Next Steps

- Now that you have formed groups, please register them on Moodle until Sunday evening
- **Next step**: acquire the data for the RAG project (until April 4th)

Requirements

- 10-20 MB of pure text data
- Example: dataset with 1.2M words, 26k lines, 10 MB of pure text data
- We can lower the requirement for private/sensible data (maybe 500k – 600k words)
- If your dataset does not comply with these requirements, come talk to us. In certain cases, we can make exceptions.

Considerations

Private Data

- Are you allowed to use the data for the project?
- Third parties will see your dataset:
  1) External services (e.g. LLM Inference Provider, Embedding API)
  2) The lecturers of this course
- Make sure that you have the permission to use the data for the project

Adequacy

- Is the data well-suited for the use case?
- Does a RAG system lead to added value for your specific dataset?