Exploring the potential of large language models (such as GPT-4) for (semi-)automatic content analysis of stances and frames in media texts

The use of pre-trained language models (PLMs) based on transformer neural networks has significantly advanced the field of natural language processing (NLP) and offers considerable potential for automatic content analysis of increasingly complex semantic concepts. Nevertheless, due to computational complexities associated with the use of pre-trained language models and the lack of best practices (and necessary programming skills) within the field of communication science, the discipline has been slow to adopt these methods (for exceptions see, for example, Laurer, 2023). The sudden popularity of ChatGPT, a chatbot allowing users with no programming knowledge to make use of large language models (LLMs) for a great variety of tasks, opens the question whether communication science can use such systems for content analysis in what is called a "zero-shot" approach in NLP (Gilardi et al., 2023; Huang et al., 2023). This would mean to simply code large scale sets of media (text) data using prompts, instead of the more complicated pretraining procedures that until recently were currently the gold standard in NLP. First experiments with adopting ChatGPT for various text coding tasks (Liu et al., 2023) raises doubts whether the conventional practice of relying on human coders in content analysis will remain preferable in the future – both from the perspectives of resources required to generate data, as well as its ultimate quality.

In our contribution, we aim to address this question by presenting the results of our methodological study comparing the potentials and weaknesses of state-of-the-art NLP method using LLM in the context of two tasks for which content analysis is often employed in communication science: (1) identification of stances towards a controversial policy issue using the concepts of a claim and argument, and (2) discerning interpretative repertoires (aka "micro frames") in the news media discourse.

We combine established PLMs with parameter-efficient fine-tuning (PEFT, cf., Hu et al., 2022) and few-shot (FS, cf., Rieger et al., forthcoming) methods to obtain efficient and performant models using as little pre-coded data as possible (e.g. 30 training examples per frame). We are able to show that this recipe (PLM+PEFT+FS) leads to better quality as well as better reliability and reproducibility of the results and to less computational costs compared to the common recipe of fully fine-tuning PLMs using a standard classification head.

We then explore the potential of using GPT-4 (the LLM used by ChatGPT) "off the shelf" for coding (cf., Ding et al., 2023) stances and frames in media texts either without any pre-coded training examples (zero shot) or with a small number of examples per category in the prompt. By experimenting with different prompts, we can also show how this impacts the quality of coding. The results indicate that the use of GPT-4 as a component in the coding process does show some promise. The model demonstrates – especially using intelligently chosen prompts – a basic understanding of our definitions of stances and frames, concepts that are not that easily identified by human coders either (in our case, it needed three rounds of training to achieve a minimum Krippendorffs alpha of 0.66). Similarly, while the coding of the stance towards weapons deliveries using zero-shot GPT-4 is satisfactory (an F1 score of 0.8), the quality of the coding varied greatly between the different frames. Here, our fine-tuned models (still) performed significantly better.

Our contribution thus aims to help communication scientists understand the potential (but also the limits) of the use of both fine-tuned PLM with few-shot learning and general LLMs like ChatGPT with zero-shot learning for common tasks in media content analysis. But the

contribution will also address the valid methodological and ethical concerns with using ChatGPT: For the moment, ChatGPT is a "black box", with the underlying LLM and the chatbot being constantly tweaked, greatly impacting questions of reliability and reproducibility. Biases within the training data of GPT-4 may also impact the coding tasks in unknown ways.

References:

Ding, B., Qin, C., Liu, L., Chia, Y. K., Li, B., Joty, S., & Bing, L. (2023). Is GPT-3 a Good Data Annotator? *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11173–11195. https://doi.org/10.18653/v1/2023.acl-long.626

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*. https://doi.org/10.48550/ARXIV.2303.15056

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. (2022). LoRA: Low-Rank Adaptation of Large Language Models (LoRA). *ICLR 2022 Conference proceedings*. https://openreview.net/forum?id=nZeVKeeFYf9

Huang, F., Kwak, H., & An, J. (2023). *Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech*. https://doi.org/10.48550/ARXIV.2302.07736

Laurer, M., Van Atteveldt, W., Casas, A., & Welbers, K. (2023). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis,* 1-17. doi:10.1017/pan.2023.20

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). *Summary of ChatGPT-Related Research and Perspective Towards the Future of Large Language Models*. https://doi.org/10.48550/ARXIV.2304.01852

Rieger, J., Yanchenko, K., Ruckdeschel, M., von Nordheim, G., Kleinen-von Königslöw, K. and Wiedemann, G. (forthcoming). Few-shot learning for automated content analysis: Efficient coding of arguments and claims in the debate on arms deliveries to Ukraine. Accepted for Studies in Communication and Media.