A REVIEW OF FINE-TUNING FOR QA AND A PROPOSAL FOR A NEW ANNOTATED CORPUS

Alice Breton 1, Adam Faci 1, Stéphane Pouyllau 1, Huma-Num IR*
1 Huma-Num IR* - HN Lab - CNRS, Paris, France (hnlab@huma-num.fr)
{name.lastname}@huma-num.fr

DESCRIPTION & GOAL

Developing reliable Retrieval-Augmented Generation (RAG) systems for specific domains like social sciences and humanities (SSH) requires question-answering (QA) datasets tailored to the linguistic and epistemological specificities of these fields. This necessity motivates a comprehensive review of existing QA datasets and corpora, ranging from general benchmarks to domain-specific corpora. We introduce a proposal for a **new synthetically generated SSH QA corpus** derived from **real user queries** in a documentary search engine called **Isidore**, supplemented by human annotation to ensure the accuracy and relevance of questions and answers.

SOTA DATASETS & FINE-TUNING METHODS FOR QA

Title	Source	Date	Type	Language	Size
HistoriQA-ThirdRepublic		2026	Politics (Debates)	Fr	1783 QA
Global PIQA Parallel		2025	Wikipedia	100 L from 65 pays 23 writing systems	13 388 QA
French Instruct dataset		2024	Instructions + Answers w/ multi-turned	Fr (Trad. from EN)	276K
CQuAE		2024	Educational	Fr	NA
Alloprof dataset		2023	Educational	Fr-Qc	29 349 QA
FrenchMedMCQA		2022	Medical	Fr	3105 QA
Squad - fr		2021	Wikipedia, SQUAD → FR	Fr (Trad. from EN)	90K tokens
Project PIAF		2020	Wikipedia (SQUAD proto.)	Fr	3835 QA
FQuAD: French Question Answering Dataset		2020	Wikipedia	Fr	60K samples
The KIAP corpus		2020	Science (Economy, Ling, Med.)	En, Fr, No	NA
TyDi QA		2020	Wikipedia	11 L (En, Ar, Bn, Fi, Ja, Id, Kiswahili, Ko, Ru, Te, Th)	204k QA
MKQA: Multilingual Knowledge Questions & Answers		2020	Google search → Q. Wikipedia → Ans.	26 L	260k QA
Chambers-Le Baron Corpus of Research Articles in French		NA	Science	Fr	1 045 872 words

SOTA for our Multilingual SSH QA Dataset

→ LuxIt : Luxembourgish Instruction Tuning Dataset from Monolingual Seed Data (2026)

- Prompt development

→ CQuAE (2024)

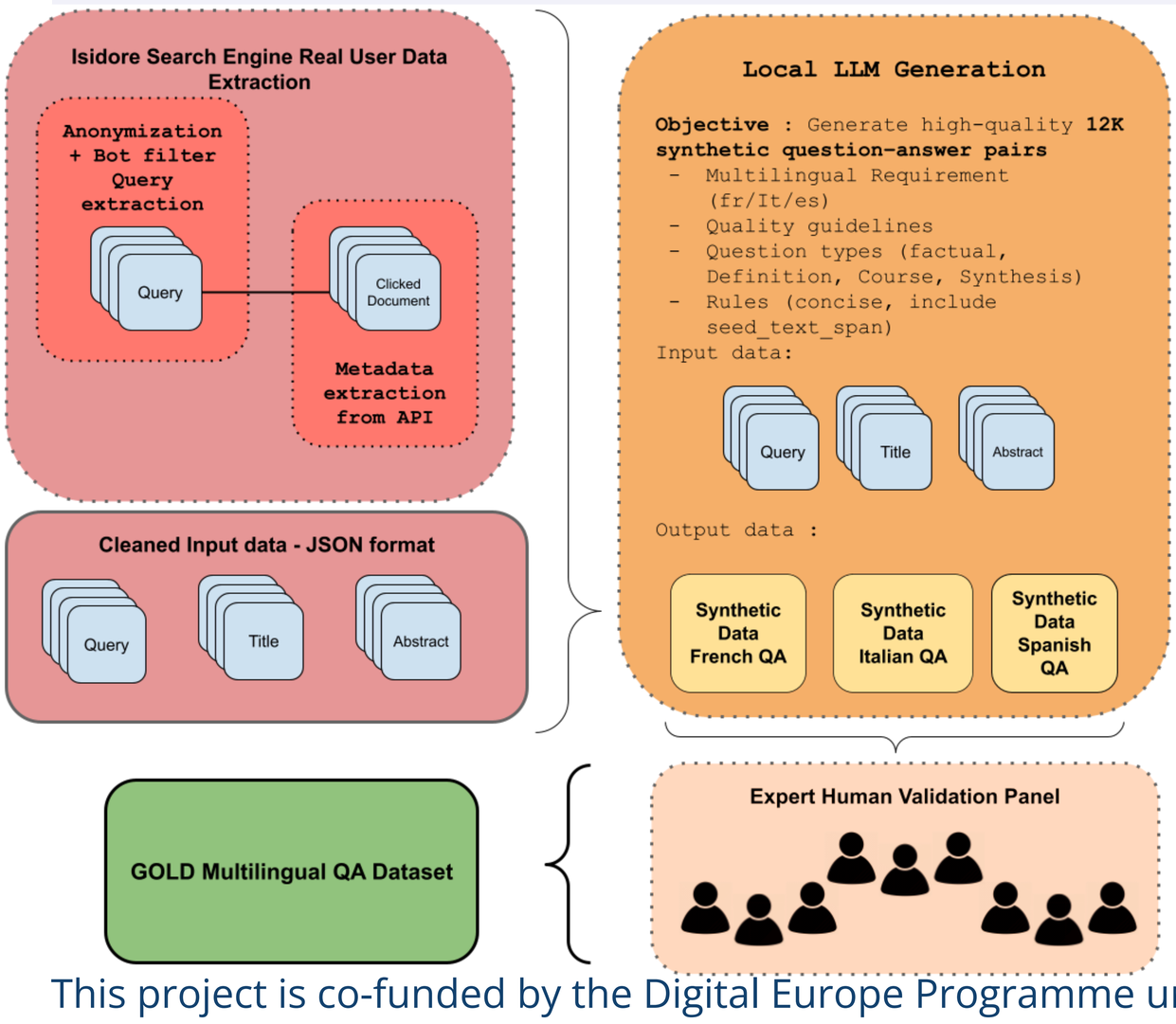
- Questions :
  - Factual: event, person, location, date... (Who did What When Where?).
  - Definition
  - Course: not a fact or a description but contains explanations or many details.
  - Synthesis: relies on different elements of the text and context.

→ LoRA (2021)

→ LaFFi (2023) - Supervised FT for QA

→ OThink-MR1 (2025) - Renforcement Learning for QA

MULTILINGUAL SSH QA DATASET PIPELINE



→ **Source:** Search logs from **Isidore** (scientific search engine) anonymized, bot-filtered, reduced to query + clicked document metadata

→ **Scale: 12K query–clicked document pairs**

→ **Generation:** Local LLM generates QA pairs in French, Italian, and Spanish from each query/title/abstract triple

→ **Question types:** Factual, Definition, Course, Synthesis each answer keeps its original source text span

→ **Validation: Expert human annotators review** and validate/reject each generated pair

→ **Output:** GOLD multilingual QA dataset → used to fine-tune a QA model for our RAG system

→ **Link:** https://gitlab.huma-num.fr/llms4eu-research-ssh-public/multilingual-ssh-qa-dataset

This project is co-funded by the Digital Europe Programme under Grant Agreement No 101198470.

Co-funded by the European Union