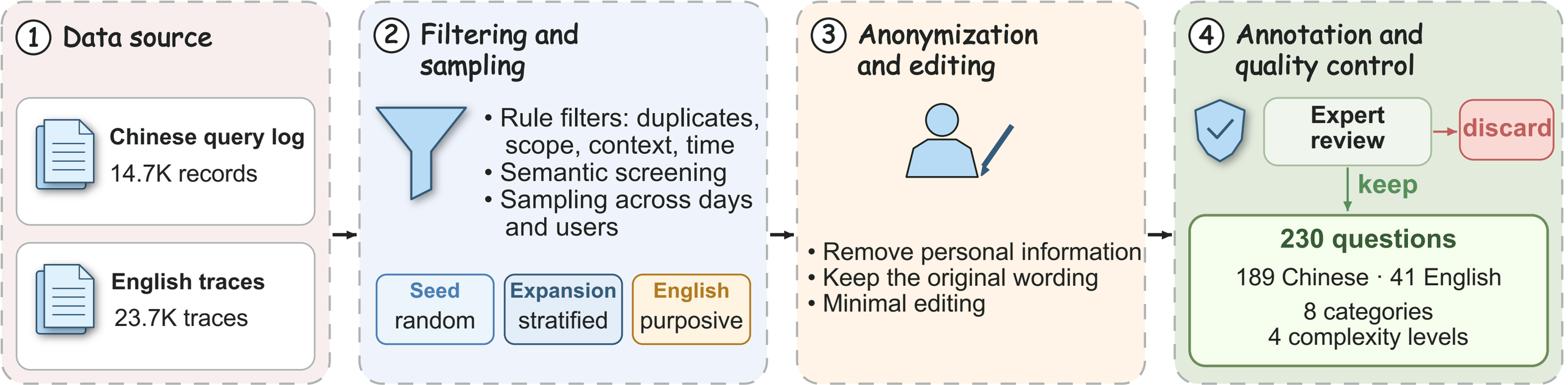

① Data source
Chinese query log
14.7K records
English traces
23.7K traces
② Filtering and sampling
• Rule filters: duplicates, scope, context, time
• Semantic screening
• Sampling across days and users
Seed
random
Expansion
stratified
English
purposive
③ Anonymization and editing
• Remove personal information
• Keep the original wording
• Minimal editing
④ Annotation and quality control
Expert review
discard
keep
230 questions
189 Chinese · 41 English
8 categories
4 complexity levels