SquirroGPT

10 Considerations for constructing a RAG – Checklist

Uncover Insights with AI-Apps from Squirro

squirro

Agenda

10 Considerations for constructing a RAG

Retrieval Augmented Generation

Meet Squirro

10 Considerations for constructing a RAG

squirro

10 Considerations for constructing a RAG

#	Consideration	Check	Notes
1	Data Access & Life Cycle Management		• Manage the entire lifecycle of the information • Data is efficiently processed, enriched, available and eventually deleted
2	Data Indexing & Hybrid Search		• Setup of hybrid search • Maintaining such a large-scale index over time
3	Enterprise Security & Access Control at Scale		• Respecting complex Access Control Lists (ACL), • Secure setup of a RAG system
4	Chat User Interface		• Adaptable chat interface • Extendible chat interface for new task classes
5	Comprehensive System Interaction		• Management of interactions between Information retrieval, user interface, LLM and entailment check • Comprehensive Information Retrieval (IR) Stack
6	Prompt Engineering		• Creating an effective prompt service • Integrating adaptive mechanisms can help refine prompts based on real-time interactions and feedback
7	Chain of Reasoning		• Sophisticated user journey progression support
8	Enterprise Integration		• Integrating a RAG into an existing enterprise setup (SDK, API , etc.)
9	Continuous Operation		• Service Level requirements, organizational setup, etc. • Talent availability to operate such systems over time
10	Cost Considerations		• Find the good balance between the LLM and the IR stack • Total cost of ownership considerations

SQUIRRO GPT

Retrieval Augmented Generation

“Autoregressive LLMs are doomed

There will be better systems that are factual, non-toxic, and controllable

Marrying them with tools such as search engines is highly non-trivial”

Yann Andre LeCun
Chief AI Scientist at Facebook & Silver Professor at the Courant Institute, New York University.

Combine an Insight Cloud with Generative AI[1]

Generative AI models hallucinate
Answers come with the "why"

Works with public data; only limited with company data
Answers are based on your own data

They have a disdain for enterprise security
Robust, secure, and respecting ACLs

& more[1]

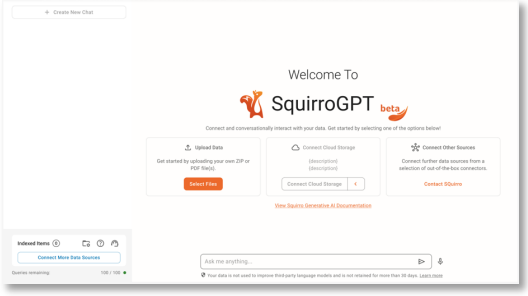
1) https://squirro.com/apps/generative-ai-trial/ , ACL = Access Control List

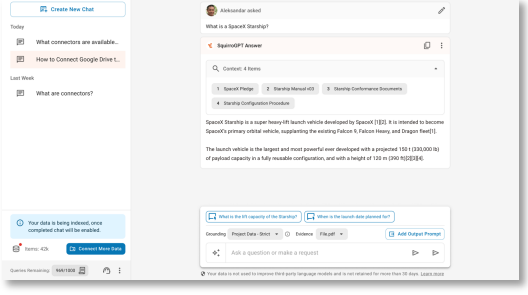
An Enterprise ready Application Stack

No Hallucination ✓ **Evidence based:**
Answers come with the "why"

Local content ✓ **Your Company Data at Scale:**
Local context and trained LLMs

Enterprise security ✓ **Robust, secure, and performant**
Enterprise ACL setup, and more





Prompt size, data ingesting, scaling, copyright, integration in Salesforce, Dynamics) etc. included.

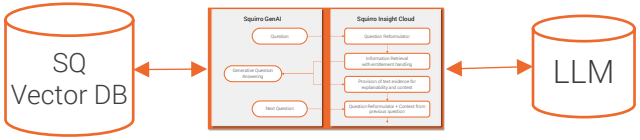
Retrieval Augmented Generation (RAG)

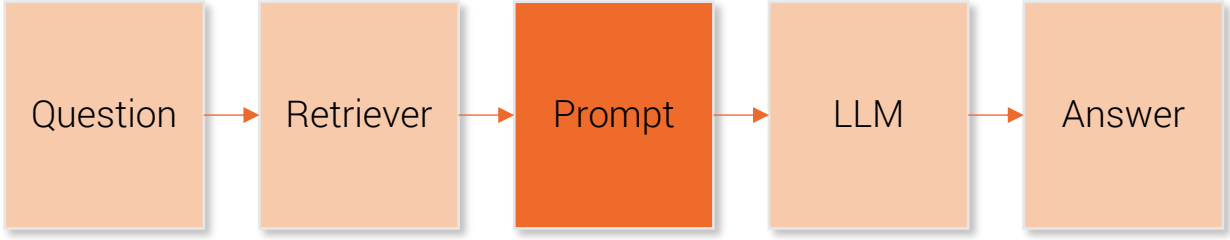
Load related context / information into "working memory" LLM context window

Google / Bing / Squirro
Classic Enterprise Search
(Retrieval)

Retrieve relevant information;
Expose to LLMs of your choice

OpenAI, Cohere, Anthropic
Large Language Model
(Generation)





Generative AI stack

Application

End-to-End applications including models or Apps without proprietary models (ability to use a best of breed approach)

Models

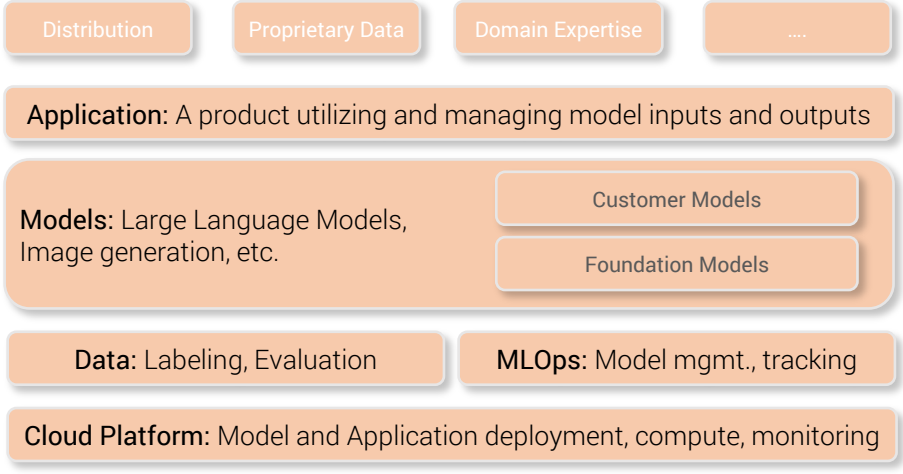
Splitting the model layer to differentiate between proprietary and open-source models as well between foundation and custom models

Data & MLOps

Models derive their value from the data they're trained on. So, there's a need to factor in data and machine learning operations (MLOps)

Cloud Platform

To run GPU based platform locally is hard; Time to switch to cloud platforms



Meet Squirro

Meet Squirro

Global Presence

- Augmented Intelligence Software Provider
- Presence in Europe, United States and Asia
- Global presence and partner network

Reference Customers

armacell — MAKING A DIFFERENCE AROUND THE WORLD
BANK OF ENGLAND
CANDRIAM — A NEW YORK LIFE INVESTMENTS COMPANY
Henkel
EUROPEAN CENTRAL BANK
Standard Chartered

A Different Approach to AI

- A powerful Insights Engine for Decision Intelligence
- Tangible, Industry/Issue Specific AI-Applications
- Build your own apps with the Squirro AI-Studio

Investors

salesforce ventures
ICOS CAPITAL
FINCH CAPITAL
HAMMER TEAM

Gartner Magic Quadrant
FORRESTER WAVE Cognitive Search & Knowledge Discovery
DEUTSCHE BUNDESBANK EUROSYSTEM
AI
ASSET ASIAN AWARDS TRIPLE A
ISO 27001 Certified

Customer Success

A self-learning system keeping you in the know & recommending what's next.

3 powerful application suites built on top of a powerful & versatile platform:

Sales Insights: Top line growth

Service Insights: Cost savings

Risk Insights: Reduce exposure

Enterprise Search: A more effective enterprise

All Solutions are SquirroGPT enabled.

Sales

CANDRIAM
armacell

Sales: Source opportunities

Service

Standard Chartered
OCBC Bank

Service: Automated case resolution

Risk

EUROPEAN CENTRAL BANK
EUROSYSTEM
DEUTSCHE BUNDESBANK
EUROSYSTEM

Risk: Threat detection for supervisory & audit

ES

MUBADALA
Henkel

Cognitive Search: A single data & knowledge fabric

For customer case studies see https://squirro.com/resources/

Thank You

Dorian Selz

dorian@squirro.com

squirro