Empowering citizens for AI: Assessing public’s (mis)conceptions about Large Language Models

Maria Zangl, Michael Bedek & Dietrich Albert

Cognitive Science Section, University of Graz

Intercultural Workshop on Data Sovereignty and Generative AI, Informatik Festival 2024

24/09/2024

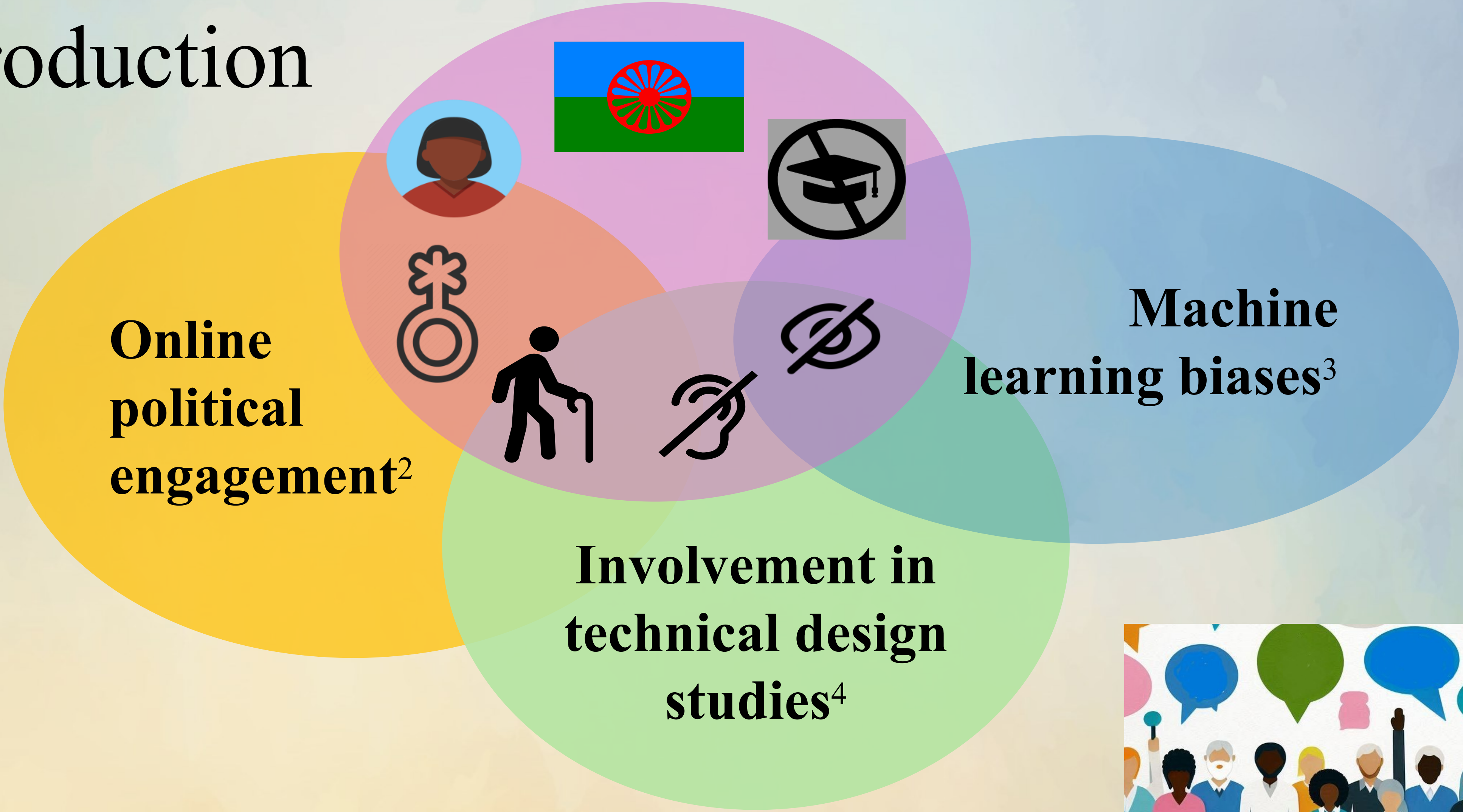
Project

Main objectives of ITHACA[1]:

- Develop and test a civic engagement platform
- Integration of AI applications
- Ensure accessibility & usefulness for all

1 EU Project (2023-2026): https://www.ithaca-project.eu/project/; https://cordis.europa.eu/project/id/101094364

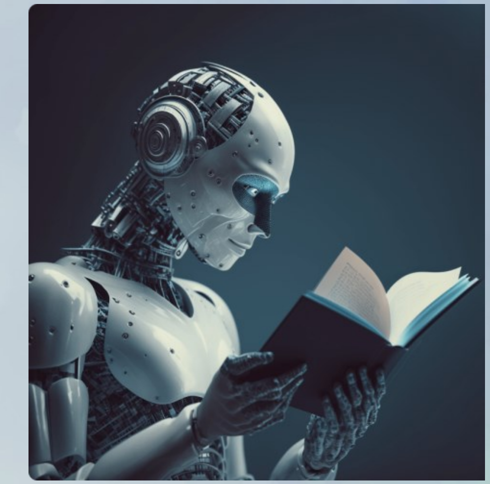
Introduction



Large language models (LLMs) for inclusivity[5]

2 Davies & Procter, 2020; Guldvik et al., 2013 ; 3 Wang et al., 2024; Kuran et al., 2020; Krupiy, 2020 4 Fischer et al., 2020; 5 Garcia Valencia et al. 2024; picture created with Microsoft's AI image generator

Introduction



- Misunderstanding LLMs
  → misuse, privacy risks & over-reliance[6]
- Experts believe the public to have misconceptions[7]
- Measurements of public‘s knowledge about LLMs are self-assessments[8]
- We need an **objective assessment of (mis)conceptions about LLMs**

6 Weidinger et al., 2022;, Navigli et al., 2023; 8 Bewersdorff et al., 2023; Henestrosa & Kimmerle, 2023; Amaratunga, 2023; 7 Bodani et al., 2023, Henestrosa & Kimmerle, 2023; Lee & Park, 2024, picture: https://www.turingcollege.com/playbooks/chatgpt-in-education

Research Questions

- **Technical** requirements for an accessible civic participation platform from diverse groups? (Study I)
- **Individual** knowledge needed for beneficial AI use? (Study I)
- Public‘s misconceptions about LLMs? (Study II)
- Do publics‘ (mis)conceptions about LLMs‘ have underlying prerequisite relations? (Study III)

Questionnaire study and focus groups (Study I): Exploring technical and individual requirements

Development of a technical requirement questionnaire

Technical requirements of
vulnerable group members (*N*=39) & municipality employees (*N*=35)

Discussion of with municipality managers (*N*=10)

Discussing **individual requirements** for AI applications with AI experts (N= 6 / 7 per session)

Outcomes of technical requirements (Study I):

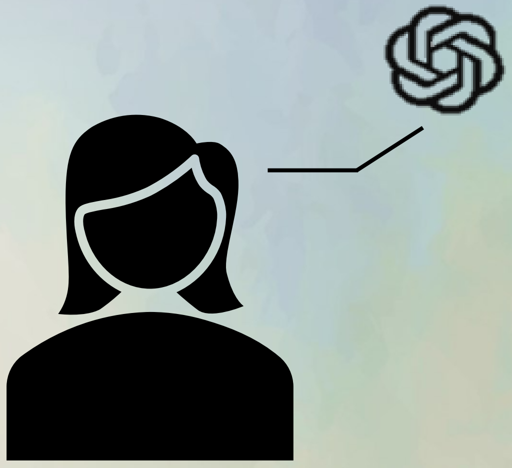
Highly desired technical applications:

- Chatbot for interaction
- Language translation
- Language simplification
- Text-to-Speech / Speech-to-Text

Picture: https://chatgpt.com/

Semi-structured interviews (Study II)

Exploring knowledge & misconceptions about ChatGPT



- What misconceptions does the public have?
- Actual misconceptions vs. experts‘ assumptions?

Semi-structured interviews (Study II):

~15 citizens who have

- at least heard of ChatGPT
- not received AI-/ data science education

Qualitative content analysis[10]: Category formation & assignment

Expected outcome:

- List of correct conceptions
- List of misconceptions

10 Mayring, 2014; Friese, S., 2019; picture created with Microsoft's AI image generator

Deducing (mis)conceptions from interviews (Study II):

(Observable) Statements from interviews

‘ChatGPT has a deep understanding about the topics it’s trained with’

‘If ChatGPT has no information stored in its database about a topic, it will inform me about this limitations’

Underlying (Mis)conceptions

- Trained on data (✓)
- Deep understanding (✗)
- Stores pre-recorded info about topic in a database(✗)
- Informs about its limitations (✗)
- Has limitations (✓)

Conceptualization of an adaptive assessment method
(Study III)

1) Identifying a theoretical concept structure of (mis)conceptions

Knowledge Space Theory[12] extensions:

competence-performance approach[13] & modeling misconceptions[14]

→formal foundations modeling through prerequisite relations of knowledge components[15]

12 Falmagne et al., 1990; 13 Korossy, 1997; Stefanutti et al., 2023; 14 Lukas, 1997; Stefanutti et al., 2020 15 Doignon & Falmagne, 2012

Theoretical concept structure (Study III):

(Mis)conceptions

Trained on data (✓)

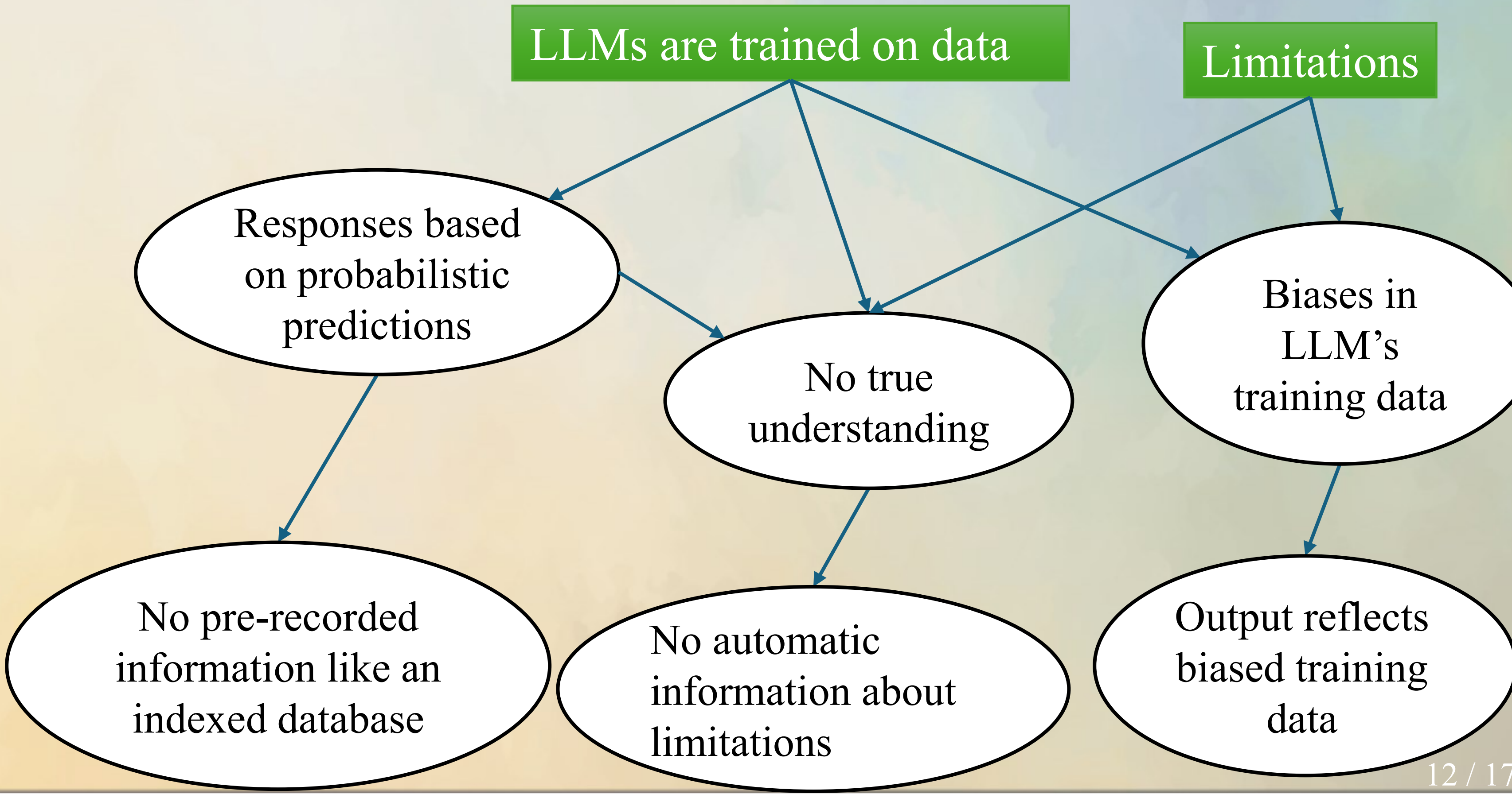
Deep understanding (✗)

Stores pre-recorded information about topic in a database (✗)

Informs about its limitations (✗)

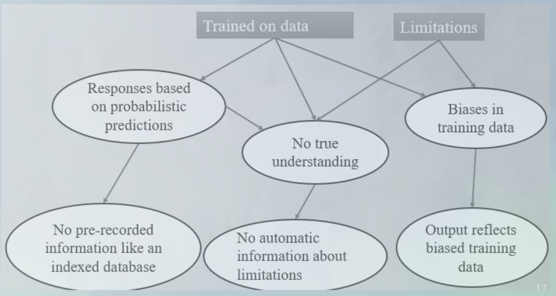
Has limitations (✓)

Potential Prerequisite Relations



Conceptualization of an adaptive assessment instrument (Study III):

1) Identifying a theoretical concept structure of (mis)conceptions



2) Item construction

- Based on identified (mis)conceptions & technical state-of-the-art

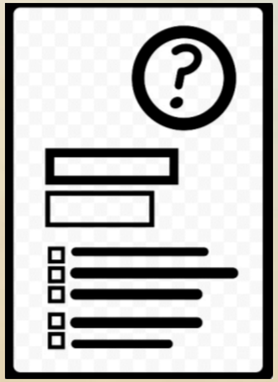
Trained on data (✓)

Deep understanding (X)

Stores pre-recorded information about topic in a database (X)

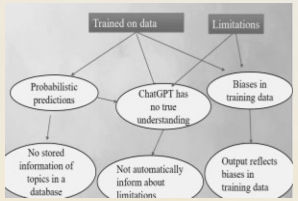
Informs about its limitations (X)

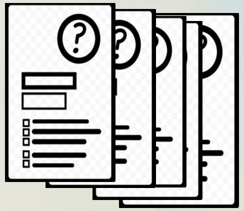
Has limitations (✓)



Validation of the theoretical knowledge structure (Study III)

3) Validating the theoretical concept structure with empirical items responses[16]





- H_1: Items with a higher level of complexity are solved less frequently than items with a lower level of complexity.
- H_2: The knowledge (correct conceptions and misconceptions) about LLMs (empirical response patterns) follows the assumed theoretical structure.

16 e.g. Kump, 2006; Ley 2006

Open Questions

ML for identifying a theoretical concept structure?

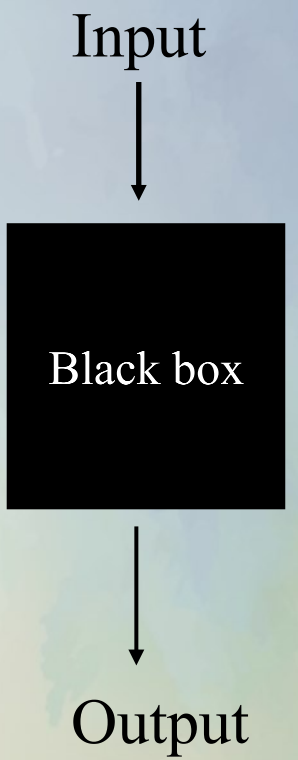
- Prerequisite relations in online learning courses using LSTM-based neural networks[17]

- Knowledge graph construction using keyphrase extraction & sentence encoders[18]

17 Xiao et al., 2021; 18 Alatrash et al., 2024,

Challenges

- Uncertainties in LLMs‘ mechanisms
  - Communicate the blackbox of exact mechanisms
- Technological advances
  - Resistant to consistent model developments



Get in touch

Thank you!

The Cognitive Science Section

ITHACA-Project Information:

maria.zangl@uni-graz.at

michael.bedek@uni-graz.at

dietrich.albert@uni-graz.at

References

Bewersdorff, A., Zhai, X., Roberts, J., & Nerdel, C. (2023). Myths, mis- and preconceptions of artificial intelligence: A review of the literature. Computers & Education: Artificial Intelligence. , Article 100143. https://doi.org/10.1016/j. caeai.2023.100143

Sulmont, E., Patitsas, E., & Cooperstock, J. R. (2019). Can you teach me to machine learn? In E. K. Hawthorne, M. A. P´erez-Qui˜nones, S. Heckman, & J. Zhang (Eds.), Proceedings of the 50th ACM technical symposium on computer science education (pp. 948–954). ACM. https://doi.org/10.1145/3287324.3287392.

Mayring, P. (2014).*Qualitative content analysis: theoretical foundation, basic procedures and software solution*. SSOAR, [Online]. Available: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-395173

H.-D. Dann, Subjective theories and their social foundation in education. Hogrefe & Huber, 1992, pp. 161–168.

Smith III, J. P., DiSessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The journal of the learning sciences*, *3*(2), 115-163.

Henestrosa, A. L., & Kimmerle, J. (2023). Understanding and Perception of Automated Text Generation among the Public: Two Surveys with Representative Samples in Germany.

Guldvik, I., Askheim, O. P., & Johansen, V. (2013). Political citizenship and local political participation for disabled people. *Citizenship Studies*, *17*(1), 76–91. https://doi.org/10.1080/13621025.2013.764219

Wegscheider, A. (2013). Politische Partizipation von Menschen mit Behinderungen. SWS-Rundschau, 53(2), 216-234. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-436995

Falmagne, J. C., Koppen, M., Villano, M., Doignon, J. P., & Johannesen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review*, *97*(2), 201.

Wang, C., Boerman, S. C., Kroon, A. C., Möller, J., & H de Vreese, C. (2024). The artificial intelligence divide: Who is the most vulnerable? New Media & Society, 0(0). https://doi.org/10.1177/14614448241232345

Kuran, C. H. A., Morsut, C., Kruke, B. I., Krüger, M., Segnestam, L., Orru, K., ... & Torpan, S. (2020). Vulnerability and vulnerable groups from an intersectionality perspective. *International Journal of Disaster Risk Reduction*, *50*, 101826. https://www.sciencedirect.com/science/article/pii/S2212420920313285

Krupiy, T. T. (2020). A vulnerability analysis: Theorising the impact of artificial intelligence decision-making processes on individuals, society and human diversity from a social justice perspective. *Computer law & security review*, *38*, 105429. https://doi.org/10.1016/j.clsr.2020.105429

Lee, U., Jung, H., Jeon, Y. *et al.* Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Educ Inf Technol* (2023). https://doi.org/10.1007/s10639-023-12249-8

Rong, Q., Kong, W., Xiao, Y., Gao, X. (2023). An Adaptive Testing Approach for Competence Using Competence-Based Knowledge Space Theory. In: Anutariya, C., Liu, D., Kinshuk, Tlili, A., Yang, J., Chang, M. (eds) Smart Learning for A Sustainable Society. ICSLE 2023. Lecture Notes in Educational Technology. Springer, Singapore. https://doi.org/10.1007/978-981-99-5961-7_18

Ley, T., Kump, B., & Albert, D. (2010). A methodology for eliciting, modelling, and evaluating expert knowledge for an adaptive work-integrated learning system. *International Journal of Human-Computer Studies*, *68*(4), 185-208.

References

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In Advances in Experimental Social Psychology, 44, pp. 247–296). Elsevier. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Fischer, B., Peine, A., & Östlund, B. (2020). The importance of user involvement: a systematic review of involving older users in technology design. *The Gerontologist*, *60*(7), e513-e523. https://doi.org/10.1093/geront/gnz163

Almatrafi, O., Johri, A., & Lee, H. (2024). A Systematic Review of AI Literacy Conceptualization, Constructs, and Implementation and Assessment Efforts (2019-2023). *Computers and Education Open*, 100173.

Friese, S. (2019). Qualitative data analysis with ATLAS. ti. Qualitative data analysis with ATLAS. ti, 1-344

Zhang, H., Chuhao, W., Jingyi, X. Yao L., Jie, C. and Carroll, J. M. (2023) Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. College of Information Sciences and Technology, Penn State University, USA. https://doi.org/10.48550/arXiv.2309.10771

Stefanutti, L., de Chiusole, D., Gondan, M., & Maurer, A. (2020). Modeling misconceptions in knowledge space theory. *Journal of Mathematical Psychology*, *99*, 102435.

Stefanutti, L., Spoto, A., Anselmi, P., & de Chiusole, D. (2023). Towards a competence-based polytomous knowledge structure theory. *Journal of Mathematical Psychology*, *115*, 102781.

Garcia Valencia, O. A., Thongprayoon, C., Miao, J., Suppadungsuk, S., Krisanapan, P., Craici, I. M., ... & Cheungpasitporn, W. (2024). Empowering inclusivity: improving readability of living kidney donation information with ChatGPT. *Frontiers in Digital Health*, *6*, 1366967.

Alatrash, R., Chatti, M. A., Ain, Q. U., Fang, Y., Joarder, S., & Siepmann, C. (2024). ConceptGCN: Knowledge concept recommendation in MOOCs based on knowledge graph convolutional networks and SBERT. *Computers and Education: Artificial Intelligence*, *6*, 100193.

BACK UP

Identifying socially vulnerable & marginalized individuals (Study I)

vulnerable / marginalized social groups

- Elderly / pensioners (e.g. 60+)
- Younger / youth (i.e. 18-30)
- Refugees & Migrants
- Roma people
- People with physical disabilities
- People with mental disabilities and/or problems (e.g. depression, addiction, etc.)
- People in rural areas
- Homeless people
- People of colour
- Women (Pregnant women &) women with young children
- Families with many children (e.g. >3)
- Single parents
- LGBTQIA+

vulnerability criteria / factors

- low income
- poor living conditions
- precarious employment / (repeated) unemployed
- lack of insurance
- low educational background
- limited educational opportunities
- low digital literacy and/or particular need for support wrt to digital platforms
- limited access to infrastructure / mobility
- limited access to cultural program and information,
- social isolation / loneliness
- structural /systematical discrimination concerning participation
- facing physical and/or verbal violence

Technical requirements questionnaire (Study I)

1) What features / functionalities would you want to have?

Possibility	I would like to have that	I cannot imagine it	I don't want to have that
Recommendations (Recommends topics or posts that might interest you, based on other posts/ topics you liked or commented on)	O	O	O
Sentiment Analysis (Determines the sentiment (positive, negative, or neutral) expressed in discussions.)	O	O	O
Toxicity sensor / Spam-/Phishing-Post detection (Detects harmful or toxic behaviors or texts in posts, chats or comments and can prevent discrimination, threats or harassments.)	O	O	O
Automated reporting and analyzation (Provides statistics about your engagement and impact of your or others' posts to understand the effectiveness of their contributions.)	O	O	O
Multimedia posts (Posting and watching videos, photos, locations or voice recordings)			

2) How should YOUR posts / comments / contributions be rated/commented by others?

Possibility	I would like to have that	I cannot imagine it	I don't want to have that
Rating (e.g. by 1-5 stars)	O	O	O
Commenting	O		
Upvoting and Downvoting posts	O		
Upvoting only	O		
Reacting with emojis	O		

3) A platform with many users and many contributions related to different topics might get chaotic or confusing very soon. **How to ensure that you get those posts/ topics that interest YOU the most?**

Possibility	I would like to have that	I cannot imagine it	I don't want to have that
Search bar (Search for content, topics, posts or tags within the entire platform by entering a keyword or query in the search bar).	O	O	O
Filter options (Filters posts or topics based on their date, location, length, popularity, …)	O	O	O
Subscriptions /Abonnements ("Following" either Users or Topics)	O	O	O
Email notifications / notification at your profile (Get regularly emails that update you on new content on the platform)	O	O	O

Do you have additional or alternative ideas / are the features or functionalities missing that should be included?

Open-answer field after every cluster of possibilities

Metrics for fairness in AI

Table 16: Fairness metrics

Metric Name	Formula
Equalized Odds and Equality of Opportunity	TPR: $P(\tilde{y} = 1 \mid y = 1, G = 0) = P(\tilde{y} = 1 \mid y = 1, G = 1)$ FPR: $P(\tilde{y} = 1 \mid y = 0, G = 0) = P(\tilde{y} = 1 \mid y = 0, G = 1)$
Overall accuracy requirement	$P[Y = \hat{Y} \mid A = 1] = P[Y = \hat{Y} \mid A \neq 1]$
Statistical Parity	$P(\tilde{y} = 1, G = 0) = P(\tilde{y} = 1, G = 1)$
Predictive Parity	PPV: $P(y = 1 \mid \tilde{y} = 1, G = 0) = P(y = 1 \mid \tilde{y} = 1, G = 1)$ PPV shows the True Positive Rate.
Overall Predictive Parity	NPV: $P(y = 0 \mid \tilde{y} = 0, G = 0) = P(y = 0 \mid \tilde{y} = 0, G = 1)$ NPV is the negative predictive value
Calibration	$P(y = 1 \,\vert\, S = s, G = 0) = P(y = 1 \,\vert\, S = s, G = 1)$
Balance for positive/negative class	$E[s \vert y = 0, G = 0] = E[s \vert y = 0, G = 1]$
Treatment equality	$\frac{FN_{G=1}}{FP_{G=1}} = \frac{FN_{G\neq1}}{FP_{G\neq1}}$
Fairness through unawareness	$X_i = X_j \rightarrow \hat{Y}_i = \hat{Y}_j$
Mutual Information	$\sum(P(\hat{y}, s) log(\frac{P(\hat{y}, s)}{P(\hat{y})P(s)})) \leq \varepsilon$

Note: S indicates a score, A a sensitive attribute, G is group index and ε an arbitrarily small non-negative number.

Table by Loi, I., Zachos, P. & Moustakas, K.

in

Zangl et al. (2023) *Trustworthy AI compliance practices, assessment and conceptualization*