

JAZZ COMPUTING

Simplifying AI Tech ...
Amplifying AI Opportunities

Jazz Computing Research

Unpacking Mistral.ai: Origins, Traction, and the Path Ahead

Company Profile & Market Positioning

Last Updated: Jan 23, 2024



© Copyright 2024 JazzComputing.com. All rights reserved.

This document and its contents are the property of JazzComputing.com and are protected by copyright laws. Unauthorized use, reproduction, or distribution of this document, in whole or in part, without the written consent of JazzComputing.com is strictly prohibited. If you have received this document in error, please notify info@JazzComputing.com immediately and delete this document from your system.

Please note that opinions, discussion, and insights expressed in these services and reports are provided for informational purposes only and should not be relied upon as legal, investment, or tax advice. No information contained in this report has been obtained from any sources protected by non-disclosure agreements (NDAs) or other confidential means, and is all based on publicly available information. The accuracy, reliability, and completeness of the information provided is strived for but not guaranteed. We do not accept responsibility for any errors, omissions, inaccuracies, or any actions taken based on this information. We recommend readers verify the information independently.

Table of Contents:

[Intro to Mistral.ai](#)

[Notable Milestones and Traction](#)

[Mistral.ai Technical Differentiators](#)

[Performance](#)

[Open Source](#)

[Capabilities](#)

[Team Expertise](#)

[Mistral.ai GTM and Monetization Strategy](#)

[Founding Team](#)

[Proof Points for Mistral.ai](#)

[Appendix I: About Jazz Computing:](#)

[Appendix II: References](#)

© Copyright 2024 JazzComputing.com. All rights reserved.

This document and its contents are the property of JazzComputing.com and are protected by copyright laws. Unauthorized use, reproduction, or distribution of this document, in whole or in part, without the written consent of JazzComputing.com is strictly prohibited. If you have received this document in error, please notify info@JazzComputing.com immediately and delete this document from your system.

Please note that opinions, discussion, and insights expressed in these services and reports are provided for informational purposes only and should not be relied upon as legal, investment, or tax advice. No information contained in this report has been obtained from any sources protected by non-disclosure agreements (NDAs) or other confidential means, and is all based on publically available information. The accuracy, reliability, and completeness of the information provided is strived for but not guaranteed. We do not accept responsibility for any errors, omissions, inaccuracies, or any actions taken based on this information. We recommend readers verify the information independently.

Intro to Mistral.ai

Mistral.ai is a new French artificial intelligence company started in early 2023 by researchers who used to work at Meta and Google. It is based in Paris. Mistral.ai has gotten a lot of attention for making large language models that anyone can use. Its latest model is called Mixtral 8x7B. People can access this model through an API or by downloading the open-source model and its associated weights through huggingface.

Even though Mistral.ai is very new, it has already raised a lot of money. This includes \$415 million in Series A round of fundraising. In less than a year after it was founded, Mistral.ai was valued at over \$2 billion. What makes Mistral.ai different from other AI companies is its focus on open-source software and foundational models that are available for anyone to build on. In addition, Mistral.ai uses a distinguished technical approach to improve the training and inference of the model through 1. a sparse mixture of experts (SMoE), 2. grouped-query attention (GQA) for faster inference, and 3. sliding window attention (SWA)¹. This is developed by their experienced founders who came from BigTech with deep large-scale systems skills.

Notable Milestones and Traction

Mistral.ai has achieved several notable milestones, including the following:

1. **Funding:** In December 2023, Mistral.ai raised \$415 million from investors. This shows investors think the company could be very successful. Mistral.ai had already raised \$105 million when it launched.
2. **New Model:** Mistral.ai released its latest model called Mixtral 8x7B. This model is very good at understanding and generating language. It is better than previous models and as good as bigger models like GPT-3.5 and Llama-2.

On the Chatbot front, the top models on the HuggingFace leaderboard have seen quite a reshuffle on the open-source front, with Mistral.ai's model now becoming the #2 model behind OpenAI's GPT-4, beating Google's Gemini, Meta's Llama-2 and OpenAI's GPT-3.5 among others and briefly beating Anthropic's Claude 2.1. It is also the only open-source model in the top-10 list. This is significant because it shows the competitiveness and quality of open-sourced Mistral.ai model as it is closing the gap with world best closed-source ones. [Check the latest ranking here.](#)

¹ Refer to Appendix II for more details about the three techniques and on further readings on the technical innovation.

Rank	Model	Arena Elo	95% CI	Votes	Organization	License
1	GPT-4-Turbo	1249	+14/-13	27399	OpenAI	Proprietary
2	GPT-4-0314	1191	+15/-14	17372	OpenAI	Proprietary
3	GPT-4-0613	1160	+13/-13	24888	OpenAI	Proprietary
4	Claude-1	1150	+14/-13	17333	Anthropic	Proprietary
5	Mistral Medium	1148	+14/-13	9370	Mistral	Proprietary
6	Claude-2.0	1131	+14/-13	11475	Anthropic	Proprietary
7	Mixtral-8x7b-Instruct-v0.1	1124	+15/-13	13485	Mistral	Apache 2.0
8	Gemini Pro (Dev)	1121	+15/-15	5304	Google	Proprietary
9	Claude-2.1	1119	+14/-13	22292	Anthropic	Proprietary
10	GPT-3.5-Turbo-0613	1116	+13/-13	28793	OpenAI	Proprietary

- GTM Approach:** Mistral.ai focuses on making models that anyone can build on. It makes its best models available through an API that developers can access. This shows a commitment to open-source software.
- Fast Growth:** In a very short time, Mistral.ai has become a key player in AI, especially in language models. It has gotten a lot of attention and is changing the AI industry. Its fast growth is impressive.

These achievements underscore Mistral.ai's technological advancements, strong financial backing, and commitment to open and responsible AI development.

Mistral.ai Technical Differentiators

A. Performance

Mistral.ai's language models, especially Mistral 7B, have shown remarkable performance. The Mistral 7B model beats other models like Llama 2 13B, which is twice as big. This makes it a strong competitor to models like OpenAI's GPT. The technical innovation in Mistral 7B is focused on two concepts:

- grouped-query attention (GQA) for faster inference:** it is an attention mechanism used in some neural network architectures, particularly in natural language processing models like transformer. In this technique, the attention is computed using multiple grouped query vectors instead of a single query vector. This allows the model to focus on different aspects of the input in parallel.
- sliding window attention (SWA):** Sliding window attention (SWA) is a technique used in attention-based neural networks to reduce the computational complexity of standard self-attention, by only allowing each position to attend to neighboring positions within a local window, rather than the entire sequence.

© Copyright 2024 JazzComputing.com. All rights reserved.

This document and its contents are the property of JazzComputing.com and are protected by copyright laws. Unauthorized use, reproduction, or distribution of this document, in whole or in part, without the written consent of JazzComputing.com is strictly prohibited. If you have received this document in error, please notify info@JazzComputing.com immediately and delete this document from your system.

Please note that opinions, discussion, and insights expressed in these services and reports are provided for informational purposes only and should not be relied upon as legal, investment, or tax advice. No information contained in this report has been obtained from any sources protected by non-disclosure agreements (NDAs) or other confidential means, and is all based on publicly available information. The accuracy, reliability, and completeness of the information provided is strived for but not guaranteed. We do not accept responsibility for any errors, omissions, inaccuracies, or any actions taken based on this information. We recommend readers verify the information independently.

In there follow up model, Mixtral, they used:

C. a sparse mixture of experts (SMoE): is a machine learning technique that allows combining multiple "expert" neural network models into one larger model in order to improve performance across different tasks or datasets.

B. GTM and Open Source Model

Mistral.ai openly releases its models and training processes. This is a big shift in the AI industry towards more transparency and accessibility.

C. Technical Capabilities

Experts praise Mistral.ai's models for being extremely capable, efficient, and flexible. This means they can be used for many different applications.

D. Team Expertise

Before starting Mistral.ai, the team worked at big tech companies where they gained expertise in training large AI models efficiently. They used this know-how to negotiate access to powerful computing resources even before they had funding. Their technical expertise and industry connection allowed them to get a 10-100X efficiency gain in model training.

Mistral.ai GTM and Monetization Strategy

Mistral.ai is monetizing its products through an open developer platform, where other companies can pay to use Mistral.ai's models via APIs. This approach allows for the commercial use of Mistral.ai's models, including the Mistral 7B model and the new Mixtral 8x7B model, while also facilitating the specialization of models on business use-cases, leveraging private data and usage feedback. Additionally, the company's focus on foundational models with an open technology angle indicates a commitment to making money from its models while maintaining transparency and accessibility to their models' internals.

Founding Team

The founding team of Mistral.ai consists of the following individuals:

1. Arthur Mensch: Co-Founder & CEO at Mistral.ai. He spent nearly three years at DeepMind, Google's AI laboratory.
2. Guillaume Lample: Co-Founder at Mistral.ai. He is one of the creators of the LLaMA language model unveiled by Meta, Facebook's parent company.

© Copyright 2024 JazzComputing.com. All rights reserved.

This document and its contents are the property of JazzComputing.com and are protected by copyright laws. Unauthorized use, reproduction, or distribution of this document, in whole or in part, without the written consent of JazzComputing.com is strictly prohibited. If you have received this document in error, please notify info@JazzComputing.com immediately and delete this document from your system.

Please note that opinions, discussion, and insights expressed in these services and reports are provided for informational purposes only and should not be relied upon as legal, investment, or tax advice. No information contained in this report has been obtained from any sources protected by non-disclosure agreements (NDAs) or other confidential means, and is all based on publicly available information. The accuracy, reliability, and completeness of the information provided is strived for but not guaranteed. We do not accept responsibility for any errors, omissions, inaccuracies, or any actions taken based on this information. We recommend readers verify the information independently.

3. Timothée Lacroix: Co-founder at Mistral.ai and the company's Chief Technology Officer. He was a researcher at Meta before Mistral.ai.

Proof Points for Mistral.ai

1. Ranking on HF board (on Jan 23, 2025, it was ranked as top #5 overall and top #1 in open source models).
2. Release of the model as an open-source API with an Apache 2.0 License (available for commercial use)
3. Partnership with Google Cloud to offer its model to GCP customers

© Copyright 2024 JazzComputing.com. All rights reserved.

This document and its contents are the property of JazzComputing.com and are protected by copyright laws. Unauthorized use, reproduction, or distribution of this document, in whole or in part, without the written consent of JazzComputing.com is strictly prohibited. If you have received this document in error, please notify info@JazzComputing.com immediately and delete this document from your system.

Please note that opinions, discussion, and insights expressed in these services and reports are provided for informational purposes only and should not be relied upon as legal, investment, or tax advice. No information contained in this report has been obtained from any sources protected by non-disclosure agreements (NDAs) or other confidential means, and is all based on publicly available information. The accuracy, reliability, and completeness of the information provided is strived for but not guaranteed. We do not accept responsibility for any errors, omissions, inaccuracies, or any actions taken based on this information. We recommend readers verify the information independently.

Appendix I: About Jazz Computing:

JazzComputing LLC is a boutique AI Advisory firm, based in Silicon Valley, founded by Dr. Lamia Youseff. We specialize in AI Investment & AI Ecosystem Advisory Services for investors. We deeply understand and leverage the distinct strengths of various entities within the AI ecosystem; from startups and research groups to established cloud AI companies and other players in the AI value chain.

We help investors make informed and strategic decisions in this fast-evolving ecosystem.

Our mission is to empower investors to identify early and evaluate properly the tech opportunities disrupting multi-\$B industries.

About Dr. Lamia Youseff:

Dr. Lamia Youseff, Ph.D., MsM, has been working in Cloud & AI / ML for ~25 years, first in academia (MIT, Stanford, UCSB), then led AI and Cloud organizations at large tech (Google, Microsoft, Facebook and Apple). She is a founding engineer of Google Cloud (a \$26B+ revenue-generating business for Google). She directly operated and managed another multi-billion-dollar business for Apple. She has 4+ patents, 20+ academic papers in top conferences and 5000+ non-self citations.

Today, she is a tech advisor for startups in Generative AI and co-leads the Stanford Generative AI meetup. She is also currently an independent consultant and the executive director at JazzComputing.com, a visiting research scientist at Stanford University in Computer Science and AI, and a research affiliate with MIT Computer Science and Artificial Intelligence Lab (CSAIL). She regularly teaches classes on business management, AI strategy, and leadership at Stanford Graduate School of Business (GSB), where she earned her Master's in Management.

© Copyright 2024 JazzComputing.com. All rights reserved.

This document and its contents are the property of JazzComputing.com and are protected by copyright laws. Unauthorized use, reproduction, or distribution of this document, in whole or in part, without the written consent of JazzComputing.com is strictly prohibited. If you have received this document in error, please notify info@JazzComputing.com immediately and delete this document from your system.

Please note that opinions, discussion, and insights expressed in these services and reports are provided for informational purposes only and should not be relied upon as legal, investment, or tax advice. No information contained in this report has been obtained from any sources protected by non-disclosure agreements (NDAs) or other confidential means, and is all based on publicly available information. The accuracy, reliability, and completeness of the information provided is strived for but not guaranteed. We do not accept responsibility for any errors, omissions, inaccuracies, or any actions taken based on this information. We recommend readers verify the information independently.

Appendix II: References

1. Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, William El Sayed

Mixtral of Experts

<https://arxiv.org/abs/2401.04088>

This paper introduces a new concept on a **sparse mixture of experts (SMoE)**, which is a machine learning technique that allows combining multiple "expert" neural network models into one larger model in order to improve performance across different tasks or datasets.

2. Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, William El Sayed

Mistral 7B

<https://arxiv.org/abs/2310.06825>

This paper introduces two new techniques to simplify transformer architecture and make it more efficient. They are:

- a. grouped-query attention (GQA)** for faster inference: it is an attention mechanism used in some neural network architectures, particularly in natural language processing models like transformer. In this technique, the attention is computed using multiple grouped query vectors instead of a single query vector. This allows the model to focus on different aspects of the input in parallel.
- b. sliding window attention (SWA)**: Sliding window attention (SWA) is a technique used in attention-based neural networks to reduce the computational complexity of standard self-attention, by only allowing each position to attend to neighboring positions within a local window, rather than the entire sequence.

3. **Mistral.ai Strategic Memo** (dated 2023/06)

<https://promorphcapital.com/wp-content/uploads/2023/06/mistral.ai-strategic-memo.pdf>

   Copyright 2024 JazzComputing.com. All rights reserved.

This document and its contents are the property of JazzComputing.com and are protected by copyright laws. Unauthorized use, reproduction, or distribution of this document, in whole or in part, without the written consent of JazzComputing.com is strictly prohibited. If you have received this document in error, please notify info@JazzComputing.com immediately and delete this document from your system.

Please note that opinions, discussion, and insights expressed in these services and reports are provided for informational purposes only and should not be relied upon as legal, investment, or tax advice. No information contained in this report has been obtained from any sources protected by non-disclosure agreements (NDAs) or other confidential means, and is all based on publicly available information. The accuracy, reliability, and completeness of the information provided is strived for but not guaranteed. We do not accept responsibility for any errors, omissions, inaccuracies, or any actions taken based on this information. We recommend readers verify the information independently.