

MuISQA Benchmark Construction

Topic

Biology, Chemistry, Geography, Medicine, Physics



Article Search and Collection

Evaluation Metrics

Stage1: Query formulation → diversity

Stage2: Passage retrieval → coverage

Stage3: Answer generation → performance



Topic and Article Collection



LLM-based Pre-annotation



Human Verification



Evaluation Metrics Design

Pre-annotation and Verification

