Unlocking Gen AI's true potential
## in the enterprise with Ema, a Universal AI Employee

Authors: Surojit Chatterjee, Souvik Sen

Introduction

The rapid rise in Generative AI (Gen AI) has been widely acknowledged as unprecedented and this emerging technology category is generating hype heights rarely seen before. From the boardroom to the dining room and every place in between, AI is the topic of conversation today. Reaching the highest ranks of public popularity, mentions of AI technology found their way into new song lyrics and were even debated in Hollywood talent discussions. The technology has moved from what was originally just text-based outputs, to creating images and films, and even uncanny replications of voices and people.

Generative AI has gripped the business world, due to both its rapid improvements in output and its potential for transformation. With Gen AI, the ceiling of automation and productivity gain possibilities has been greatly raised. But, for most enterprises eager to take advantage of this promising technology, the value so far has been limited. Despite all the hype, there are some major concerns and challenges for the effective and safe use of Gen AI in the enterprise.

Generative AI
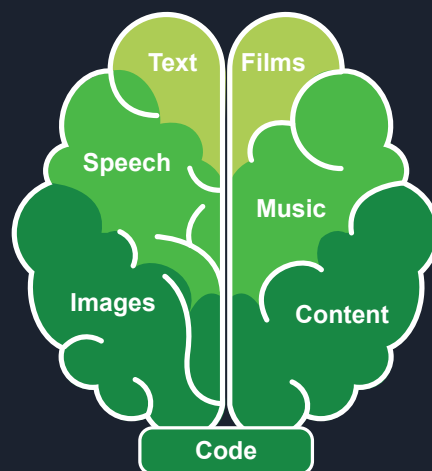
The key challenges enterprises face when deploying Gen AI are: (i) inconsistent quality of answers (known as hallucinations); (ii) high risk of data leakage, IP & copyright issues, and (iii) complexity of a fragmented Gen AI tech stack. Despite the continuous influx of new tools into the market, these fundamental issues persist, resulting in a diminished return on investment and consequently causing enterprises to hesitate in adopting them. Indeed, many enterprises are opting to pilot Gen AI experiments. But they are holding out adoption until new technology solutions arrive to solve these fundamental roadblocks.

> **Will 2024 be the year of turning the Gen AI hype train into true value for the Enterprise?**

What is Generative AI?

Generative AI (Gen AI) refers to a type of artificial intelligence (AI) that can generate new content, from usually text-based instructions, also known as prompts. These prompts are in natural language, allowing non-technical users to use the tools and capabilities of companies such as OpenAI, Anthropic, Mistral, MidJourney, and many others. Gen AI models are trained on large data sets for specific purposes, such as writing, creating images, videos, etc. With vastly different amounts and/or types of training, each model creates an output that is unique to it, even if they are created by the same parent company. These large generative models (e.g., GPT 4, Claude 2.1, Gemini, Mistral-S, Llama2) are also referred to as Large Language Models (LLMs). LLMs are a specialized class of AI models that use natural language processing (NLP) to understand human prompts and can reason and generate content. These large language models are often called foundational models as well.

Gen AI evolves as it continues to train on more data. Unlike traditional AI, Gen AI is trained on unlabelled data sets using unsupervised learning. It requires lots of computing power to create and train the underlying neural networks to mimic how humans act, react, and create. This ability to "create" something new from something that exists is very different from how AI has been used in the enterprise to date.

The rise of GenAI is largely because people can use natural language to prompt the AI, making it more accessible to many. As a result, the potential use cases for it have multiplied. Across a multitude of industries today, GenAI is being used as a creative assistant for writing, research, coding, designing, and so much more. But this is just the beginning.

Gen AI landscape

Today, Gen AI tools can be split into four categories. Each category provides a capability that is used in the world of work to complete various tasks. These tasks vary in frequency, complexity, and business impact. Here is a breakdown of the four categories of tools and the activity they are useful for:
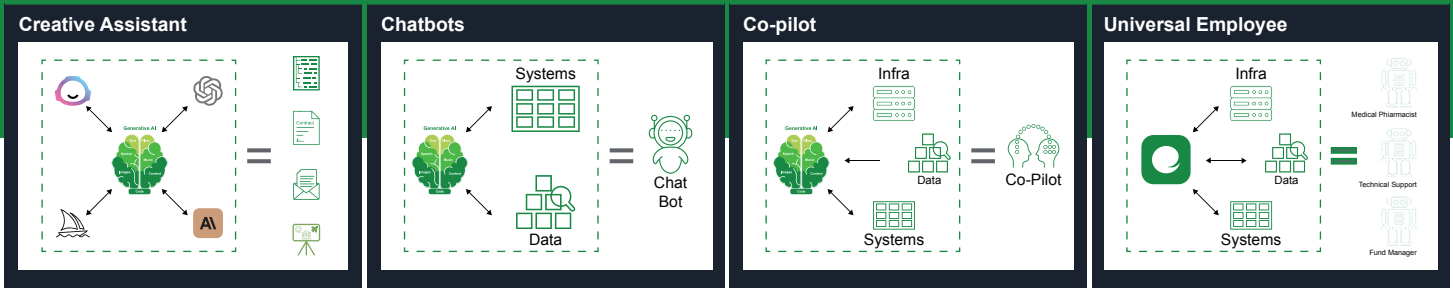
Creative Assistants - Creative Assistants captured the world's imagination by showcasing their ability to perform a diverse set of tasks, from writing articles to easily explaining complex concepts to a five-year old. These Gen AI-driven tools excelled in generating versatile content in a coherent and human-like manner and quickly shot to fame across the world.

In the realm of content creation, Creative Assistants have shown their worth. They can draft blog posts, marketing materials, images and reports at remarkable speeds, saving businesses valuable time and resources. Additionally, they can be used for data analysis, code creation, and even helping with creative brainstorming by providing new perspectives and ideas. However, their usefulness within an enterprise setting is inconsistent. While Creative Assistants may excel at certain tasks, they fall short in others. Their output lacks the depth of expertise that a human specialist can provide, making them more suitable for one-off tasks like generating initial drafts or simplifying concepts rather than continually augmenting an enterprise employee base.

Chatbots - Chatbots represent a natural advancement in the realm of Gen AI tools, particularly in the domain of chat and voice bots. These AI-powered bots are designed to engage in natural language conversations with users, whether through text or voice, and have witnessed a remarkable transformation with the integration of Gen AI technology.

Chatbots have evolved to become significantly more intelligent and versatile. They are no longer limited to predefined scripts or basic responses. Instead, they can understand context, interpret user intent, and provide relevant and informative answers to a wider range of questions. This capability is made possible by the foundational knowledge a large language model (LLM) already has.

One of the remarkable features of modern Chatbots is their capacity to adapt their tone and

style to match specific question types or user preferences. They can switch from formal to casual language, inject humor when appropriate, or even mimic the tone of a specific character or brand persona. This adaptability not only enhances the user experience but also makes these agents more relatable and engaging.

Co-Pilots -
Possibly the most widely used evolution in the realm of Gen AI tools. These AI systems are designed to work alongside people, acting as conversational interface alternatives to usual user interfaces. Co-pilots assist with simple tasks, problem-solving, and decision-making. Co-Pilots are highly specialized and can provide valuable insights into the specific application they are deployed in.

Co-Pilots have found their niche in specific applications, where they aid the masses in making informed decisions and navigating intricate processes. Unlike Creative Assistants, Co-Pilots offer consistent value to enterprises by enhancing productivity, reducing errors, and helping with problem-solving tasks. Their tangible impact on efficiency and decision quality makes them a valuable asset in the workplace.

As organizations harness the power of Co-Pilots, they are witnessing improvements in efficiency across the micro tasks they are being deployed to be used with. Examples include email responses, summarizing conversations, and expert questioning to assist in decision-making. Each task on its own just saves minutes per use, however, Co-pilots are used multiple times regularly across almost all employees making these small savings add up. Co-pilots have many benefits, but they cannot work across multiple applications and datasets and are not able to perform tasks that require workflow generation.

Universal AI Employee -
The Universal AI Employee represents a groundbreaking advancement in the world of Gen AI tools. Unlike the previously mentioned categories, which serve specific functions or task types, the Universal AI Employee is a versatile and all-encompassing AI system that can perform a wide range of complex tasks across various domains.

At its core, the Universal AI Employee is designed to emulate the capabilities of a human employee in a specific role - they can engage in conversations, comprehend context, take continuous human feedback, reason, and make informed decisions. They excel in areas requiring substantial (readily accessible) knowledge to dynamically create and evolve workflows that span across multiple enterprise applications or datasets.

While a Universal AI employee can assist in administrative tasks, automate processes, provide data-driven insights to support decision-making, and even collaborate with human employees on complex projects, they are also suited for process and role types that require a greater level of expertise. Think complex, specialized workflows within eCommerce, Healthcare, Life Sciences, Financial Technologies, and many more.

Ema

How to drive value with Gen AI

As with any AI & automation tooling, it is key to solve problems that are measurable in their current and future state to be able to track benefits. The business case for Gen AI will be catalyzed by the amount of time or money being saved. It's unlikely any CFO will sign off on investment in a tool just because of its hype in the market. So how should we think about Gen AI and the value that it can bring?

Firstly, we should start with the types of processes that are a good fit for this type of technology. We believe that ideal processes prime for impact from the use of Gen AI are those that require a level of creativity. This is because, unlike the previous automation technologies that required a strict set of rules in order to provide a specific output, with Gen AI, for the first time, we have an opportunity to not have to know the exact outcome we require.
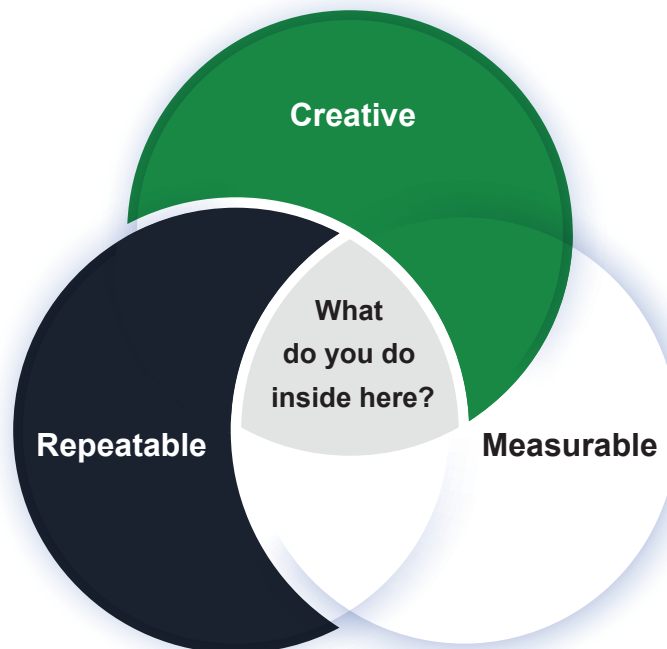
We must stress, though, that being creative doesn't mean that you are painting the next Van Gogh masterpiece, or writing the next Taylor Swift mega-hit. Creativity in this context means ability to generate content or summarize existing content.

Some examples of which are below:

Customer Support Agent responding to customers' questions based on past tickets or knowledge base articles

A medical professional reviewing medical history and authorizing a requested treatment

A marketing exec working on a product description

A technical support representative answering a technical question

A financial advisor reviewing the overnight news and providing a summary to his colleagues/clients

A procurement specialist writing an RFP or a responder writing a response to an RFP question set

This list goes on, but the common trait here is that very few other forms of AI and automation software to date have managed to make a significant dent in these types of tasks from a productivity perspective. This is why leveraging Gen AI in these types of processes, as opposed to many others, is the right approach if we aim to make the biggest impact in the enterprise. Choosing the right process type, though, is never enough.

There are two other areas we must still consider when finding true value:

Firstly, we have to be able to measure the impact of Gen AI improvements, and we all know if you can't measure something, you can't improve it. Therefore we have to ensure that before we start making improvements, we can measure the baseline -just like in many other AI & automation initiatives.

Baseline measures could include:

1. Average Handle Time
2. Average Task Duration
3. Accuracy Percentage
4. Speed to Answer
5. Hourly, Daily, Weekly, Monthly Volumes
6. Sales Volumes
7. Right First Time
8. Customer/ Employee Satisfaction

Ema

Privileged and Confidential. Ema Unlimited Inc., 2024 | 6

Secondly, and most importantly, the task/process needs to have a high level of repeatability in order to generate a return on investment. High level repeatability means the general direction of the task you want to complete is defined, but the specifics such as a need for an input template, a templated response, etc are not needed to be defined exactly, as we are dealing with creative processes.

Issues with Adoption

Early adoption of Gen AI has been difficult for the Enterprise. Initial issues ranged from concerns around current solutions' inability to show ROI, being slow & complex to deploy, and having a high risk of sensitive data leakage.

For an Enterprise to make widespread use of Gen AI it has the age-old choice of buy vs. build. The challenge with the build option this time, perhaps more so than ever, is the relative lack of maturity of the entire Gen AI stack. The current Gen AI toolchain is fragmented between many different complex components with hundreds of different vendors providing each component. To make sense of it all and to stitch together a coherent solution enterprises need super-specialized personnel. Unfortunately, there is an acute shortage of talent who have this level of expertise. Enterprises are often finding that Gen AI projects are too complex to implement and too slow to deploy.

Gen AI tech stack is fragmented. Building applications is slow, complex and needs specialized talent.
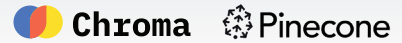
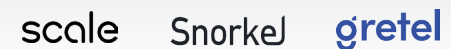**Gen AI Tech Stack (example offerings)**

Orchestration Libraries

LangChain   FIXIE

Data Mgmt

UNSTRUCTURED.IO   LlamaIndex

Vector DB

Chroma   Pinecone

Model Training/Tuning Infra

Weights & Biases   anyscale   Hugging Face

Data Labeling/Synthetic Data

scale   Snorkel   gretel

Compute & Inference

Azure   CoreWeave   Google Cloud   aws   Lambda

Foundation Models

GPT-4   LLAMA 2   Falcon   Claude   Mistral AI   PaLM 2
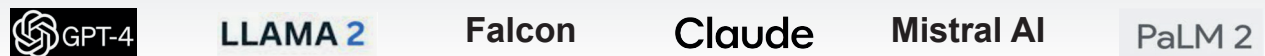
Up until advent of Gen AI, modern chat & voice bots were delivering scripted answers. This acted as a safety blanket meaning a conversational interface never went off script. This is especially useful when it doesn't understand something. Despite this being a big frustration for the end customer, it was safer for the enterprise to say "I don't know" than provide an inaccurate answer. The lack of an accurate and factual answer being provided by early Gen AI models often makes it difficult to safely deploy and extract ROI.

Potential data leakage is another major concern for enterprises. Many of the existing Gen AI tools offer no way to obfuscate sensitive data from leaking to public large language models - a risk not worth the potential ramifications.

Finally, the challenge around significant value creation is a difficult conundrum to solve with inconsistent usage. On tasks that were not historically repeated on a regular enough basis the savings are not deep enough to matter - a hour saved, here or there, a few times a week, by a few people, isn't enough for many to even allow a full roll out of Gen AI, with many organizations (and even some entire countries) saying no to Gen AI use.

Introducing EMA,
your Universal AI Employee

Gen AI has the promise to revolutionize the enterprise, but for all the reasons described above it is challenging for enterprises to realize ROI from their investments. As we took on building Ema, we asked ourselves two main questions: what makes it easy to work with human employees, and what are the things humans can't do well but AI could do better? That's how the concept of building a Universal AI Employee was born - an AI Employee who is simple to activate, accurate, and can be trusted in the enterprise.

Let's first talk about how we made Ema amazingly simple to activate. Typically building a Gen AI application requires specialized talent to do a huge amount of work on a complex toolchain. Ema introduces a **Conversational Operating System** that helps enterprises to focus on business logic rather than sift through a myriad of Gen AI libraries, frameworks, vector databases, orchestration frameworks, models, and training infrastructure. With a simple conversation, you can activate a new *Persona* who can perform unique tasks for your enterprise. These Personas collaborate seamlessly with human colleagues, and can automatically generate the workflows needed to complete tasks in a given role.
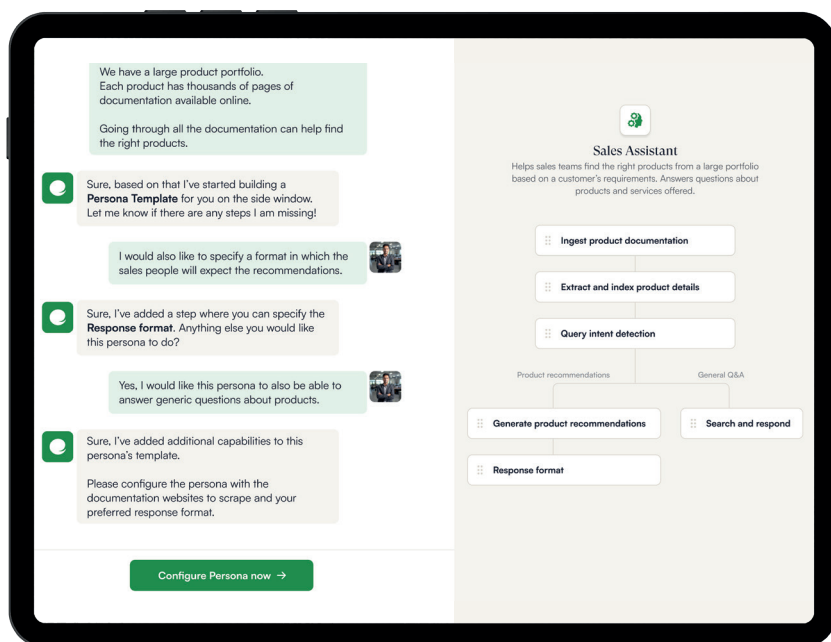
Ema's Persona design is inspired by how humans perform complex tasks in the enterprise. To perform a complex task in the enterprise, a human employee would first figure out the steps needed to do the task, and then the dependencies between the steps. Post that a human employee will plan out how to do each step optimally and then execute on all the steps. In summary, to perform a complex task, humans will generate a new workflow and execute it.

Ema's Personas are powered by our flagship **Generative Workflow Engine™** (GWE). At its core, LLMs predict the next word(s) based on the context provided. However, a complex task in the enterprise requires not just generating the next word(s), but dynamically figuring out the next steps to take. **Generative Workflow Engine™** (GWE) can generate new workflows with a conversational description of the problem by breaking the problem down into subtasks. Each subtask in the workflow could be performed by deterministic computer algorithms, custom workflows from Ema's workflow library, calling into traditional RPA (Robotic Process Automation) routines, knowledge management or another machine learning task. GWE decides and orchestrates how to optimally perform each subtask utilizing multiple enterprise applications. GWE has long-term memory that allows it to take feedback from human users and incorporate it to continuously improve the performance of the Persona(s) deployed in the enterprise.

Activate AI employees, instantly
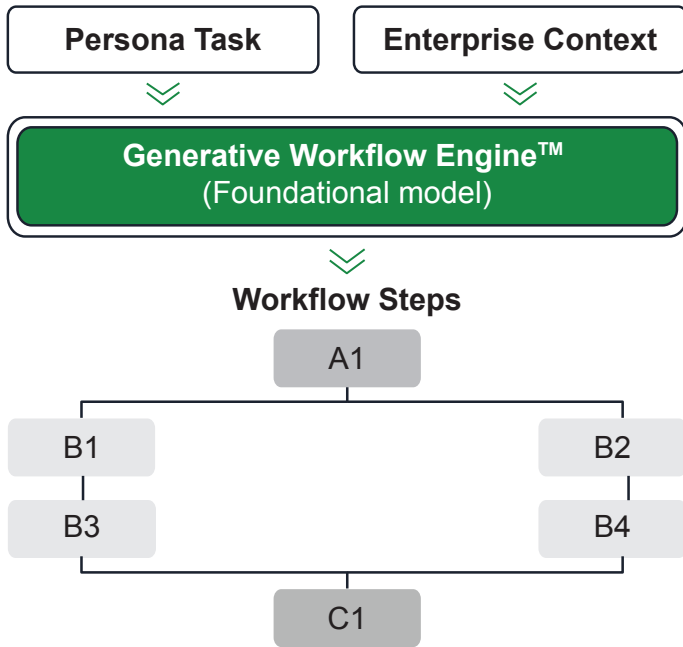
Generative Workflow Engine™

Generates new workflows with a simple conversation

Breaks down any complex task into dependent subtasks

Performs every subtask optimally

Persona Task | Enterprise Context

Generative Workflow Engine™
(Foundational model)

Workflow Steps

A1
B1
B2
B3
B4
C1

Ema has built standard personas for common roles in the enterprise such as Customer Service Specialist, Data Analyst, Sales Assistant etc. Ema's platform also enables building specialized personas for very specific and bespoke tasks in the enterprise within a few hours. Customers and partners can easily configure or conversationally generate new personas. One can think of Personas as specific roles of an employee in an Enterprise.

One employee, infinite roles

Activate AI employees for any task, specific or bespoke, simple or complex, with an effortless conversation.

Ema works for you

Seamlessly collaborating with human colleagues, making them more effective and productive

Customer Support

Data Analyst

Sales Assistant

Ema

Specialized Personas

Employee Assistant

Many more

As previously noted, current Gen AI applications suffer low accuracy caused by hallucinations and entail very high computational costs. Ema solves these problems with its proprietary **EmaFusion™** model. Ema's GWE determines which sub-tasks for a generated workflow need to be performed via LLMs, and sends each such subtask to the EmaFusion model. EmaFusion™ determines how to perform each subtask while maximizing accuracy and minimizing costs. EmaFusion is a *fusion of experts* model exceeding 2 Trillion parameters, that

utilizes most of the current foundational models - GPT 4, GPT 3.5, Claude, Gemini, Mistral, Llama2, and more. EmaFusion also allows customers to integrate bespoke models - Ema hosted private models trained on your enterprise data, domain-specific models, etc. Customers can also BYOM (Bring Your Own Model) and easily integrate with EmaFusion. It maximizes accuracy and reduces cost by intelligently blending output from multiple models that are suitable to perform a certain type of sub-task.

Increase ROI with unrivaled accuracy

Maximizes accuracy at the lowest possible cost
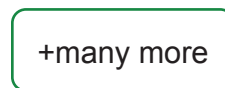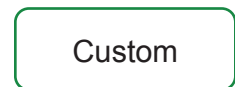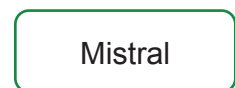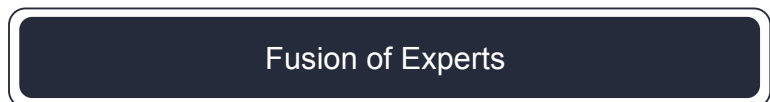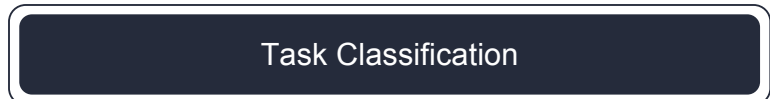
Intelligently blends the output from multiple LLMs (foundational and custom)

Future proof - Avoid lock in to any one LLM

EmaFusion™

Task Classification

Fusion of Experts

| GPT 4 | Gemini | Mistral |
| Llama2 | BYOM* | Custom |
| +many more |

*BYOM - Bring your own model

Ema

AI is often only as good as the data it has access to. A big challenge in deploying AI applications in the enterprise is giving the application access to data spread across the enterprise. For example a specific AI application may need to access data across CRM systems, cloud drives, HR systems and AWS databases. Such data integrations, ingestions, and transformations add to the complexity and delay in deploying AI applications. Ema's secure pre-integrations with hundreds of enterprise applications and data sources make installation, configuration, and deployment of Ema a straightforward process.

+ Hundreds More

We built Ema with trust at its core. Ema is compliant with leading international standards such as **SOC 2, ISO 27001, HIPAA, and GDPR**. Robust security measures, including redaction and safe de-identification of sensitive data, audit logs, real-time monitoring, robust encryption of all data at rest and in transit and explainability across all output results, guarantee that your enterprise data remains private, secure, and compliant. Ema never trains models on one customer's data to benefit other customers.

Robust data protection preventing sensitive data leakage to LLMs

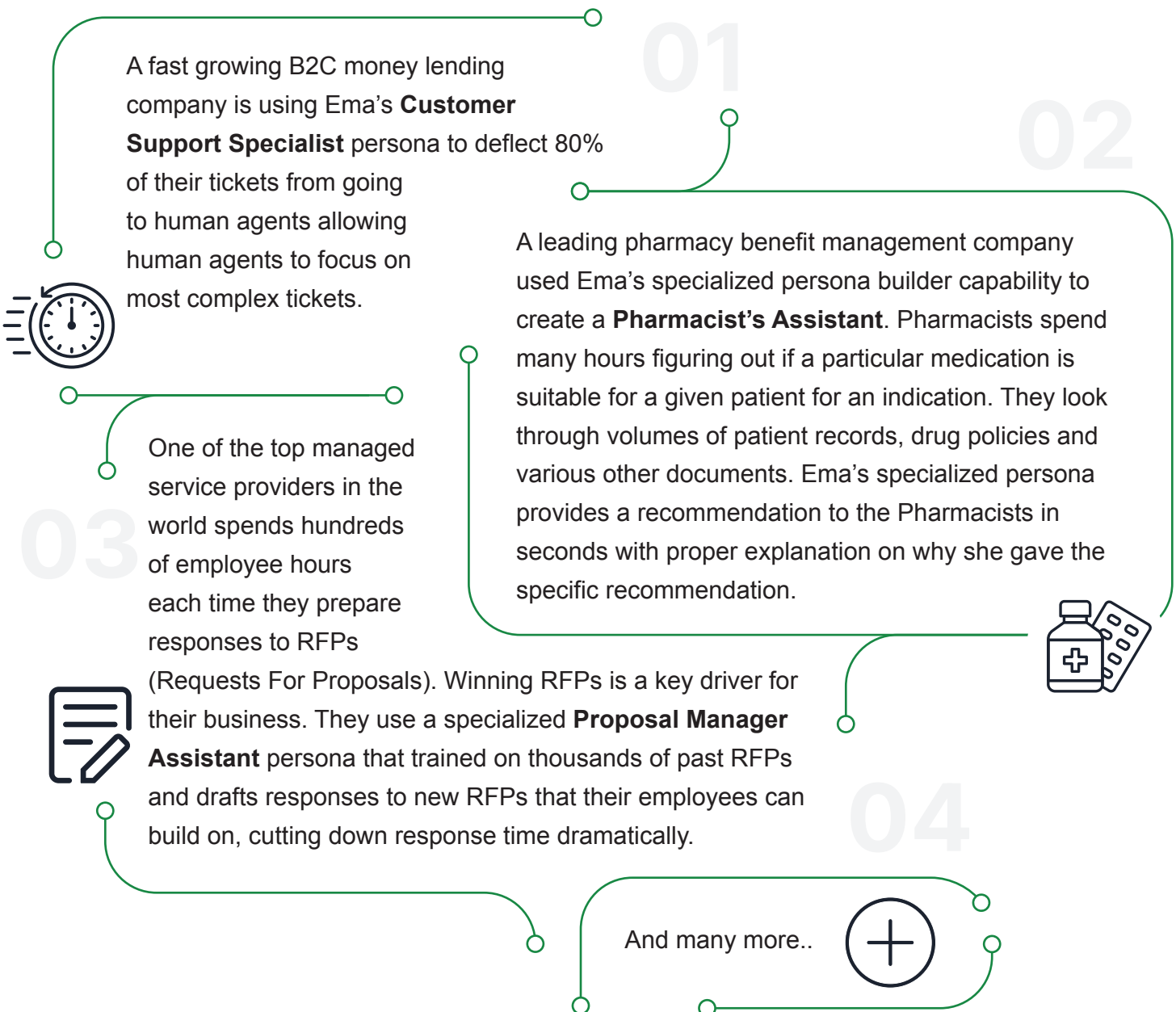**Frequent Pen Tests** keep data secure

Copyright checker ensures you are never in violation

Private data stays private, only trains models you use

Ema allows a variety of deployment options to enterprises - from multi-tenant SAAS to single-tenant and on cloud-prem options. You can choose your deployment configuration based on your enterprise needs and budget.

Customer Application Examples of Ema's Personas

The proof is in the pudding when it comes to Gen AI tools. While there is enormous promise, they are complex to implement and there are many pitfalls that make it hard to extract the right ROI. As explained earlier, Ema's Universal Employees are simple to implement, produce highly accurate results and can always be trusted with private data. Below we give some examples of how Ema's AI Employees or Personas are being used in the real world by top enterprises in the world:

01

A fast growing B2C money lending company is using Ema's **Customer Support Specialist** persona to deflect 80% of their tickets from going to human agents allowing human agents to focus on most complex tickets.

02

A leading pharmacy benefit management company used Ema's specialized persona builder capability to create a **Pharmacist's Assistant**. Pharmacists spend many hours figuring out if a particular medication is suitable for a given patient for an indication. They look through volumes of patient records, drug policies and various other documents. Ema's specialized persona provides a recommendation to the Pharmacists in seconds with proper explanation on why she gave the specific recommendation.

03

One of the top managed service providers in the world spends hundreds of employee hours each time they prepare responses to RFPs (Requests For Proposals). Winning RFPs is a key driver for their business. They use a specialized **Proposal Manager Assistant** persona that trained on thousands of past RFPs and drafts responses to new RFPs that their employees can build on, cutting down response time dramatically.

04

And many more..

Conclusion

Generative AI is the future of AI & Automation for many organizations, but not all Gen AI tools are created equal. Gen AI needs to quickly and easily showcase its value, with minimal effort from the customer side. This is especially true in a world where AI skills for many organizations are in short supply. It's highly possible that they will miss out on this next great wave of innovation, due to a lack of means to deploy these tools in the manner they are comfortable with. This is where EMA steps in, making it the Gen AI tool of choice for you. With Ema, you can deploy Gen AI tooling for a variety of enterprise use cases by quickly creating new personas conversationally. Ema's personas are simple, accurate and trusted.

If you want to differentiate yourself from others starting out on their Generative AI journey, and realize its impact quickly, then you need to hire Ema. Reach out at ema.co to learn more.