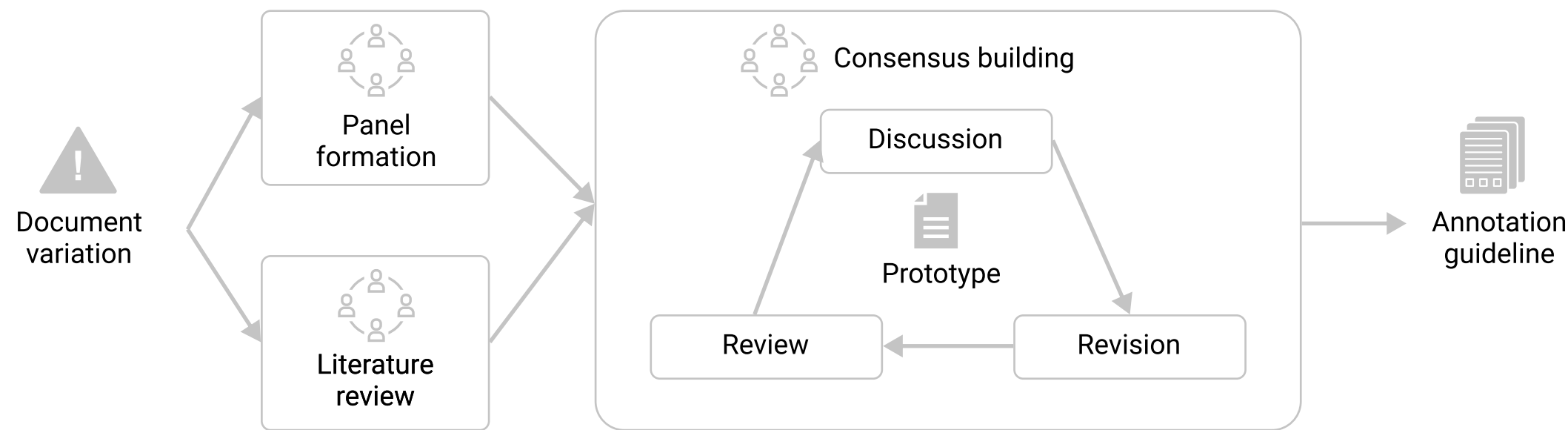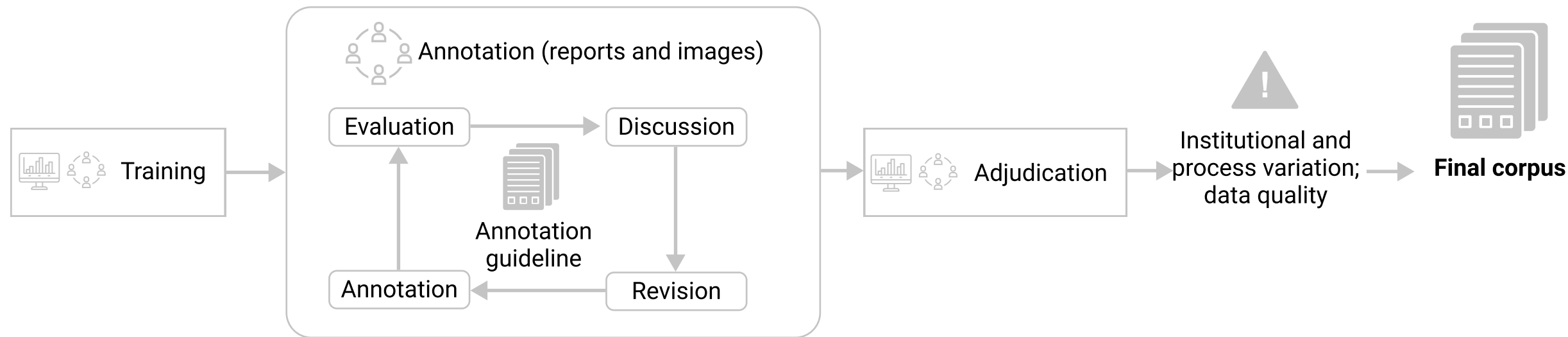# 4. Guideline development

5. Corpus annotation

Recommendations for Data Quality

Document variation

- Compare document metadata, e.g., practice settings, sections
- Adoption of standards, e.g., HL7 CDA
- Compare the cosine similarity between two documents represented by vectors
- Conduct a sub-language analysis to assess syntactic variation

Human factor and cognitive bias

- Training
- Consensus development
- Calculate the degree of agreement among abstractors

Recommendations for Process Documentation

Annotation guideline: background of the study, annotation tool, annotation instructions, concept definitions, meeting notes for consensus development and team discussion, annotation examples, reference. More details can be found at: https://github.com/OHNLP/annotation-best-practices