

Is Agentic RAG Worth it?

Pietro Ferrazzi

Fondazione Bruno Kessler

pferrazzi@fbk.eu

Motivation

Working with R&D of Alkemy SPA, we found that most of the projects involving GenAI were based on RAG, and there was a shift happening toward Agentic RAG. We wondered whether it was worth to go there

This work is the result of a collaboration between

Pietro Ferrazzi^{1,2}, Milica Cvjeticanin³, Alessio Piraccini⁴, Davide Giannuzzi⁵

¹Fondazione Bruno Kessler, Trento, Italy

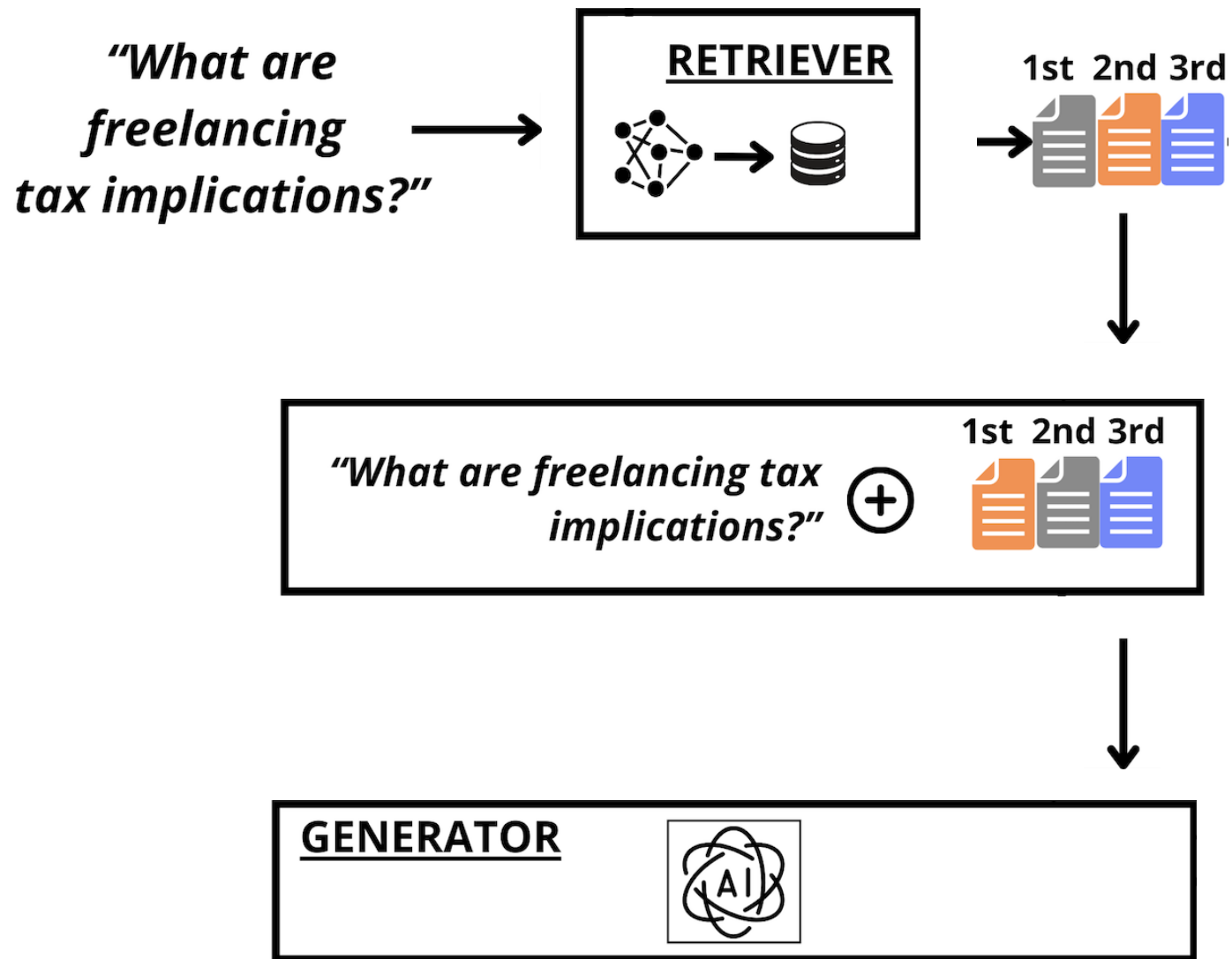
²University of Padova, Italy

³Cargill Geneve, Switzerland

⁴Alkemy, Milan, Italy

⁵Komebi Studio, Milan, Italy

Naïve RAG



| Naïve RAG issues

Retrieval performed even for queries that do not require it

| Naïve RAG issues

Retrieval performed even for queries that do not require it

Queries and docs in the KB differ, causing weak retrieval

| Naïve RAG issues

Retrieval performed even for queries that do not require it

Queries and docs in the KB differ, causing weak retrieval

Noisy or suboptimal retrieval

| Naïve RAG issues

Retrieval performed even for queries that do not require it

Queries and docs in the KB differ, causing irrelevant retrieval

Noisy or suboptimal retrieval

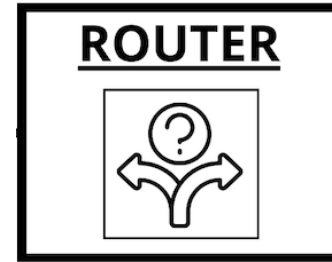
How to solve these issues?

ENHANCED RAG!



| From Naïve to Enhanced RAG

Retrieval performed even for queries that do not require it



Route the queries to the retrieval step only if required

Queries and docs in the KB differ, causing weak retrieval



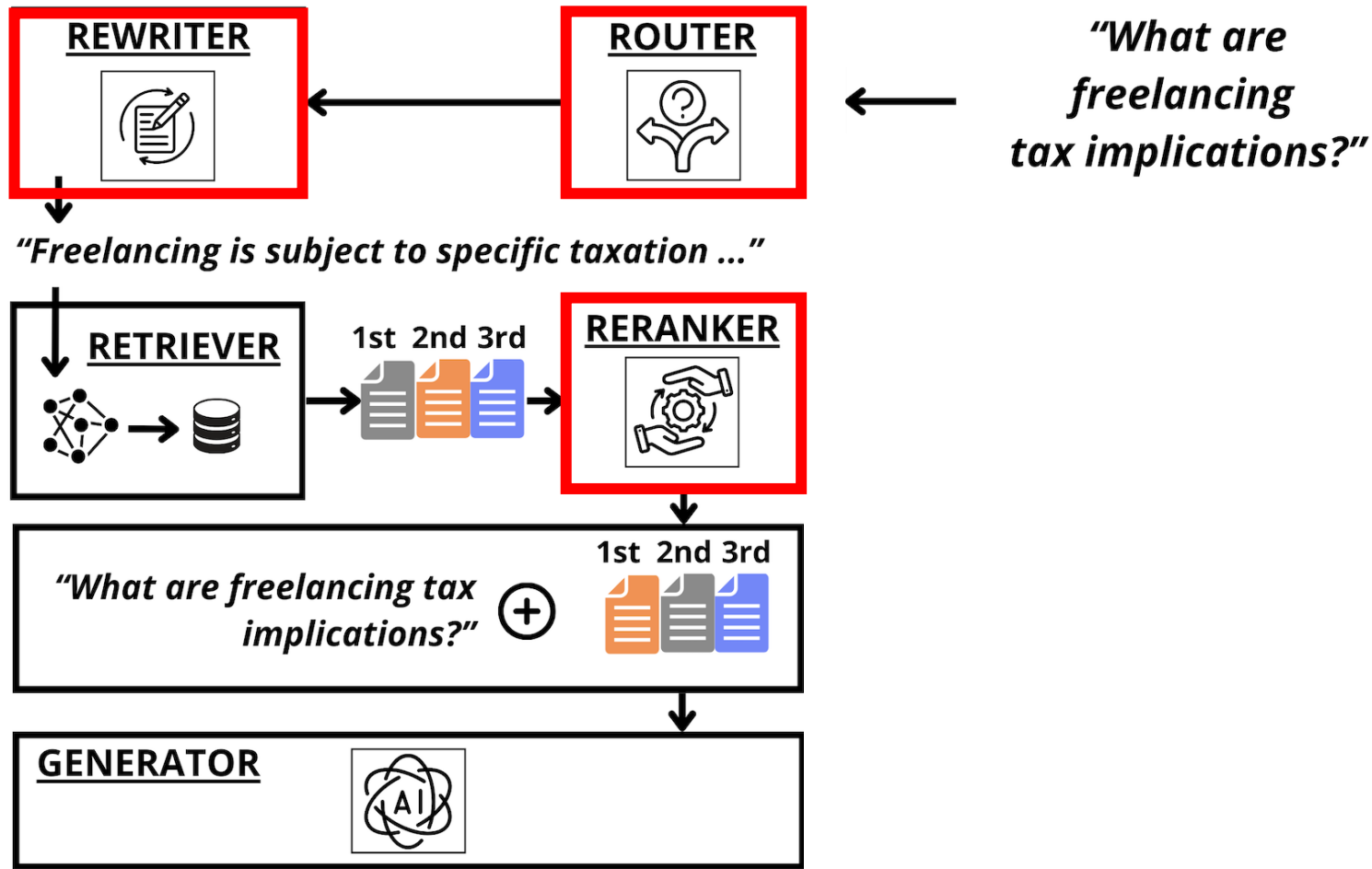
Rewrite the query to make it consistent with the documents

Noisy or suboptimal retrieval

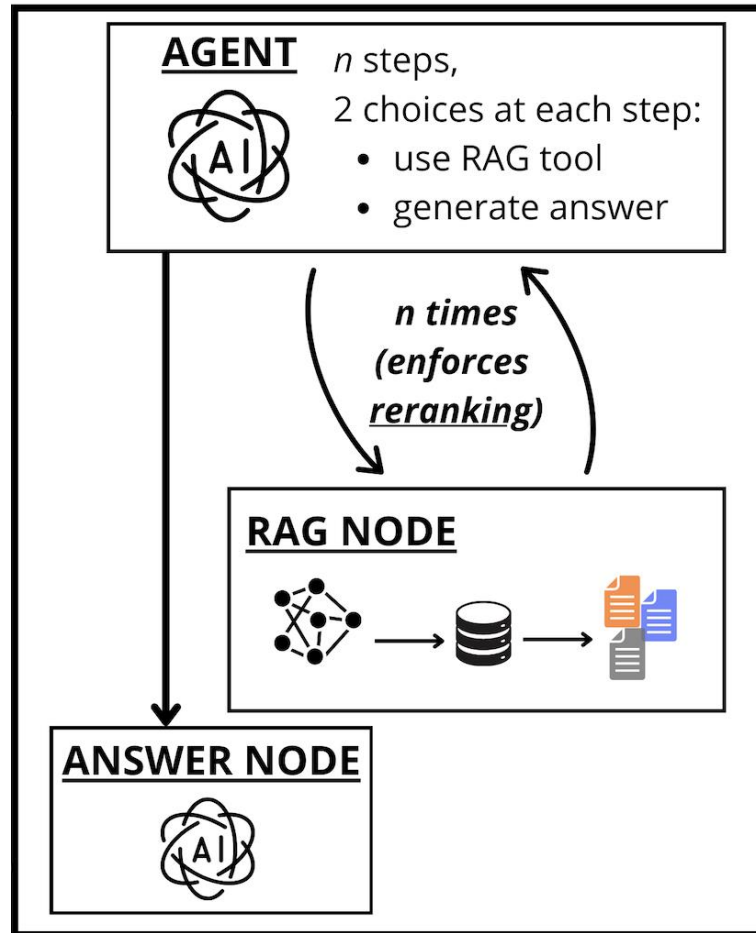


Reorder the retrieved documents to refine the importance list

From Naïve to Enhanced RAG



| The hot one: Agentic RAG



← ***"What are
freelancing
tax implications?"***

| The hot one: Agentic RAG

Research and Implementation Question:

When designing RAG systems, should practitioners adopt **Agentic** architectures **or** more traditional **(Enhanced)** pipelines?

Evaluation dimensions

Experimental schema:

- 1) Find a Naive RAG shortcoming more complex systems try to solve
- 2) Find SOTA approaches to address them in Enhanced and Agentic systems
- 3) Analyse and compare the performances

Evaluation dimensions

Experimental schema:

- 1) Find a Naive RAG shortcoming more complex systems try to solve
- 2) Find SOTA approaches to address them in Enhanced and Agentic systems
- 3) Analyse and compare the performances

Naïve RAG shortcoming	What we evaluate	Implementations	
		Enhanced	Agentic
Retrieval is performed even for queries that do not require it	Accuracy of RAG usage for in-scope and out-of-scope queries	Semantic routing system	Agent decides whether to do retrieval or not
Queries and documents in the KB differ in format or semantics, causing weak retrieval	Impact of query rewriting techniques	Hyde-based query rewriting	Agent rewrites query as it wishes
Noisy or suboptimal retrieval	Impact of retrieved document list refinement techniques	Encoder-based re-ranker	Agent can redo retrieval multiple times
The underlying LLM is too weak / too slow / is too expensive	Impact of selecting more / less powerful models	Test different LLMs	Test different LLMs

1) Find TYPICAL TASKS for RAG APPLICATIONS

TASK



Question Answering

**Information
Retrieval / Extraction**

1) Find TYPICAL TASKS for RAG APPLICATIONS

TASK	DESCRIPTION
	<i>Users can ask any type of question to be answered via knowledge retrieval</i>
	<i>Domain-specific queries to be answered by grounding responses on expert knowledge</i>
	<i>Find resolved blog posts that address the same question posed by the user, providing user-friendly summary</i>
	<i>Seek evidence on users' statements finding docs. Return final assessment with a summary of the references</i>

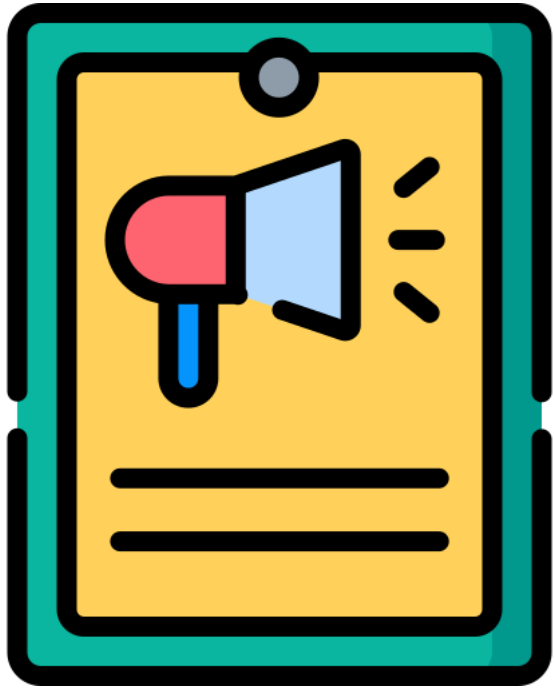
Evaluation data

2) Find DATASETS on these TASKS

DATA	DOMAIN	TASK	DESCRIPTION
NQ	General		<i>Users can ask any type of question to be answered via knowledge retrieval</i>
FiQA	Finance		<i>Domain-specific queries to be answered by grounding responses on expert knowledge</i>
CQAD-EN	Grammar forum		<i>Find resolved blog posts that address the same question posed by the user, providing user-friendly summary</i>
FEVER	Wikipedia		<i>Seek evidence on users' statements finding docs. Return final assessment with a summary of the references</i>

| Results

**Come and see
our POSTER**



**Have a look at our
PREPRINT**

