mistral.ai strategic memo

Generative AI is a transformative technology

The last year has seen a spectacular acceleration in generative AI: systems able to generate text / image conditioned on text and images. Those systems can help humans:

- produce superb creative content (text, code, graphics)
- read, process and summarise unstructured content streams thousands of times faster than humans
- interact with the world (exposed through natural or application interfaces) to execute workflows faster than ever before.

The power of generative AI was suddenly demonstrated to the general audience with the release of ChatGPT; this kind of product has been in the making by only a few small teams across the world — the few researchers of these teams are now the limiting factor to create new economic actors in the field.

Generative AI is about to boost productivity in all sectors and create a new industry (10B market size as of 2022, projected to be 110B by 2030, with an estimated growth rate of 35% per year), by seamlessly enhancing the human mind with machine capabilities. It is a transformative technology for the world economy that will change the nature of work to bring positive societal changes.

An oligopoly is shaping up

Generative AI technologies are based on years of research made in many parts of the industry and academia. The final breakthroughs, i.e. scaling training to internet-wide data and aligning models with human feedback, finally made these technologies usable by many; these breakthroughs were made by very few actors, the largest of which (OpenAI) appears to have hegemonic intention over the market.

These very few actors train generative models and hold them as assets; they serve it to thousands of third-party productivity enhancing products, in addition to also serving first-party chatbot-like products. Dozens of third-party startups are created every month to build various interfaces to these generative models.

We believe that most of the value in the emerging generative AI market will be located in the hard-to-make technology, i.e. the generative models themselves. Those models need to be trained on thousands of very powerful machines, on trillions of words coming from high quality sources, which is one factor that sets a high barrier to entry. The second important barrier lies in the difficulty to assemble an experienced team, something that mistral.ai will be in a unique position of doing.

All major actors are currently US-based, and Europe has yet to see the appearance of a serious contender. This is a major geopolitical issue given the strength (and dangers) of this new technology. mistral.ai will become a European leader in productivity and creativity enhancing AI, and guide the new industrial revolution that is coming.

Current generative AI do not meet market constraints

OpenAI and its current competitors have embraced a closed technology approach, which will dramatically reduce their market reach. In that approach, the model is kept secret and is only served through a text to text API endpoint. This raises the following important concerns for businesses:

- Businesses wishing to use generative AI technology are forced to **feed their valuable business data and sensitive user data to a black-box model,** typically deployed in the public cloud. This creates safety issues: models kept secret cannot be inspected to guarantee their outputs to be safe, thereby preventing them to be deployed in safety-critical applications. It also raises legal problems, in particular the one of falling under extraterritorial reach when sending personal data out of a company's legal territory.
- Only exposing the output of models, instead of exposing the model entirely, makes it harder to connect with other components (retrieval databases, structure inputs, images and sounds). Hundreds of products are currently built by interconnecting model outputs and inputs to create composed capacities (memory, vision, etc.). **Those products would work much better and faster were the models available as white boxes** (see for instance the Flamingo model, that combines white box vision and text models into a text+vision model)**.**
- **The data used to train** the model is kept secret, implying that we rely on a machine that has unidentified sources, and can produce uncontrollable outputs. Filtering efforts to address this issue are only a slim and breakable guarantee that the model will not output sensitive content on which it may have been trained. As of April 2023, this issue formed the basis of ChatGPT ban in Italy.

Disrupting the market from Europe

By creating mistral.ai, we intend to train state-of-the-art models with counter-positions to closed-model current offerings. **Our vision is to become a leading actor in the field, while developing a very valuable business around integrating these models in the European industry and beyond.**

mistral.ai will become a research leader in the generative AI field, eventually offering the best technology within 4 years. For this, we will first focus on several key differentiators, and then expand to a full-scale R&D effort choosing the best solutions for making new steps toward human-usable AI.

Specialising in the European market as a first step will create a defendable effort in itself — technological counter-positioning will further contribute to our appeal. Many, if not most, talents in the field of LLMs originate from Europe; as we have extensively tested, a large number of them can be convinced to join forces in our project.

Technological counter-positioning

Our early differentiators, that constitute dead angles in our competitor's strategy, are the following:

- **Take a more open approach to model development.** We will release models with a permissive open-source-software licence, that will be **largely above the competition in that category.** We will distribute tools to leverage the power of these white box models, and create a developer community around our trademark. This approach is an ideological differentiator from OpenAI; **it is a very strong point argument for hiring highly coveted top researchers, and a strong accelerator for development**, as it will open the doors for many downstream applications from motivated hackers. It will improve our reach for business development. We will balance our open-source strategy with our economic interests, keeping **the strongest and most specialised models reserved for negotiated access.**
    - We will dedicate 1% of our funding to a non-profit foundation in charge of open-source community development.
- Whether open-source or licensed, the internals (architecture and trained weights) of our models will always be accessible to our customers. This will allow **tighter integration with customer's workflows, whose content can be fed into different parts of the deep model, instead of serialising all content into input text, fed to black box APIs.**
- **Increase the focus on data sources and data control**. Our models will be trained on high quality data content (in addition to scraped content) for which we will negotiate licence agreements. This will allow us to train models much better than currently available ones (e.g. Llama). Using deeply involved technology (mixtures-of-experts and retrieval-augmented models), we will service models with optional data sources access: for a paid premium, a certain model can be specialised on finance/law/etc (this provides a substantial performance boost). With similar technologies, our models will be able to provide on-the-fly differentiated data access to employees with different views of the company intellectual property.
- **Propose unmatched guarantees on security and privacy.** Our models will be deployable on private clouds and optionally directly on device, effectively making privacy a non-concern by removing problematic flows. For this, we will direct our R&D effort towards training small but super-efficient models, effectively proposing models with the highest quality/cost ratio of the market. Our open-source strategy will also be a guarantee of audibility when deploying our models in critical industries (in particular the dual and the health sector).

Business development

On the business side, we will provide the most valuable technology brick to the emerging AI-as-a-service industry that will revolutionise business workflows with generative AI. We will co-build integrated solutions with European integrators and industry clients, and get extremely valuable feedback from this to become the main tool for all companies wanting to leverage AI in Europe.

Integration with verticals can take different marketing forms, including licensing full access to the models (including the trained weights), specialisation of models on demand, partnering with integrators/consulting companies to establish commercial contracts for fully integrated solutions. As detailed in our roadmap, we will explore and identify the best approach in parallel to technological development.

What it takes to become a leader in AI

The rarest team

The founding team is composed of lead researchers in the field, formerly employed by DeepMind and Meta, seasoned repeat French entrepreneurs and influential public leaders

- Arthur Mensch — CEO — Former staff research scientist at DeepMind
    - Lead author of several major contributions to LLM: Chinchilla, Retro, Flamingo
- Guillaume Lample — Chief scientist — Former senior staff research scientist at Meta
    - Lead Llama, major contribution of Meta to the field of large language models
- Timothée Lacroix — CTO — Former staff software engineer at Meta
    - Tech lead of Llama
- Jean-Charles Samuelian
    - CEO of Alan
- Charles Gorintin
    - CTO of Alan
- Cédric O
    - Former French Secretary of State for Digital Affairs

The first five employees that are already identified will be experienced researchers from major tech companies. They are extremely motivated by the European and open-source angle, and by the perspective of leaving companies that undergo constant reorganisations due to the speed of development of generative AI.

Infrastructure and data sources

Training a competitive model requires at least an exa-scale cluster for a few months. We intend to rent such capacity for a full year, to allow the development of both open-source and commercial models, with various capacities.

We have already negotiated competitive deals for renting computational power in Tier 1 cloud service providers (we are planning to reserve 1536 H100 starting in September, with a summer ramp up). As mistral.ai has a strong European grounding, we will also be working with both emerging European cloud providers as they grow their deep learning offers.

Having trained models at large-scale before has provided us know-hows that will allow us to gain a factor 10-100 in training efficiency compared to public methods — our founders and early employees know exactly what to do to train the strongest model for a given computational budget.

Our early investors are also content providers in Europe, and will open all necessary doors for acquiring high quality datasets on which our model can be trained and fine-tuned.

Access to major clients for usage exploration

The founding team is already organising business exploration with major French and European industrial actors. A small product-oriented team (6 people at the end of first year) will start developing leads while the technical team trains the valuable technological bricks. The model team will remain 100% focused on the hard technological brick to avoid distraction.

Business development will start in parallel to the first model family development, using the following strategy

- Focus on exploring needs with a large industrial actors, accompanied by third-party integrators that will be given full access to our best (non-open source) model
- Co-design of products with a few small emerging partners that focuses on generative-AI products.

Exploration with businesses will be used to drive the design of the second generation of models.

Roadmap

First year

We will train two generations of models, while developing business integration in parallel. The first generation will be partially open-source and rely on technology well mastered by the team. It will validate our competence near clients, investors and institutions. The second model generation will address the significant shortcomings of current models to become safely and affordably usable by businesses.

Train the best open-source standard models

At the end of 2023, we will train a family of text-generating models that can beat ChatGPT 3.5 and Bard March 2023 by a large margin, as well as all open source solutions.

Part of this family will be open-sourced; we will engage the community to build on top of it and make it the open standard.

We will service those models with the same endpoints as our competitor for a fee to acquire third-party usage data, and create a few free consumer interfaces for trademark construction and first-party usage data.

Customise for business needs and differentiate

In the following six months, these models will be accompanied by semantic embedding models for content search, and multimodal plugins for handling visual inputs. Specialised models, retrained on negotiated high quality data-sources will be prepared.

Business development will start in parallel to the first model family development: we intend to have proof-of-concept integration by the end of Q1 2024.

On the technical side, during Q1-Q2 2024, we will focus on two major aspects that have been under-estimated by incumbent companies:

- **Train a model small enough** to run on a 16GB laptop, while being a helpful AI assistant
- **Train models with hot-pluggable extra-context,** ranging in the **millions of extra words**, effectively merging language models and retriever systems.

In parallel, training and fine-tuning datasets will be constantly enriched through partnerships and data acquisition.

At the end of Q2 2024, we intend to

- be distributing the best open-source text-generative model, with text and visual output
- own generalist and specialist models with one the highest value/cost
- have scalable and usable and varied APIs for serving our models to third-party integrators
- have privileged commercial relations with one or two major industrial actors committed in using our technology

Next stages

Competing and overcoming actors like OpenAI will require major investments in later stages (GPT-4 cost a few hundred million dollars). Our purpose in the first year is to demonstrate that we are one the best teams of the world in the AI race, able to ship models and model affordances that rival the largest actors. Our experience as researchers in LLM will allow us to be much more capital-efficient in the early stage than companies discovering the field or pivoting towards it.

One of the North stars of mistral.ai will be safety: we will release models in a well-staged way, making sure that our models can only be used for purposes aligned with our values—for this, we'll offer beta access to a "red team" to uncover inappropriate behaviours and correct them.

We will thus convince major public and private institutions to trust us for constructing the safe, controllable and efficient technology that we need to make humanity benefit from this science breakthrough — effectively bringing institutions and states for the Serie A round. In that round (Q3 2024), we expect to need to raise 200M, in order to train models exceeding GPT-4 capacities.

Strong financing will allow us to train models on larger infrastructures, thereby establishing us as a research leader in AI that will be the go-to provider of the European industry.