[sfeir]

WENVISION

Gen AI for CIOs and CTOs

Release 8

TECHWAVES - #GENAI FOR CIO & CTO

RELEASE 8 - April 2024 - Updates signaled by:

MAJ - R8

“GENERATIVE AI IS NOT A THREAT TO DEVELOPERS, QUITE THE OPPOSITE”

EDITOR’S NOTE

Olivier Rafal
Consulting Director - Strategy

olivier@wenvision.com

Many amongst us have read articles positing that the advent of generative Artificial Intelligence will usher in an era of unprecedented developer productivity, to the extent that it might eventually render certain jobs obsolete, or perhaps even challenge the very existence of the developer profession. Do you believe it? Personally, I remain skeptical.

It is undeniable that generative AI offers a marked enhancement to developers' efficiency. It represents an invaluable aide, proficiently executing the most monotonous tasks: scripting commands to interact with an API, integrating fields within a web interface, or transmuting code from one dialect to another — tasks that are standardized, labor-intensive, and devoid of any intrinsic value.

This digital assistant might transcend its current capabilities, autonomously generating portions of the requisite application code, perhaps from basic prompts or even from impromptu hand-drawn schematics — a contemporaneous embodiment, augmented by AI, of the ethos of Model Driven Development. This paradigm, conceived in the dawn of the 21st century, postulated the notion of crafting the very skeleton of software predicated on its UML representation.

MDD never managed to establish itself as a successful model or gain widespread adoption as it was overly ambitious, excessively rigid, and lacking in performance, among other issues. Owing to their flexibility, efficacy, and ease of use, development tools incorporating generative AI stand a far better chance of success.

WHY AI WON'T REPLACE DEVELOPERS' POSITIONS

This, however, does not imply that the emergent AI-infused development tools will lessen the need for developers in the forthcoming years. Here are three considerations on this matter:

- AI necessitates guidance. The AI generates code upon the developer's request. The developer is the one who comprehends the specifications set by the Product Owner, possesses an understanding of the software's architecture and its operational context, and must refine the prompts based on these factors.
- One must meticulously assess the code produced by AI. Regardless of whether one employs generative AI or not, a developer always inspects and evaluates the code. Without exception.
- **There remains an ever-growing demand for code creation.** To believe that an uptick in productivity will lead to job losses in the imminent future is to mistakenly assume that our capacity for code production has a finite boundary. Especially in this era, where technology and business are intrinsically intertwined, where software is omnipresent, and above all, where its evolution is ceaseless.

Not only do developers have many promising days ahead of them, but generative AI will also substantially ease their work. This will undoubtedly please Chief Information Officers who will implement it.

GIVING VOICE TO THE EXPERTS

Consulting

Marie Fontaine
Head of GenAI

marie@wenvision.com

The pace of evolution, both technologically and from a user perspective, of Generative AI is unparalleled. Keeping up with the advancements in the field and understanding its potential impacts on our businesses have become challenging. The story begins in 2020 when OpenAI opened access to GPT-3 in Beta, but it was with ChatGPT that the general public became aware of the potential of this tool.

This feeling of being overwhelmed is met with a sense of immediacy. The projected productivity increases are so profound that no business can afford to overlook this technology, especially in the context of software development. Research conducted by GitHub reveals that developers using Copilot take 56% less time to complete a task. McKinsey anticipates a productivity boost of at least 30% in software engineering. This emergence presents new challenges for companies: the selection of appropriate tools, harmonizing the integration of this technology, and ensuring support for their team members.

At SFEIR and WENVISION, we've been observing this development for several months and are eager to share our findings via TechWaves - #GenAI 4 Dev. Formulating a strategy for Generative AI necessitates familiarization, devising a plan grounded in real-world scenarios, and creating an appropriate organizational structure.

Engineering

Florent Legras
Engineering Director

legras.f@sfeir.com

LLMs offer a refined solution that **significantly enhances existing applications.** For business-oriented users, they provide a tangible opportunity to access genuine self-service Business Intelligence through interactive dialogue. For developers, these advanced tools **optimize productivity** and offer effective solutions to specific challenges.

Implementing an LLM within an enterprise requires a transition from a basic proof of concept to a production-ready application. This shift is of utmost importance. Although a **POC may showcase the elementary functionalities of an LLM system**, it might overlook essential aspects such as scalability, performance, and security measures. **Transitioning from a POC to a deployable application can be challenging due to the variable nature of LLMs.**

These applications must be carefully crafted and conceptualized in a comprehensive context, integrated into a data platform within an overarching information system. A well-devised architecture ensures the scalability, maintainability, and extensibility of the application. A production-ready application should include mechanisms for monitoring and managing models, which entails performance tracking, issue identification, and ensuring the application's security and compliance. Without these integral features, an LLM application might pose more challenges than benefits.

“CODING WITH GENERATIVE AI? I TOOK ON THE CHALLENGE”

Engineering

Didier GIRARD
Co-CEO SFEIR & WENVISION

didier@wenvision.com

Over the past year, several activities have restricted my time spent on coding, dwindling my coding sessions to just a few hours a month.

I believe it is vital to code on a regular basis, and due to a lack of practice, I started losing some reflexes, much like a musician who stops playing their instrument and gradually loses dexterity.

This inspired me to embark on a challenge: **could I develop an application based on generative AI, despite my somewhat rusty skills?**

To accomplish this, I used Replit and its generative AI, a convenient tool that necessitates no installation. My goal was to reconstruct an application I had crafted in Flutter several years ago to counteract my dyslexia: when I am weary, my disability becomes more apparent, leading to "keyboard missteps".

The original Flutter application translates my text from French to English, and then back from English to French. It serves as my personal assistant. For translation, it employs the Deepl API.

I'm fond of the Go language, so I opted to work with Replit in Go. The user interface was also crafted by Replit, initially using a basic form, and later building upon Bootstrap.

“I accepted the challenge of coding an application with Replit and its generative AI. The outcome? A quick and efficient development! AI won't replace developers, but it will enable them to code faster, just as the electric plane has helped carpenters.”

The development was straightforward and took about 2 to 3 hours.

Encouraged by this rapid advancement, I chose to introduce a new capability to my assistant: text generation. To achieve this, I utilized OpenAI's GPT-4. The goal was for my assistant to both review and help write my texts. With Replit, integrating OpenAI's API only took me a few dozen minutes. Right now, the integration is quite simple, but ultimately, my assistant fulfills my requirements.

My experimentation with Replit confirmed what I had suspected. The introduction of generative AI to help developers represents a significant step forward. In just a few quarters, the most efficient developers will be such tools.

However, certain skills are crucial to use generative AI effectively in development. From my experience, albeit limited to a few hours, I believe debugging ability is fundamental. **The AI doesn't always get everything right the first time, and being able to swiftly pinpoint issues and request the AI to produce a new solution is vital.**

I urge you to try out these tools. While the roles developers will undergo significant transformations with AI, it will not replace them. Just as the electric planer didn't make carpenters obsolete, AI will simply enable us to code more rapidly and efficiently.

To be continued...

For those of you who would be curious, here's the GitHub repository: https://github.com/dgirard/wizwizwiz17/tree/main

Admittedly, it might not be the most elegant code, but consider it is the work of an AI that's been coding for less than a year. In the years ahead, it will only get better!

FASTER, HIGHER, STRONGER

MAJ - R8

Consulting

Aurélien Pelletier
Platform director
aurelien@wenvision.com

A few months before the Paris Olympics, the new generation of Large Language Models (LLMs) has embraced the Olympic motto.

Faster

The majority of new LLMs now come in various sizes, each offering a cost/speed/relevance ratio suited to different use cases. This is the case, for example, with Google's Gemini 1.0 (Nano, Pro, Ultra), Anthropic's Claude 3 (Haiku, Sonnet, Opus), and Mistral (Small, Large).

Grog introduces a new processor that generates text ten times faster.

Larger volumes of data

Google's Gemini 1.5 announces a context window of 1,000,000 tokens, nearly eight times larger than GPT-4's (128K tokens). This would allow for the ingestion of the entirety of The Lord of the Rings, an hour of video, or 30,000 lines of code in a single go.

Multimodality: Beyond Text

Conversations are no longer limited to text but can now include images, audio, and even video. This versatility opens the door to many innovative applications, such as medical image analysis, voice recognition, or the generation of multimedia content.

Rise of Open Source

Open Source models are also becoming more powerful. Mistral released Mixtral, for Mixture Of Expert, while Databricks introduced DBRX as the most powerful of the open models..

Stronger

According to chatbot Arena where Hugging Face organizes "a battle of the LLMs," Claude 3 Opus has dethroned GPT-4. And our daily use confirms this feat.

Now everyone is waiting for OpenAI's response and a probable GPT-5...

THE LATEST GENERAL AI NEWS FROM THE PAST FEW WEEKS

MAJ - R8

Consulting

Baptiste Pugnaire
AI / Data Consultant
baptiste@wenvision.com

Battle at the Summit

The race for innovation in the field of generative AI is heating up. Leaders are multiplying their announcements:

- **Anthropic** unveils Claude 3 (03/04), promising performances superior to GPT-4. Mistral announces its "Large" model (04/26), a first with non-open-source weights, accessible only via API.
- **Mistral** had started training LLama 3 in January, and the release of the first small models is imminent.
- **Meta** avait commencé l'entraînement de LLama 3 en janvier et la sortie des premiers petit modèles serait imminente.
- **Google** announces Gemini 1.5 (02/15) with a record context window, as well as the open-source models Gemma (02/21).
- Whisper, OpenAI's TTS model is now available on Microsoft Azure (13/03). On another note, a blog post announcing the release of GPT-4.5 for June had leaked (03/13).

In this frantic race, innovations are succeeding one another, and each player seeks to gain an advantage. However, while the tech giants are multiplying their announcements, companies are awaited, and it is now imperative to demonstrate tangible results to avoid "over promise, under deliver".

OpenAI Revolutionizes Video Creation with Sora

It was the major announcement that shook the world of video creation, OpenAI unveiled Sora on February 16th. Sora ("sky" in Japanese) is a video generation model, it allows, from a simple prompt, the creation of photorealistic and coherent video clips of several tens of seconds. Hollywood studios such as Warner, Paramount, and Universal have already been invited to technical demonstrations of this new model.

Go big AND go open source

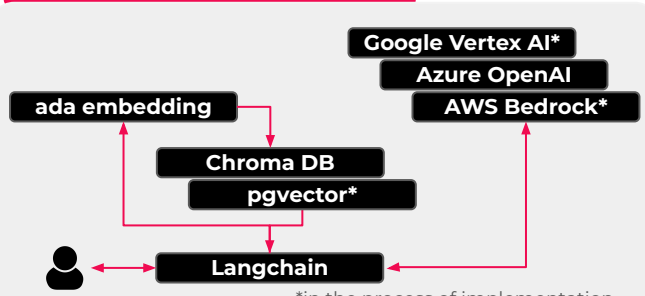
Until recently, the term "open-source model" often implied "small size" (max 70 billion parameters), but recently xAI, Databricks, and Cohere have released powerful and relatively massive models: Grok-1 (314B), DBRX (132B) and Command R+ (104B). This shift is excellent news for the AI research and innovation community.

Microsoft Consolidates Its Grip on the AI Ecosystem

Recently, Microsoft has recruited Mustafa Suleyman and Karén Simonyan (formerly of Inflection AI and DeepMind) to lead its Microsoft AI division. Historically a privileged partner of OpenAI, the company has been multiplying agreements, notably with Mistral AI, whose latest model is accessible only on Microsoft's cloud.

FEEDBACK: VEOLIA SECURE GPT

THE IMPLEMENTED ARCHITECTURE



*in the process of implementation

In the spring of 2023, Veolia's IT group management decided to create Secure GPT, with the goal of mastering this technology and avoiding the use of public tools. The ambition was to first open it to the 5,000 IT collaborators in the group. Given the results (adoption rate, scalability, cost control, etc.), the application is now accessible to some 200,000 Veolia collaborators worldwide, several months ahead of schedule.

Secure GPT currently offers 3 services:

- A conversational chat, with an adjustable level of creativity
- Translation into 95 languages
- The ability to upload PDFs and text files for querying in chat (for the duration of a session)

The initial developments utilized Google Cloud for the front end and Azure OpenAI for the LLM and embedding, orchestrated via Langchain on Google Cloud Run. Veolia is currently working on expanding its platform, with the implementation of AWS Bedrock (Llama2, Claude2, Stability) and Google Vertex AI (Palm2), as well as the deployment of pgvector in PostgreSQL, which will offer persistence for querying Veolia's knowledge bases.

In the long term, the challenge for Veolia will be to invest in what will bring value to its collaborators and will not end up in the roadmap of SaaS publishers. In other words, Veolia will not focus on optimizing a CRM, but will concentrate, for example, on solutions specific to the efficient operation of its plants.

"It's a revolution in IT.

There will inevitably be disappointments, setbacks on this or that PoC, but overall, GenAI is going to drastically change our way of working and developing software. Our duty is to master these new technologies to support Veolia's transition."

Les enseignements de Fouad Maach, Head of Architecture and Industrialization for Veolia Group IS&T

Which skills are required?

The right person can do the job, but it's crucial to quickly parallelize: one can start with a very small team to initiate development, but scaling up requires the creation of a real organization, where people are trained and their skills are enhanced. There's no strict need to be a data scientist, but understanding basic concepts is essential along with having back-end development skills and strong Cloud competencies - and adding front-end and DevOps skills for scaling up. Developers also need to adapt to a paradigm shift: moving from classic algorithmic programming to Langchain agents that handle a part of the decision-making process. This is a change that needs to be accepted and embraced.

Which other advices can you give?

Take care of the legal aspect: the 'legal' team was a great help to us. Things would have been problematic without their validation. It's also important to support the users: we might think that everyone knows how to use GenAI, but we realized that there was a need for training; we are going to deploy an e-learning module.

Langchain is at the heart of the architecture; so, has the framework been satisfactory?

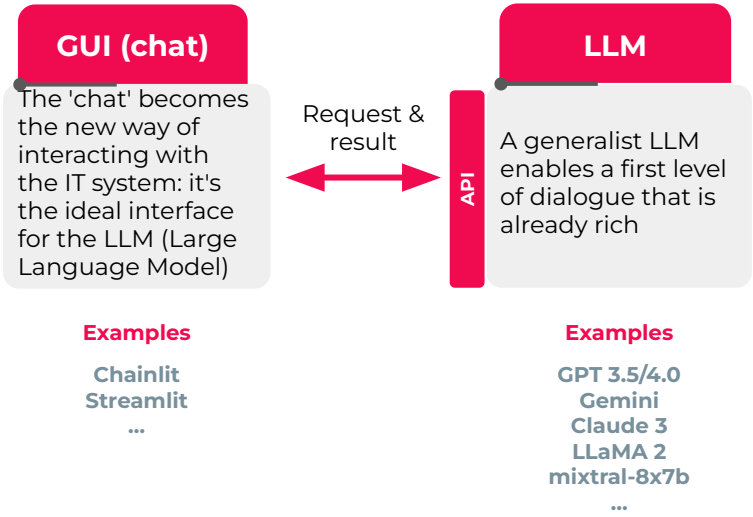
Yes, it is simple to use and very powerful. It meets the expectations of agnosticism well: you can change the LLM without any worries. There is a small learning curve, but after that, it's all good!

TECHWAVES #GENAI FOR DEV, USER GUIDE

[sfeir]

WENVISION

HOW TECHNOLOGIES POSITION THEMSELVES AMONG EACH OTHER: OUR VISION OF A “POC LLM APP STACK”



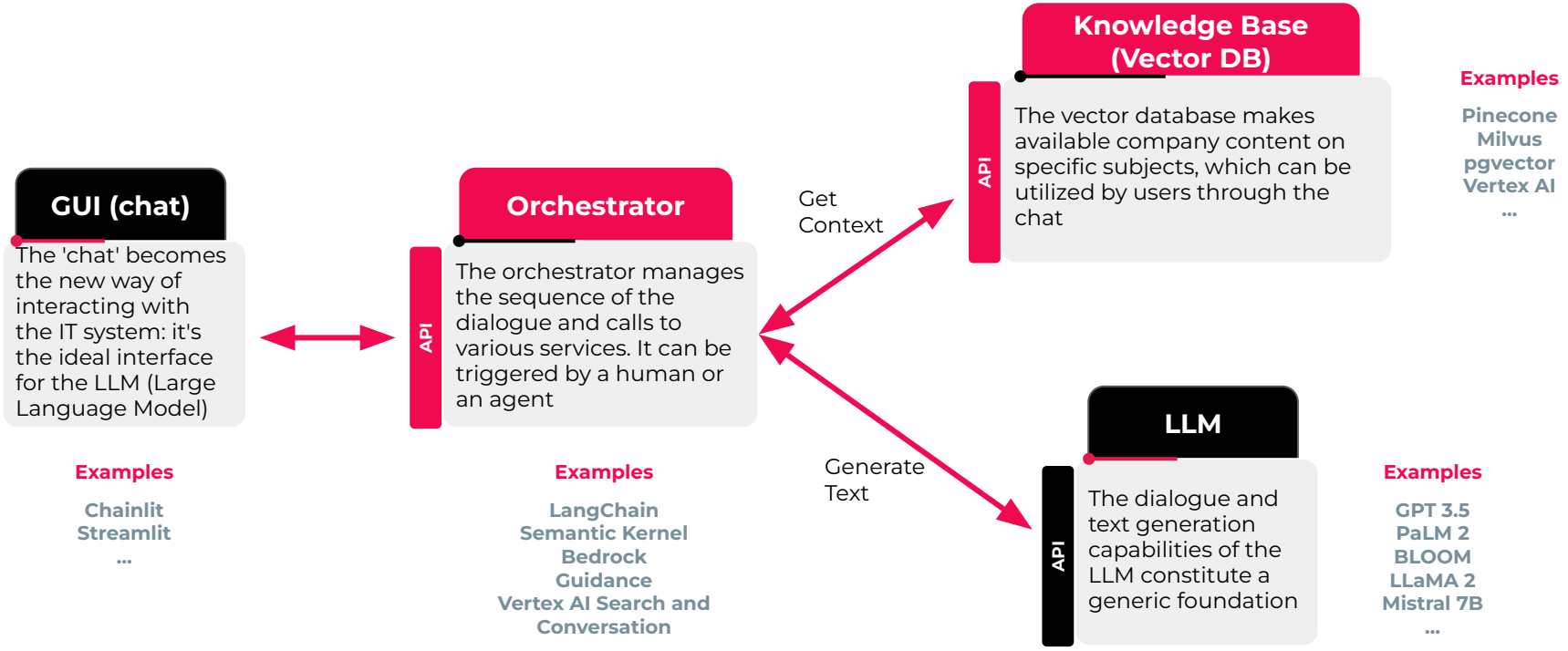
STARTING WITH A POC

The first challenge for a company is to provide its employees with a secure version of a conversational agent like ChatGPT. To do this, it needs to deploy a general-purpose LLM (Large Language Model) instance in its environment and connect it to a chat-like interface. This initial step allows for the validation of initial hypotheses, such as the most appropriate language model or the adjustment of certain parameters that affect the model's creativity (such as temperature). This is indeed a Proof of Concept (PoC) with a limited user population.

This initial step also serves as a means for companies to gather use cases from employees through the questions asked, thus improving future versions of this product through various fine-tuning techniques.

While the risk of data leakage to the outside is mitigated, the work on the relevance of the information transmitted by the LLM remains to be further developed in subsequent development cycles.

HOW TECHNOLOGIES POSITION THEMSELVES AMONG EACH OTHER: OUR VISION OF AN 'MVP LLM APP STACK'



BUILDING INITIAL BELIEFS ABOUT A CUSTOMIZED LLM WITH COMPANY DATA

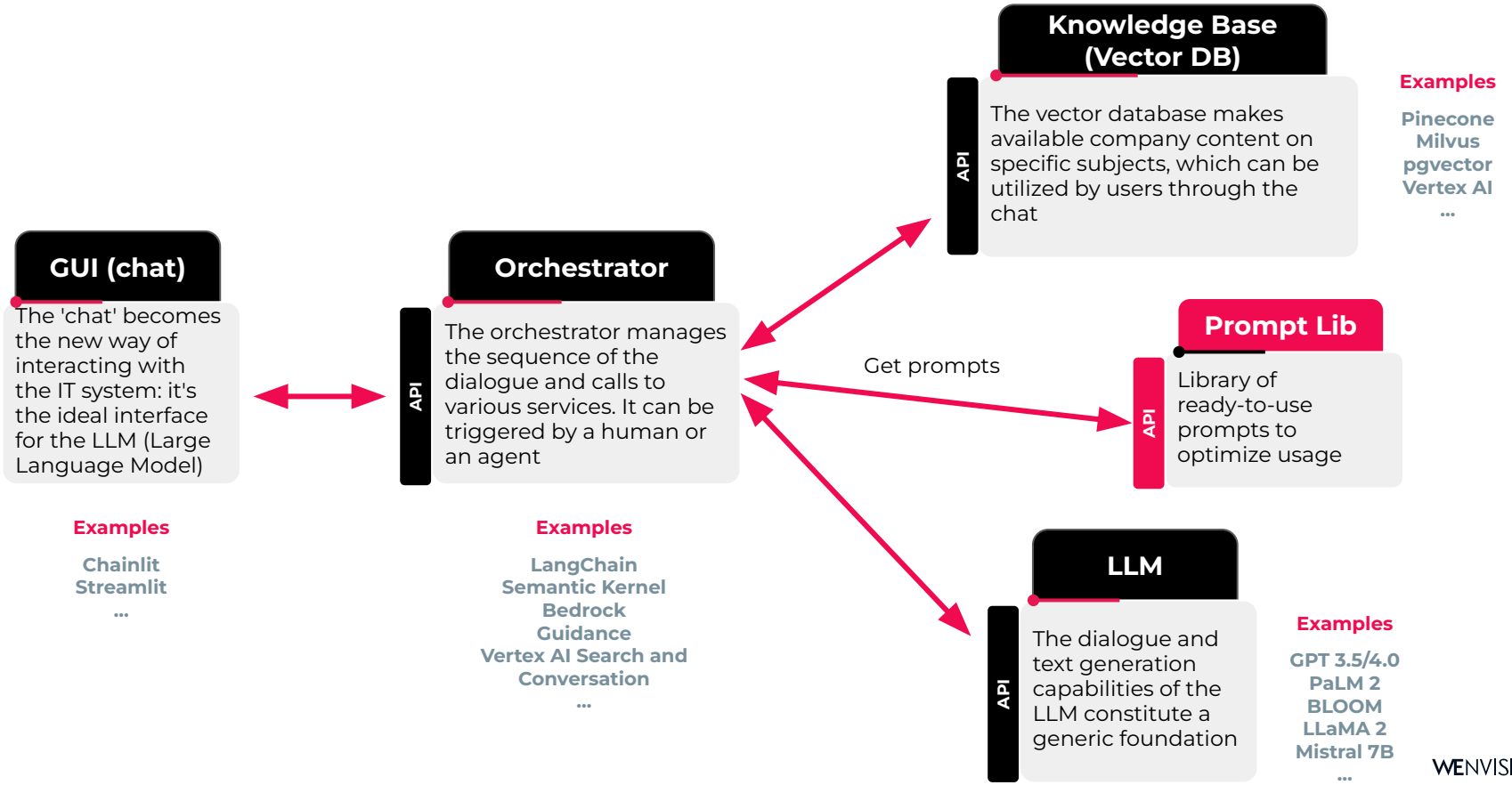
The main objective of the MVP stage is to personalize the generated content using company data. To achieve this, two new components come into play in this architecture: vector databases and an orchestrator.

The knowledge base, encapsulated within the vector databases, is separated from the LLM (Large Language Model), which only functions as a writer, synthesizing natural language text. The orchestrator ensures the sequencing of different services to maintain the continuity of the dialogue. The orchestrator can also retain information sources and provide them to the interface, allowing the user to retrieve the document on which the response was based. This stage allows for experimenting with the interaction between a data corpus selected by the company and an LLM.

The quality of the results will largely depend on the prompts formulated as input in the interface, and thus on the prompt engineering skills of the users.

HOW TECHNOLOGIES POSITION THEMSELVES AMONG EACH OTHER: OUR VISION OF A 'CLASSIC LLM APP STACK'

MAJ - R5

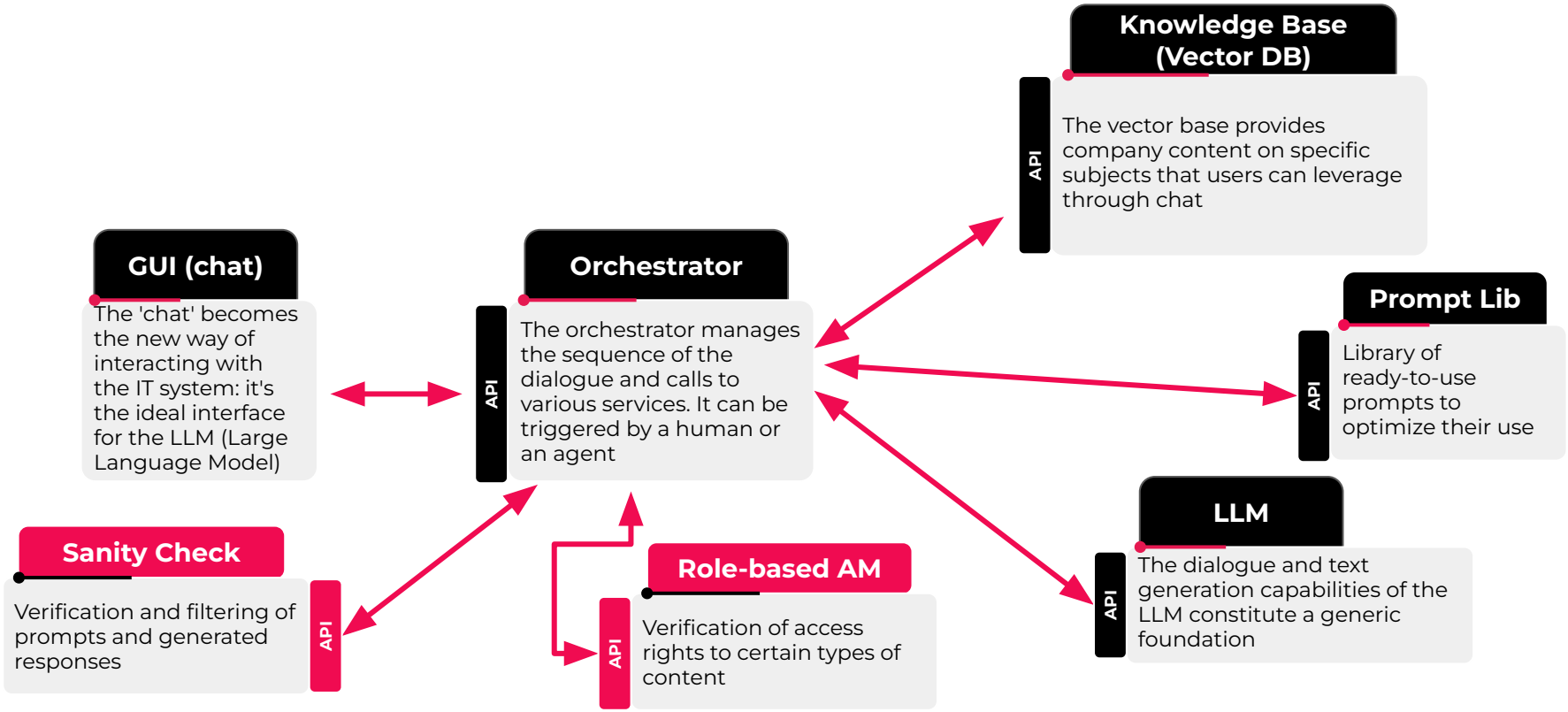


BENEFITING FROM PROMPT ENGINEERING WITH A CLASSIC LLM STACK

The intelligence lies in the prompt. The major addition here is the "Prompt lib" or prompt library component, which catalogs ready-to-use prompts and optimizes the use of the LLM (Large Language Model). The role of the prompt engineer becomes crucial during this phase. Their mission is to formulate high-quality prompts to achieve the best results, ensuring that the AI generates precise, relevant, and tailored responses to specific needs.

By integrating this component after analyzing the initial usage of the application by users, prompt engineers can focus on high-value use cases.

HOW TECHNOLOGIES POSITION THEMSELVES AMONG EACH OTHER: OUR VISION OF AN 'ADVANCED LLM APP STACK'



INDUSTRIALIZING AND SUSTAINING WITH AN ADVANCED LLM STACK

MAJ - R6

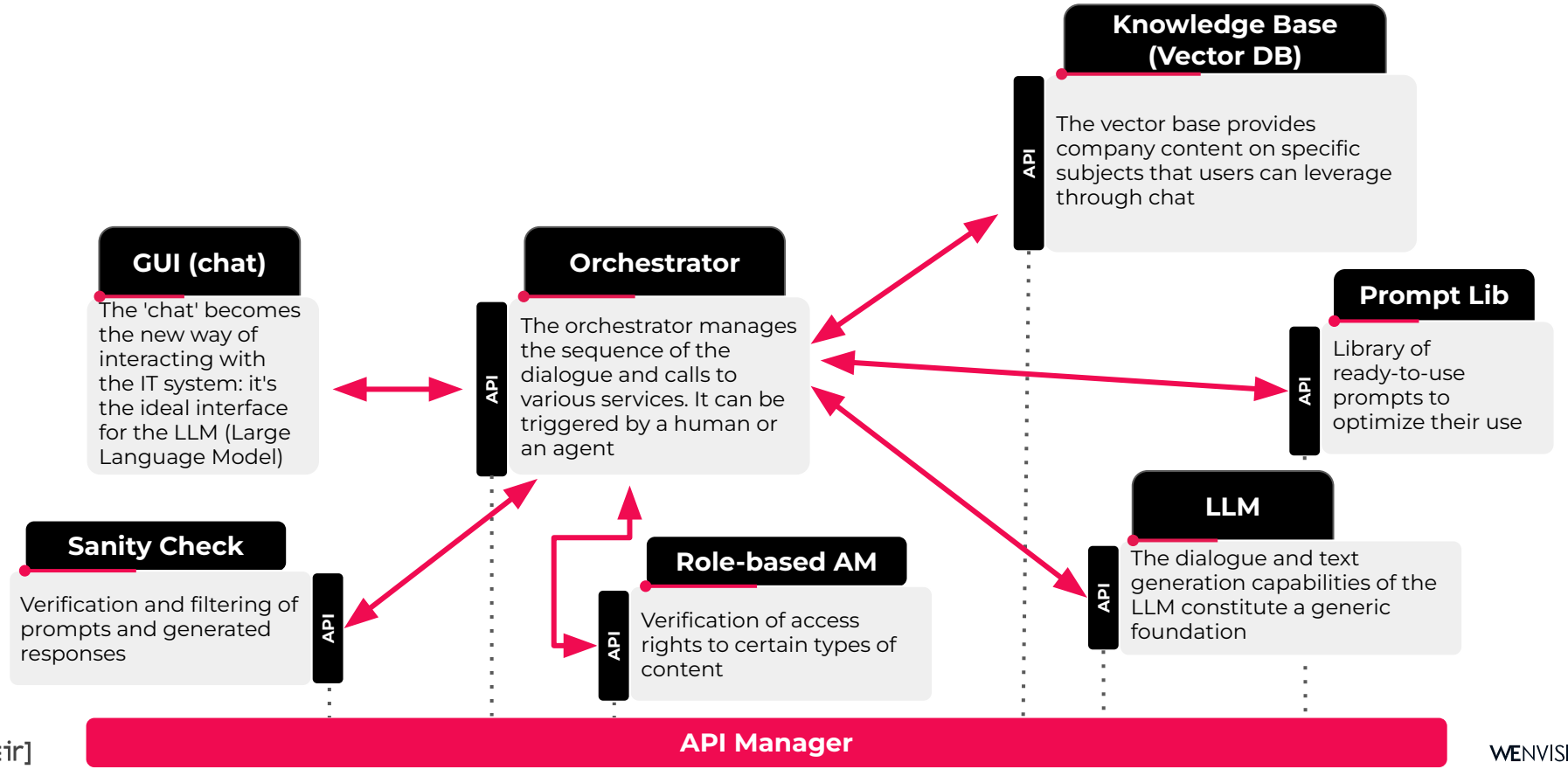
The architecture of an application based on LLMs (Large Language Models) must be able to address the challenges posed by its industrialization and scale of use, particularly in terms of governance and compliance.

Control over data distribution should not be limited to external factors; it is imperative to apply the data access management rules practiced by the company within the LLM application. This is why the orchestrator should be able to query the company's data access management tool.

Ethics and compliance are also significant considerations during the industrialization phase of such an application. A component like a "sanity check" ensures adherence to the company's ethical standards both at the input and output of the model. It also allows for monitoring the long-term quality performance of the model.

HOW TECHNOLOGIES POSITION THEMSELVES AMONG EACH OTHER: OUR VISION OF AN 'ENTERPRISE LLM APP STACK'

MAJ - R5



CREATING ASSISTANTS TO ENCAPSULATE AND MASTER COMPLEXITY

MAJ - R8

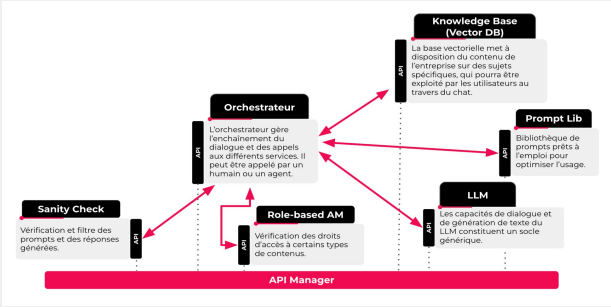
GUI (chat)

The 'chat' becomes the new way of interacting with the IT system: it's the ideal interface for the LLM (Large Language Model)

API

Assistants

The assistant compiles all the parameters to configure and the data to use to respond to a specific use case (e.g., studying legal documents, managing customer relations, resolving an incident, etc.).



THE GENAI PLATFORM: 3 CORE SERVICES AND THE CAPABILITY TO INDUSTRIALIZE SPECIALIZED ASSISTANTS

A GenAI platform enables meeting generative AI needs in the most efficient way possible. It opens up a world of possibilities for innovation and enhancing business operations, while keeping all options open in terms of providers, models, use cases, and integration with the company's IT and processes.

Generalist Chat Assistant

- Natural Language Chat via an Interface Connected to an LLM
- This assistant enables employees to have access to a highly generalist assistant, not specialized to the company's context, capable of addressing use cases such as drafting meeting minutes, composing emails, and generating summaries from text

Dynamic Document Assistant

- Natural Language Chat via an Interface Connected to an LLM with the option to upload a document into the conversation. The document will not be stored persistently, and the assistant can only interact with it during the chat session
- This use case leverages the capabilities of an LLM to extract and structure textual information from a document

Preconfigured Specialized Assistant

- A natural language assistant specialized in a specific industry vertical, addressing one or more specific use cases. This assistant will be pre-configured by an administrator through a dedicated interface
- This interface will allow for selecting the LLM, pre-writing instructions and recurring prompts, and loading a document corpus, such as a knowledge base, or connecting to a source like a CRM

APPLICATION OF #GENAI PRINCIPLES TO DATABASE MANAGEMENT SYSTEMS (MVP)

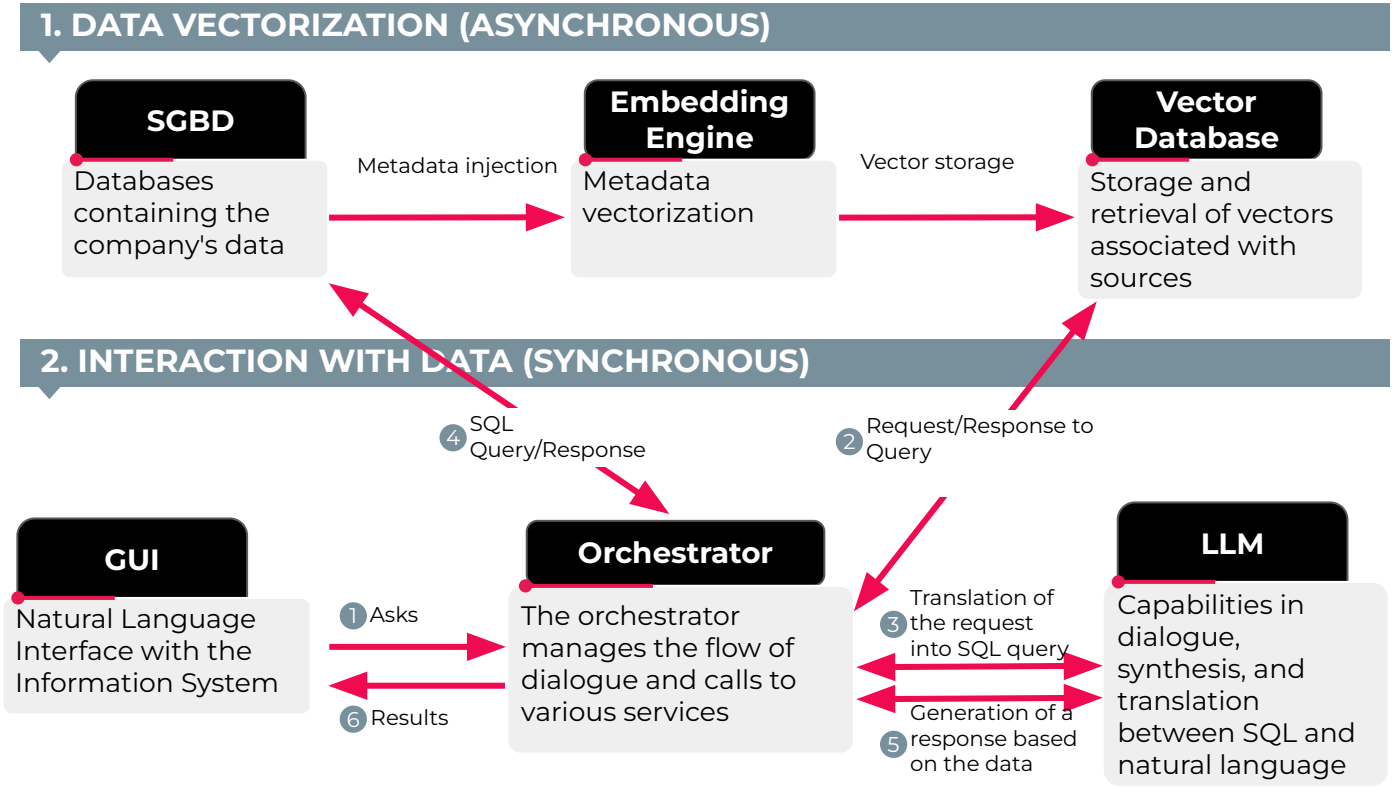
MAJ - R5

The dialogue is an ideal way to query enterprise data. It's no longer about querying data but collaborating with an assistant that responds to requests made in natural language.

Such an architecture breaks down into two major domains: firstly, data preparation, which happens asynchronously, and secondly, dialogue management in real-time.

However, caution must be exercised with this approach:

- It requires well-populated databases and intelligible metadata.
- It introduces security risks like SQL injection.



FINE-TUNING VS RETRIEVAL AUGMENTED GENERATION (RAG)

MAJ - RS

FINE-TUNING : DEFINITION

Fine-tuning involves customizing an LLM for specific tasks or particular domains. The process entails taking a pre-trained model and retraining it on a smaller, targeted dataset, often composed of task-specific texts. By adjusting the weights and parameters of the original model, fine-tuning enables the LLM to better comprehend and generate more accurate responses tailored to specific contexts.

USE CASE FOR FINE-TUNING

Adapting a model to a specific context, for example:

- Assisting a developer
- Adopting legal or literary sentence structures
- Understanding domain-specific vocabulary. Fine-tuning is a valuable process for achieving these adaptations.

THE PROS

- The ability to adapt the model to its context (e.g., retail, construction, law, etc.) or its task (e.g., customer support, providing instructions, etc.) is essential for tailoring the model's responses to the specific requirements and domains of the business or application. This adaptability ensures that the model can provide relevant and accurate information within a given context or for a particular task.

THE CONS

- Cost and Time: The process is relatively expensive and time-consuming.
- Difficulty in Forgetting: It's nearly impossible to make the model forget what it learned, except through extensive and costly fine-tuning.
- Lack of Access Control: There's no access control, making all model-learned information accessible to all users.

RAG: DEFINITION

The "RAG" leverages the built-in search capabilities of LLMs to query a vector database (see associated sheet) or documents in order to extract relevant information in response to a given query. The retrieved information is then used to enhance text generation, providing a more contextual and informed response, drawing on the company's knowledge rather than just the model's general knowledge.

USE CASE FOR RAG

Enriching the model with company data (while preserving access control and the ability to delete or update content), for example:

- Financial data
- Product descriptions
- Documentation
- Technical frameworks

This process can enhance the model's understanding and ability to generate content specific to the organization's needs.

THE PROS

- Quick process (possibility of continuous updates: as seen in language models that rely on Google or Bing to add current content to their responses)
- Access to company documents and data
- Ability to have granular control over information access based on roles
- Requires a lightweight generic language model

THE CONS

- This has no impact on how the language model will respond: this issue should be addressed in the prompt (prompt engineering is required).

CONCLUSION

RAG for the substance, then fine-tuning for the style, if necessary.

Fine-tuning may seem enticing and is promoted by publishers; however, it proves to be complex and risky to handle. If the goal is to enhance the relevance of responses, RAG will be faster and easier to implement, immediately adding value in conjunction with lightweight models designed for natural language interactions. Fine-tuning can come into play for shaping interactions and responses (if an optimized prompt library isn't sufficient).

MAJ - R6

RETRIEVAL AUGMENTED GENERATION (RAG): BEST PRACTICES

BASIC LEVEL

1. Prompt Optimization and Emphasis on Simplicity

Employ effective prompt engineering techniques to condition the model, starting with simple prompts.

2. Determining the Ideal Chunk Size

Identify the optimal size of data chunks to ensure efficient retrieval and processing.

3. Use Summaries for Conciseness

Apply summarization techniques to data chunks to provide the model with a concise representation of information.

4. Rigorous Data Management

Pay meticulous attention to data management, verification, versioning, and cleaning of data sources and pipelines. Data quality is essential for achieving high-quality Retrieval Augmented Generation (RAG).

5. Regular Evaluation and Continuous Adaptation

Frequently assess the performance of content retrieval and generation using use-case-specific metrics. Remain open to adaptation and continuous improvement, while also considering simpler alternatives to vector databases, such as PGVECTOR.

INTERMEDIATE LEVEL

1. Metadata Filtering and Relevance

Add metadata to chunks to facilitate result processing.

2. Embeddings Management

Develop strategies to manage documents that are frequently updated or newly added.

3. Reliability and Trustworthiness

Ensure the accuracy and reliability of generated content by using citations and applying techniques such as confidence estimation, uncertainty quantification, and error analysis.

4. Hybrid Search Techniques

Integrate various search techniques, such as keyword-based search and semantic search.

5. Query Transformation and Trade-offs

Explore strategies like hypothetical document embeddings, query decomposition into sub-queries, and iterative query evaluation to bridge information gaps. Consider trade-offs between precision, recall, cost, and computation to optimize the retrieval and generation process. Also, experiment with advanced chunking strategies and re-ranking of retrieved documents to improve overall performance.

ADVANCED LEVEL

1. Model and Embeddings Refinement

Refine the model by continuing training on a more specific dataset or optimizing embeddings to better represent data relationships.

2. Personalized Embeddings

Personalize embeddings by creating a matrix to emphasize relevant aspects of the text for your use cases.

3. Query Routing

Use more than one index or tool and route sub-queries to the appropriate index or tool.

4. Multi-Retrieval

Combine results from multiple retrieval (and generation) agents to improve overall quality and fidelity.

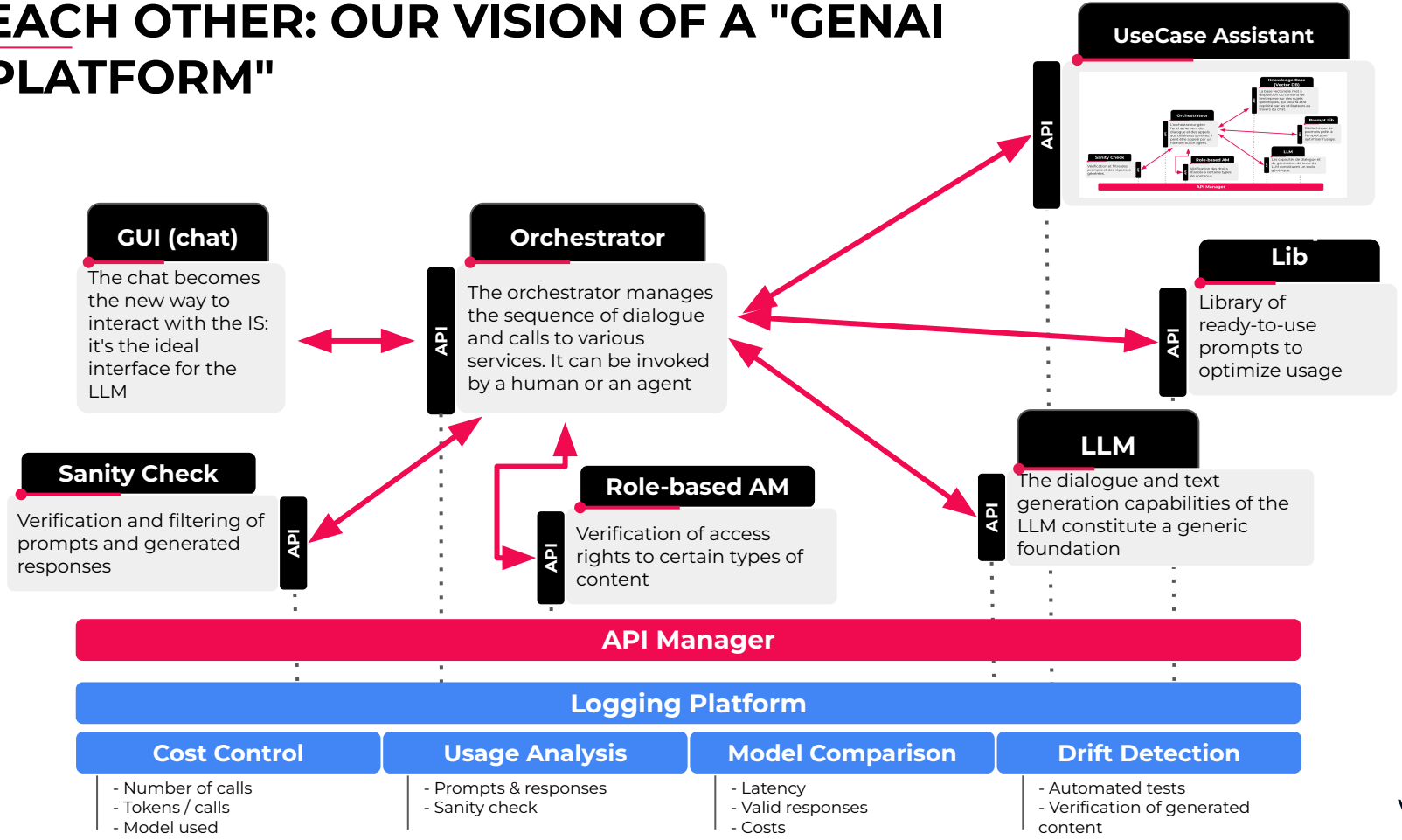
5. Contextual Compression and Self-Querying

Apply compression techniques to reduce context size while maintaining relevance. Use self-querying by using the model's output as a query to retrieve additional information and enhance the initial response.

For more information, please refer to Shyamal Anadkat's post on X (formerly Twitter) at https://twitter.com/shyamalanadkat/status/1746645405276975581 [Last access: 17.01.2023]

HOW TECHNOLOGIES POSITION THEMSELVES AMONG EACH OTHER: OUR VISION OF A "GENAI PLATFORM"

MAJ - R7



TAKING LLMs TO PRODUCTION

Sanity Check

It is crucial, for professional use of LLMs and GenAI in general, to monitor the generated content. It's essential to ensure that users are not exposed to offensive content or content that does not comply with the company's rules. **Sanity checks are integrated into some platforms, like Azure Open AI or Google Vertex, which by default offer content analysis**. It's also possible to adjust the alert level based on certain criteria. The alert level must be adaptable depending on the use case and be monitored.

Cost control and usages

Monitoring GenAI algorithms, like any technology, is crucial for fully understanding the costs involved. Additionally, logging the application's usage is critical. Logs enable retrospective analysis of technical KPIs (number of requests, tokens, latency, etc.) as well as business KPIs (types of requests, usage frequency, validity of the generation, etc.). **To choose the appropriate models, it is necessary to have comprehensive logs to calculate a suitable ratio between cost, latency, and relevance for each use case.**

Model comparison

With the proliferation of available models and the rapid development of new ones, it's crucial to maintain up-to-date knowledge of their capabilities. Depending on specific needs, **it's necessary to choose wisely among the available tools, taking into account expected latencies and the relevance of the responses.** All these criteria guide companies toward the most suitable tool. **Logs allow for reflection on usage and help guide this choice. Architectures must be independent of models and allow for an easy transition to new models to keep pace with the evolution of the field.**

Drift detections

Once models are deployed in production, it's imperative to monitor their usage and, most importantly, the relevance of their responses. **Some models may evolve over time, so it's essential to ensure their consistency with the intended use.** Several approaches can be considered, such as regularly running a set of tests with result verification, which can be done using different methods (embedding, other LLMs, temperature set to 0, etc.). Another solution is to freeze the model version to ensure it won't change, although this functionality is not available on all platforms.

MAJ - R7

FINOPS FOR GENAI

IT costs have always been at the center of business concerns, but the concept of **FinOps emerged with the advent of the cloud model**, **where the usage AND architecture of solutions have a direct impact on the costs charged**, due to their variability. When deploying a GenAI platform, the aspects related to induced costs are multiple, but one can imagine a framework that allows evaluating not only the impact of GenAI initiatives, their financial viability, the optimization options available to teams, but also the costs specific to each type of deployment. It is noteworthy that, just as generative AI is relatively recent, **the notions of FinOps for GenAI are also in the process of being established.**

GEN AI Delivery

In this segment of the framework, we group aspects related to the effective provision of AI models, as well as enablement and training campaigns for teams on AI. **It's essential to ensure that collaborators, regardless of their use of Generative AI (whether as consumers or platform creators), have the necessary knowledge to confidently benefit from it.**

GEN AI Platform Cost Allocation

Here, the focus is on how the costs associated with the GenAI platform are distributed. This involves precisely mapping expenses and allocating the costs of shared services. **This step is central for a company's ability to understand the costs of its GenAI platform and to make informed decisions.**

GEN AI Cost Optimisation

In cases where a model is being trained or enhanced (augmented), **it is necessary to continuously optimize costs to improve process efficiency without sacrificing quality.** Similarly, this section can include all ancillary steps related to improving the quality of the data used to feed the implementation.

GEN AI Pricing model & Business Value

A deep understanding of the values obtained and actions taken in an optimization approach allows for **assessing the financial viability of deployed genAI solutions and enables the validation** (or invalidation) **of investments made in this field.**

Human Cost + **Finops** = **Business Value**

FIVE BEST PRACTICES TO ADOPT TO AVOID SHADOW AI

MAJ - R6

"Shadow AI" refers to the use of artificial intelligence tools and technologies within an organization without formal authorization or control from IT or the company. This occurs when employees or business units decide to adopt AI solutions outside the framework established by the company for the adoption of such technologies.

Unauthorized use of AI can take various forms, such as using unapproved third-party AI software, accessing cloud-based AI services without IT approval, or utilizing generative AI tools like ChatGPT/Bard/Perplexity to perform various professional tasks.

The emergence of Shadow AI is often driven by the pursuit of efficiency and the need to quickly address specific needs. However, **this practice can pose significant risks to the company, including data security, regulatory compliance, and IT governance.**

5 BEST PRACTICES TO ADOPT TO AVOID SHADOW AI

Implementation of an Internal Gen AI Platform	Open Communication	Quick Approval Processes	Education and Training	Establishing Policies to Access Gen AI
Invest in the development of an internal platform specifically designed to meet your company's needs in generative AI	Encourage open communication between departments to understand their needs. Ensure that available resources meet requirements, thereby reducing the reliance on unapproved AI tools	Simplify the approval processes for new AI tools to prevent employees from turning to Shadow AI due to slow approvals	Raise employee awareness about the risks associated with Shadow AI through regular training sessions, so that everyone understands the implications of unauthorized AI usage	Establish clear policies regarding employee access to generative AI tools, specifying where, when, and how they can use them, taking into account the potential benefits and risks to data security and confidentiality

WHICH GOVERNANCE FOR A RESPONSIBLE AI?

MAJ - R8

Data Governance

Ensuring the quality, security, freshness, and usability of data throughout the organization

- Data Quality
- Access and Confidentiality
- Knowledge of Assets Data
- Architecture
- ...

&

AI Governance

Ensure that the use of AI systems is ethically responsible, transparent, secure, and compliant with regulations

- AI System Security
- Transparency and Explainability
- Ethics and Accountability
- Human Oversight
- ...

- Roles and responsibilities
- Regulatory compliance
- Training & awareness
- Process & standards

DATA GOVERNANCE PLAYS A MAJOR ROLE IN IMPROVING EFFICIENCY

MAJ - R5

Strong data governance is not just an accessory of GenAI, but a **fundamental cog in its functionality and value**.

QUALITY

The efficiency of a model directly depends on the quality of the input data. Effective data governance can maintain and improve data quality by ensuring that the correct, relevant, and consistent data are retrieved and used in the generation process.

RELEVANCE

A data governance framework must ensure that the most relevant data are extracted for the task at hand. For instance, **properly cataloged data can help quickly identify the right data sets** for fine tuning.

COHERENCE

It is essential that the data used in the models be consistent. Data governance strategies ensure this consistency across different data sources, making the results more reliable, especially for the RAG (Retrieval-Augmented Generation) process.

ACCESS

If a model accesses sensitive data and uses it, data governance is essential to ensure that this information is handled correctly, in order **to avoid violations, leaks, or other misuse of sensitive data**.

DISTORSION

Effective governance allows for the identification and reduction of biases in the data used in models. **Biases can lead to phenomena of hallucinations, where the model generates information that may be erroneous or inappropriate**. Governance ensures that the data are balanced and representative.

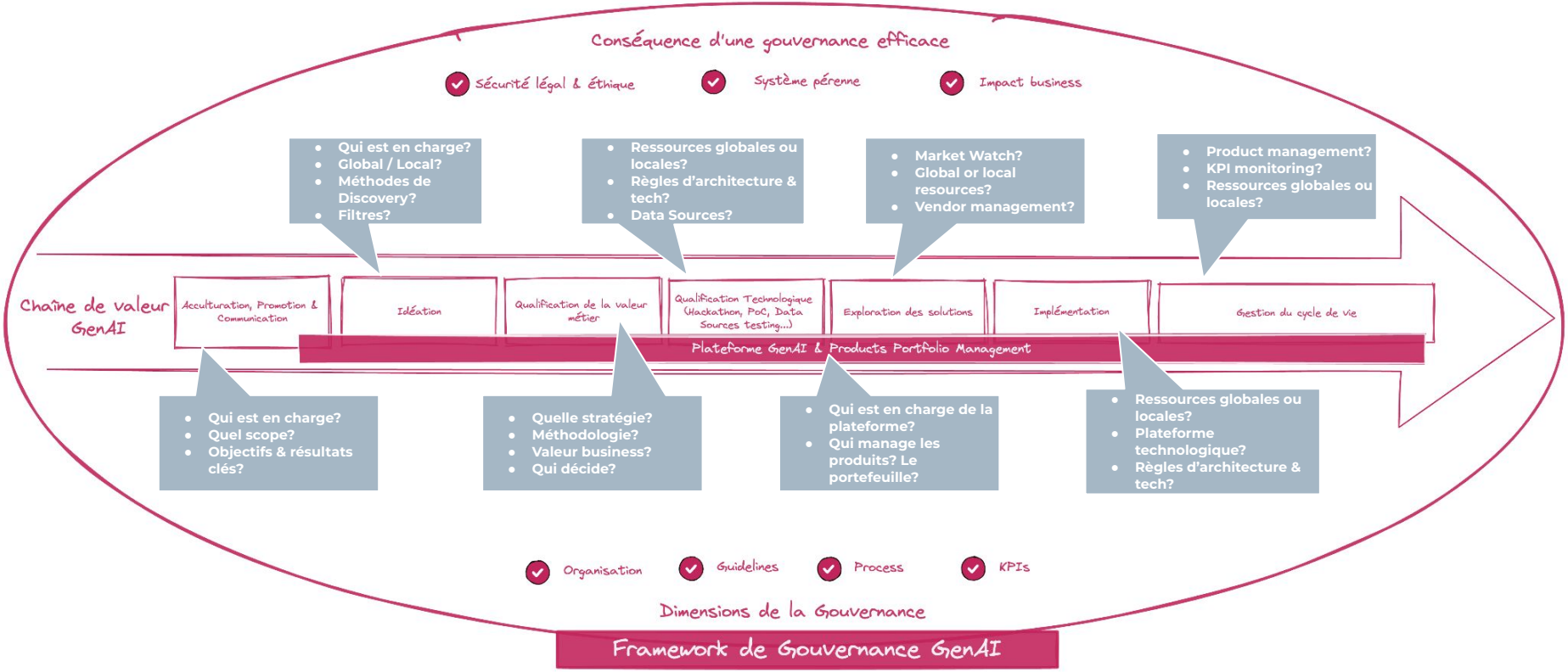
CONFORMITY

Data governance ensures that the **models comply with all relevant data privacy laws and regulations**, thereby avoiding legal complications.

KEY QUESTIONS THROUGHOUT THE VALUE CHAIN

MAJ - R8

Facing the challenges posed by GenAI requires establishing a **robust governance framework**. This framework must, at each stage of the value chain, provide answers to essential questions to ensure the implementation of **responsible AI**.



FOSTERING A CULTURE OF DATA GOVERNANCE IS ONE OF THE KEYS TO THE SUCCESS OF GENERATIVE AI

MAJ - R8

Gen AI promises revolutionary advancements capable of generating innovative content from vast amounts of data. However, to realize this potential, one element must be considered: effective data governance. The presence of poor-quality data or data with biases can compromise the reliability of results and the sustainability of algorithmic models. This is where training in data governance becomes indispensable.

TRAINING IN DATA GOVERNANCE TO AMPLIFY THE IMPACT OF GENERATIVE AI

Training in data governance involves every employee in data collection and adherence to quality standards. By making everyone a participant in this process, costly errors can be avoided, and the efficiency of generative AI technologies can be maximized.

TRAIN TO PROMOTE ETHICS AND INDIVIDUAL RESPONSIBILITY IN DATA USAGE

The growing importance of data also highlights ethical challenges related to biases in data sets that can influence AI outcomes. Training employees on these issues promotes responsible data management by encouraging users to adopt a critical perspective towards the tools and data they handle.

BUILD A DATA CULTURE

Finally, establishing a data culture through training is important so that each stakeholder can understand the role data plays in the strategic and operational success of a business. Through these trainings, it's also about creating a momentum around data and its stakes to raise awareness among all its users.

COMPARING LLMs: PUBLIC BENCHMARKS

Definition

Language models are evaluated on various tasks to assess their effectiveness and **facilitate comparisons**. Since LLMs can perform a wide range of tasks, it is crucial to evaluate their performance based on the specific requirements of our applications. To do this, open-source benchmarks allow us to test and compare model performance.

We have categorized these benchmarks into several categories based on the type of problem the model needs to solve. In our applications, we typically require a specific use of LLMs (such as code generation, augmented search, etc.). Based on this, we can assess model performance to select the one that best suits our needs.

Our Expert Eye: the Few-Shot Evaluation

In some evaluation datasets, models have been more or less assisted in the generation of their responses. This is what is known as few-shot learning, a technique that involves **providing the model with examples of questions and answers to guide it toward the expected type of response.** This helps steer the model towards the task we want it to accomplish.

For example, the **MMLU** benchmark was evaluated in 5-shot mode, which means that the model received 5 examples of answers before being evaluated. This is preferred for testing a model's ability to provide the expected response format. In contrast, the **HumanEval** benchmark is in 0-shot mode, which means that the model is simply asked to solve the problem without any example solutions. This test is rarely used on its own because it is much more versatile.

Different categories of benchmarks allow to identify the best LLM based on its usage.

World of Knowledge

LLMs can be used to answer questions about **general knowledge or specific documents**. These problems are challenging because the questions can be complex, the answers implicit, and the facts hard to identify. To address this, benchmarks have been created, such as **TriviaQA (Joshi et al., 2017)**. This benchmark compiles a set of questions to assess a model's ability to answer questions in natural language.

Code Generation

LLMs can be used for code generation. To evaluate the relevance of this application, a set of benchmarks has been created. The **HumanEval benchmark (Chen et al., 2021)** compiles a collection of code generation tasks. For each task, the goal is to test whether the model is capable of generating code that passes unit tests. The tasks are diverse and cover a wide range of development tasks.

Mathematical Problems

LLMs have made significant progress in various tasks, but they are not yet capable of solving complex multi-step mathematical problems. Specific benchmarks like **GSM8K (Cobbe et al., 2021)** have been created to assess the relevance of LLMs for simple mathematical problems. These benchmarks are useful for evaluating LLM models' understanding and their "reasoning" abilities.

Aggregated Benchmarks

LLMs are trained on a massive amount of information, including specialized subjects. However, it is still challenging to determine the extent to which these models can learn and apply knowledge from various domains. To evaluate them, generalized benchmarks have been created, such as the **MMLU (Hendrycks et al., 2020)**. This benchmark covers 57 subjects, ranging from social sciences to philosophy to physics problems.

Fresh LLM

MAJ - R7

One of the major challenges of Large Language Models (LLMs) lies in their inability to manage and integrate new and updated knowledge. **Once trained, LLMs do not benefit from real-time updates and tend to produce false or outdated answers**. In response to this issue, significant progress has been made by several researchers at Google. Presented in the paper **"*FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation*" (2023)**, their solution is based on the development of a new benchmark titled FreshQA and a new methodology, FreshPrompt.

FreshQA is a new dataset designed to test the ability of LLMs to handle questions requiring up-to-date knowledge, including those that evolve rapidly or are based on false premises. Concurrently, **FreshPrompt is a contextual learning method that integrates relevant and recent information extracted directly from Google Search into the prompts of LLMs.**

Unlike Gemini or Copilot, which use search engine results to answer questions without necessarily integrating this information contextually or updating the model itself, **a Fresh LLM knows how to directly refresh its knowledge**. FreshPrompt allows for a more sophisticated and targeted integration of recent data by extracting and incorporating relevant information from web searches directly into the prompts of LLMs. This means that LLMs can generate responses not only based on their initial training but also enriched by the most recent external data.

FreshQA

Designed around **600 questions divided into four categories**, FreshQA is a **dynamic benchmark** that challenges LLMs on topics requiring evolving knowledge. **The model was trained on over 50,000 questions asked by users.** Two modes of evaluation were possible: a "relaxed" mode that focuses on the accuracy of the main answer, and a "strict" mode that verifies the factualness of all statements made in the response. **The results reveal the limitations of LLMs in dealing with changing knowledge, false premises, and multi-step reasoning, regardless of the model size.**

FreshPrompt

FreshPrompt improves LLMs by integrating the latest web information, thus bypassing their limitations on knowledge updating. **It starts with the collection of varied data via Google Search.** This collection includes not only direct answers but also more nuanced elements like related user queries. **Key information from these data (sources, dates, titles, and standout keywords, etc.) is then extracted, organized chronologically in the LLM's prompt, and enriched by targeted demonstrations. This process thus guides the model towards logical, precise, and updated responses**, even for questions with erroneous premises.

Quelles perspectives pour les entreprises ?

The introduction of the FreshPrompt concept invites a rethinking of LLM use within companies. **To remain competitive, they must ensure that their LLMs are provided with "fresh" internal information**—whether it be sales figures from the previous day or even real-time sales data, as well as recent news and internal events. **The need for updated internal data goes far beyond the scope of using Google searches as an information source for LLMs.** By updating their data, companies could not only enrich the quality of analyses produced by LLMs but also increase their ability to proactively respond to market developments.

TECHWAVES “GEN AI FOR CIO & CTO”

[sfeir]

WENVISION

METHODOLOGY FOR OUR TECHWAVES

The Techwaves are the product of a proven methodology at SFEIR and WEnvision, dedicated to identifying the most relevant technologies within a domain. These technologies empower companies to realize their ambitions without amplifying their technical debt.

To do this, we evaluate various criteria, such as user-friendliness, the growth of a community, the presence of training modules or even books, among others.

This task is feasible thanks to internal crowdsourcing within the group, which equally enriches our technological surveillance and our analyses.

The findings from these analyses enable us to place the technologies on our curve, segmented into six sectors.

1.

EXPERIMENTATION

Deals with developers and innovative companies, start-ups, looking for ways to gain a decisive edge through technology. They test a lot, carrying out "proofs of concept."

2.

TAKEOFF

The technology is adopted by start-ups and CIOs who are willing to bet heavily on it in order to gain a distinguishing advantage. For them, it is vital that the project succeeds.

3.

GROWTH

Slightly more cautious companies, yet still wanting to quickly leverage innovative technologies, in turn adopt the technology.

4.

PLATEAU

Technology has found its market. The more conservative ones, who wish to go with proven technologies, have adopted it. The market is able to meet the demand.

5.

DECLINE

Fewer and fewer new projects are adopting this technology. Its usage is declining. Using this technology can be equated to accruing debt.

6.

THREAT

Using this technology for a new project is to be avoided. Few players are still involved with it.

GENERAL LLMs

MAJ - R8

Definition

A Large Language Model (LLM) is a type of machine learning model that has been trained on a large volume of text data. This enables LLMs to "understand" existing content and generate new content, such as text, code, scripts, song lyrics, emails, letters, and more.

LLMs are still in the process of development, but they have the potential to revolutionize a wide range of industries, particularly in software development, marketing, sales, customer service, and new product development functions.

Use Case

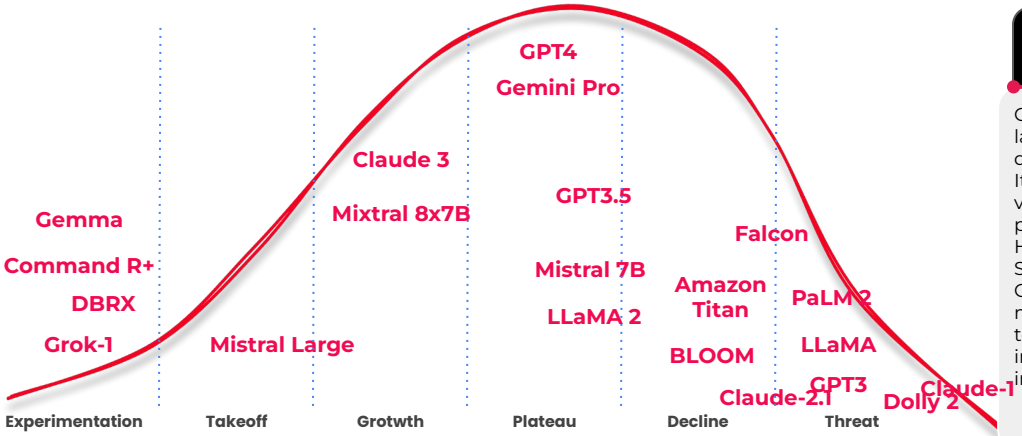
LLMs can be used in various domains, including natural language processing (text classification and sentiment analysis), automatic translation, creative writing, and question answering. They are also valuable for code and text generation, aiding in data analysis, fraud detection, and customer segmentation.

Our Expert Eye

The field of Large Language Models (LLMs) is constantly evolving, with leaders such as OpenAI's GPT-4, Anthropic's Claude, and Google's Gemini currently dominating the market. However, open-source models have their own advantages, especially in terms of customization and on-premises deployment.

These technologies are rapidly advancing, and it's crucial to stay up-to-date with the latest trends and regulations, including the European Union's AI Act. It's essential to develop flexible solutions that allow for easy replacement of one LLM with another.

LLMs are not universal, so it may be wise to use different models for different tasks. To make the most of LLMs, companies need to stay informed about best practices and technological advancements while complying with current regulations. The key to success lies in adaptability, compatibility, and choosing the right model for each specific task.



Focus sur...

ANTHROP\C

Claude 3 is the latest language model developed by Anthropic. It comes in three versions with increasing performance levels: Haiku, Sonnet, and Opus. Similar to GPT-4 or Gemini Pro, Claude 3 is multimodal, enabling it to process both text and images in its interactions.

Google

In December, Google announced the launch of Gemini, new multimodal AI models capable of understanding and generating text, as well as understanding images and videos. They come in three sizes: Nano, Pro, and Ultra, with increasingly advanced skills. Note: the Ultra model is not yet available.

MISTRAL AI_

Mistral AI is a French startup specializing in AI, offering the Mistral-7B and Mixtral-8x7B models. These are currently the most powerful open-source models and have demonstrated the strength of the Mixture of Experts (MoE) approach. Their latest model, Mistral Medium, promises even better performance.

ORCHESTRATORS

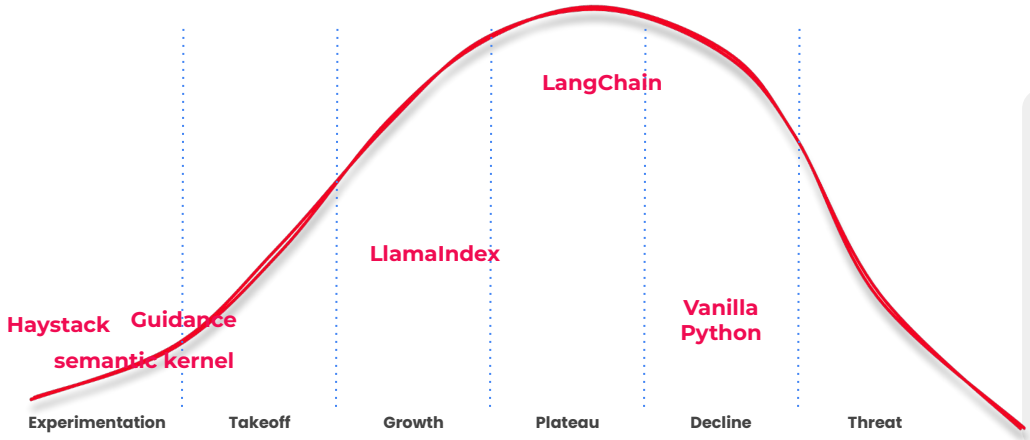
MAJ - R8

Definition

LLM orchestrators are designed to streamline interaction between users and LLMs by handling complex tasks that typically require multiple API calls. They assist in structuring conversations, managing dialogue state, handling inputs and outputs, and controlling interactions with the model.
These orchestrators are valuable tools for fully leveraging the capabilities of LLMs, making it easier to interact with the models and providing additional features for more efficient and tailored use based on the specific needs of users.

Use case

These tools can be used to build chatbots capable of generating text, translating languages, answering questions, and more. They also enable the automation of tasks that are currently performed manually and integration of LLMs with vector databases to create more powerful applications. In the end, orchestrators enable the creation of autonomous agents capable of planning and executing actions.



Our Expert Eye

These tools have only been available for a short time but have quickly become essential for building applications based on Large Language Models (LLMs). They not only accelerate development but also add additional functionalities to LLMs, making them indispensable when creating a tool. Whether it's for prompt management, context optimization, API call chaining, or document vectorization for integration into the model, orchestrators simplify and speed up the integration of these functionalities.

One of the most popular orchestrators is Langchain, an open-source initiative that has seen remarkable growth, going from 0 to over 49,000 stars on GitHub in just a few months. This framework allows connection to the most commonly used models and the execution of API call chains with external APIs. Thanks to an active and engaged community, new components are regularly added, enriching the tool's functionalities almost daily.

Regarding hyperscalers, Microsoft has released Guidance as open source. Although less used than Langchain, it remains a solid competitor for the future.

Focus sur...

LangChain

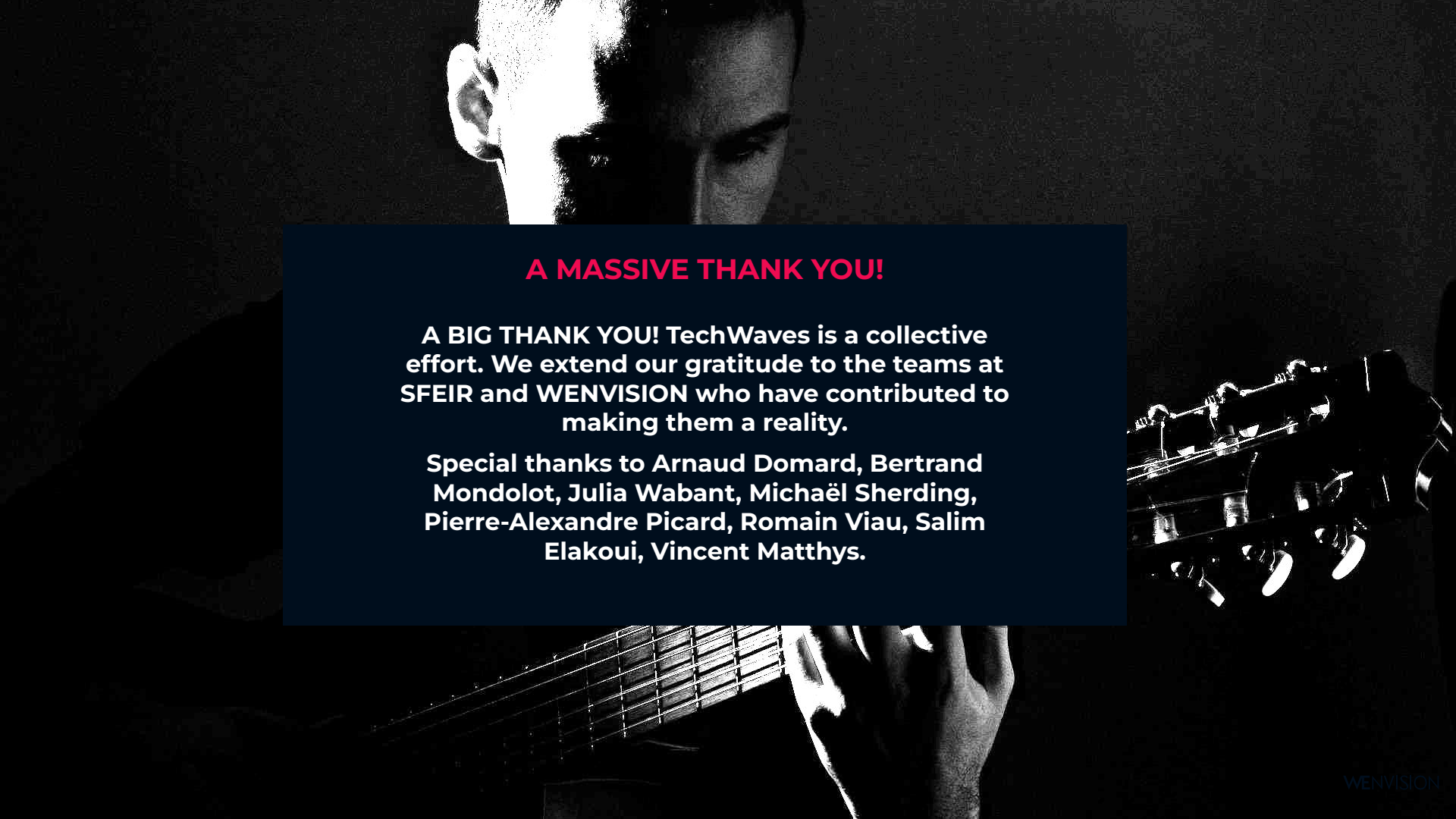
LangChain is an open-source framework for creating applications using Large Language Models (LLMs). It offers a high-level API and supports multiple LLMs to facilitate the development of applications such as chatbots, document summarization tools, code analysis tools, personal assistants, etc.

Microsoft

Microsoft Guidance is an open framework for controlling LLMs, with a simple syntax. It supports various output structures, prototyping, and caching. Potential applications include text generation, translation, and user interaction in chatbots.

Google

Vertex AI Search and Conversation (formerly Gen App Builder) from Google Cloud enables users to quickly create enterprise generative AI applications using models, search tools, and conversational AI technologies. It allows for the creation of chatbots, digital assistants, etc., in just a few minutes.



A MASSIVE THANK YOU!

A BIG THANK YOU! TechWaves is a collective effort. We extend our gratitude to the teams at SFEIR and WENVISION who have contributed to making them a reality.

Special thanks to Arnaud Domard, Bertrand Mondolot, Julia Wabant, Michaël Sherding, Pierre-Alexandre Picard, Romain Viau, Salim Elakoui, Vincent Matthys.

[sfeir]

WENVISION

TECHWAVES - #GENAI FOR CIO & CTO

www.sfeir.com

www.wenvision.com

Techwaves GenAI

The analysis of experts from SFEIR and WEnvision on GenAI solutions

If you're feeling lost and don't know where to start with GenAI, the teams at SFEIR and WEnvision have compiled market solutions to provide you with a condensed and ranked overview on a technology trends curve of GenAI solutions.

Techwaves are updated every 6 weeks and can be found on:
consulting.wenvision.com and sfeir.com