Industrial AI Canvas (RAG Edition)

Project name:

Date:

Version:

Data curation

Which cleaning operations are necessary?

How many gold standard annotations are necessary? How can they be acquired?

Can the assistant be specialized for a first topic? Which data subset is necessary for that?

How can different prompts be compared and evaluated?

How can standardized processes for model evaluation/data analysis look like?

What data is IP-sensitive?

Does the data have to be anonymized?

Tools for NLP data curation:
Kern.ai refinery
Argilla
LabelStudio
Renumics Spotlight
Galileo

Problem Classes

Can agents be used to increase robustness / capability?

Can the query be broken down into sub-queries?

Can metadata help to retrieve context materials?

Which metrics can be used to evaluate the retrieval performance?

Which metrics can be used to evaluate the generated result?

Observability/ validation tools for LLM:
Arize Phoenix
Evidently
Langfuse
Deepchecks
Langkit

User Interaction

Classical chatbot interface or custom retrieval application?

Can the user control the prompting?

Is there any additional context to consider?

Which latency is required?

How can the user give feedback on the answer?

Tools for building chatbot apps:
Gradio
Streamlit

Operations

Does the solution serve external users?

What failure modes can be critical ?
(data leakage, brand abuse, legal issues)

Which internal/external regulations apply to the solution?

Which processes have to be run through to productionize a system e.g. software security assessment?

Which stakeholders are necessary to uphold the operation of the system?

What kind of maintenance is needed for iinfrastructure and application layer?

How often does the model have to be updated? How much effort is involved?

Which metrics have to be monitored to detect data drift and system performance decline?

What is done in case of the system's performance declining?

Testing and monitoring:
Giskard
guardrails
LLM-Guard

Methods

Is the problem multi-modal?

Should there be fine-tuning in addition to or instead of RAG?

Do pre-trained models exist for the task?
Which model should be used?
(GPT-4, GPT-3.5, Llama-2, Mistral, vicuna)

Which prompting technqiues should be used?

Tools / APIs for LLM models:
Hugging Face
OpenAI
Microsoft Azure
Aleph Alpha

Value Propositions

What are tasks the user has to fulfill?

How can the assistant support the user in these tasks (examples)?

How can the business benefit from analyzing user queries?

What key competitive advantage can be generated? Now and in 5 years.

How can the business value be measures?

Integration

How does the assistant integrate into the application layer?

Does the assistant have to run on premise?
Should the model be self-served?

Which cloud providers are already available?

Cloud infra:
Azure
AWS
Google Cloud

Vector DBs:
Postgres
Qdrant
Milvus
Chroma
Weaviate

Data Sources

Which data sources are available?

Which data formats are essential? Are there readers and scrapers available for this data formats?

How is the data stored, and how can it be accessed?

What are the access rights on the data samples?

Tools / APIs for data scraping and formating:
Chaindesk

Resources

What will the project cost? What is the cost for a PoC?

What does the compute infrastructure cost for inference and retrieval?

What are the costs for tooling and IT infrastructure?

Which roles are available to support system development and operation?

Who are important decision makers regarding budget and technical operations?

Tools for building RAG systems:
Langchain
Haystack
Llama-index
SpaCy
Jina

Designed by: Renumics GmbH

renumics