data later reassembled by LLMs and tools based on it is not a black box. The theoretical and experiment-based reconstruction demonstrates that it is possible to develop a reverse construction approach for internet-based datasets.
The first step is to reveal the genre or other linguistic elements of the source that are suggested by the conversational LLM tool. The second step is to determine the frame and/or situation of genre production. This approach does not promise a stable result but is helpful for the methodological analysis of the information provided by LLM-based tools and suggests further development.

Further investigation proposes the analysis and revision of "web ontologies" based on datasets that are available for understanding the results of LLM operation. This research can enhance our approach to the "representativeness" of data and introduce a theoretically grounded and socially applicable approach to AI tools.

References:

1. Rossi, Luca, Harrison, Katherine, Shklovski, Irina. (2024) *The Problems of LLM-generated Data in Social Science Research.* Sociologica 18.2 (2024): 145-168.
2. boyd, danah. (2014). *It's Complicated: The Social Lives of Networked Teens*. Yale University Press.
3. Marwick, Alice E. (2013). *Status Update: Celebrity, Publicity, and Branding in the Social Media Age*. Yale University Press.
4. Abidin, Crystal. (2018). *Internet Celebrity: Understanding Fame Online*. Emerald Publishing Limited.
5. Goffman, Erving. (1959). *The Presentation of Self in Everyday Life*. Anchor Books.
6. Suchman, Lucy A. (1987). *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press.
7. Simondon, Gilbert. (1958). *Du Mode d'Existence des Objets Techniques*. Aubier.

8. Markham, Annette N. (2003). *Life Online: Researching Real Experience in Virtual Space*. AltaMira Press.
9. Webb, Taylor, Keith J. Holyoak, and Hongjing Lu. (2023) *Emergent analogical reasoning in large language models*. Nature Human Behaviour 7.9: 1526-1541.