



# From Concept to Reality: Mastering LLMs from POC to Production



# About us

- Datasaur offers:
  - Powerful NLP Labeling product
  - Private LLM development platform allowing clients to compare and build with 200+ AI Models
- Second-time founder, built AI systems at Apple and Yahoo
- StartX F19, YC W20

## Backed by



# Trusted by customers

We support hundreds of companies and universities globally across healthcare, legal, fintech, eCommerce, and more.



# Agenda



- **2024 Trends**

- Pilot/POC → Production
- Importance of a healthy, competitive ecosystem
- Calculating ROI

- **Production Techniques**

- Don't rely on a single model
- What you can do - LLM distillation
- Future-proofing deployments as next-gen models
  - Ground Truth Data
  - Regression Testing

# 2024 Trends

A large blue arrow pointing right, set against a black background on the left and a light blue gradient on the right.

# A Shift is Happening...



- **2023-mid 2024**
  - Year of Pilots / POCs
  - Does this technology work? For us?
  - What are the best use cases?
  - Executive reports, board meetings
- **2024+**
  - What are we going to implement?
  - What's the impact on the bottom line?
  - When will we see the ROI?

# What's the Best Model™?

May: OpenAI GPT 4o

# What's the Best Model™?

~~May: OpenAI GPT-4o~~

**Jun: Claude 3.5 Sonnet**



# What's the Best Model™?

~~May: OpenAI GPT-4o~~

~~Jun: Claude 3.5 Sonnet~~

**Jul: Llama 3.1**

# What's the Best Model™?

~~May: OpenAI GPT-4o~~

~~Jun: Claude 3.5 Sonnet~~

~~Jul: Llama 3.1~~

**Aug: Gemini 1.5 Pro**

# Explosion of Model Options

| Rank* (UB) | Model  | Arena Score | 95% CI | Votes  | Organization | License             | Knowledge Cutoff |
|------------|--|-------------|--------|--------|--------------|---------------------|------------------|
| 1          | <a href="#">ChatGPT-4o-latest (2024-08-08)</a>   | 1314        | +6/-5  | 11555  | OpenAI       | Proprietary         | 2023/10          |
| 2          | <a href="#">Gemini-1.5-Pro-Exp-0801</a>          | 1297        | +4/-4  | 20674  | Google       | Proprietary         | 2023/11          |
| 3          | <a href="#">GPT-4o-2024-05-13</a>                | 1286        | +2/-3  | 78496  | OpenAI       | Proprietary         | 2023/10          |
| 4          | <a href="#">GPT-4o-mini-2024-07-18</a>           | 1274        | +5/-3  | 20089  | OpenAI       | Proprietary         | 2023/10          |
| 4          | <a href="#">Claude 3.5 Sonnet</a>                | 1271        | +3/-3  | 48546  | Anthropic    | Proprietary         | 2024/4           |
| 4          | <a href="#">Gemini Advanced App (2024-05-14)</a> | 1266        | +4/-3  | 52249  | Google       | Proprietary         | Online           |
| 5          | <a href="#">Meta-Llama-3.1-405b-Instruct</a>     | 1263        | +5/-4  | 19909  | Meta         | Llama 3.1 Community | 2023/12          |
| 7          | <a href="#">Gemini-1.5-Pro-001</a>               | 1260        | +3/-3  | 70339  | Google       | Proprietary         | 2023/11          |
| 7          | <a href="#">Gemini-1.5-Pro-Preview-0409</a>      | 1257        | +3/-2  | 55650  | Google       | Proprietary         | 2023/11          |
| 7          | <a href="#">GPT-4-Turbo-2024-04-09</a>           | 1257        | +3/-3  | 85076  | OpenAI       | Proprietary         | 2023/12          |
| 11         | <a href="#">GPT-4-1106-preview</a>               | 1251        | +3/-3  | 92780  | OpenAI       | Proprietary         | 2023/4           |
| 11         | <a href="#">Mistral-Large-2407</a>               | 1249        | +4/-5  | 12394  | Mistral      | Mistral Research    | 2024/7           |
| 11         | <a href="#">Claude 3 Opus</a>                    | 1248        | +2/-3  | 156550 | Anthropic    | Proprietary         | 2023/8           |
| 11         | <a href="#">Athene-70b</a>                       | 1247        | +6/-4  | 12128  | NexusFlow    | CC-BY-NC-4.0        | 2024/7           |
| 11         | <a href="#">Meta-Llama-3.1-70b-Instruct</a>      | 1246        | +5/-4  | 14622  | Meta         | Llama 3.1 Community | 2023/12          |

200+ foundation models and counting

# Importance of a healthy, competitive ecosystem

"GPT-4o mini scores 82% on MMLU and currently outperforms GPT-4 on chat preferences in LMSYS leaderboard. It is priced at 15 cents per million input tokens and 60 cents per million output tokens, an order of magnitude more affordable than previous frontier models and more than 60% cheaper than GPT-3.5 Turbo."



- OpenAI blog, July 18, 2024

"Today, we're announcing a series of improvements across AI Studio and the Gemini API:

Significant reduction in costs for Gemini 1.5 Flash, with input token costs decreasing by 78% and output token costs decreasing by 71%."



- Google Gemini Blog, Aug 8, 2024

# How do we measure ROI for an LLM?



- **What's the cost savings/revenue generation potential?**
  - Ex: Doctor saves 30 minutes a day writing up reports
  - Ex: Rapid new pharmaceutical drug discovery
- **What's the cost of implementing this solution?**
  - Build/buy technology
    - *Unit costs*
    - *Difficulty of hiring GenAI experts*
  - Maintenance/operation
    - *Hallucinations*
  - Onboarding/training

# Looking Beyond Cost/Accuracy



## What are the other parameters to the problem?

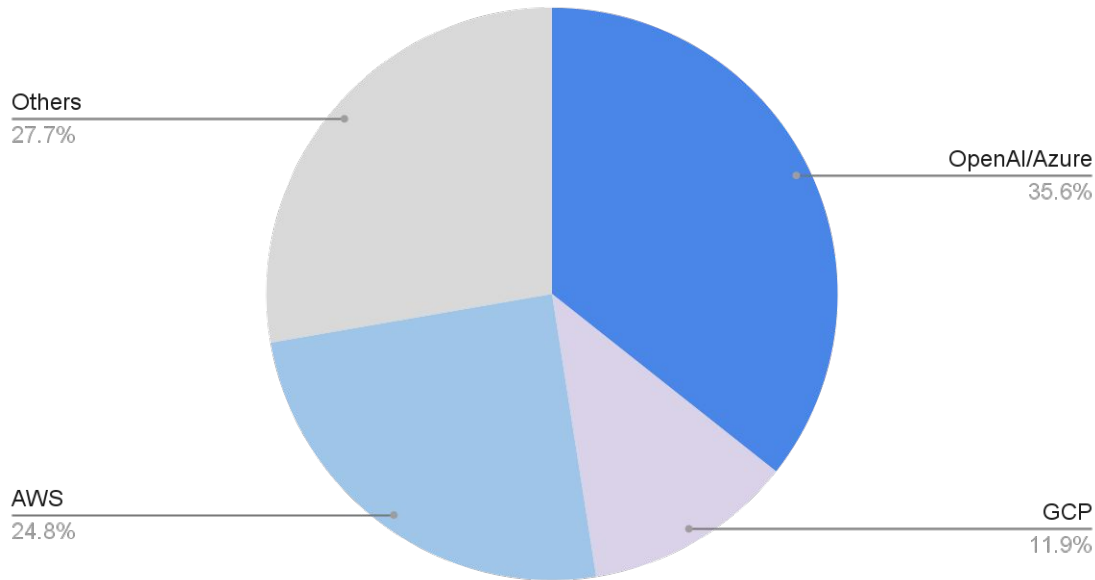
- Timeline (build vs. buy)
- Security
- Preferred cloud
- Data integrations
- SLAs (How long can an answer take?)

# Production Techniques

A large blue arrow pointing right, set against a black background on the left and a light blue gradient on the right.

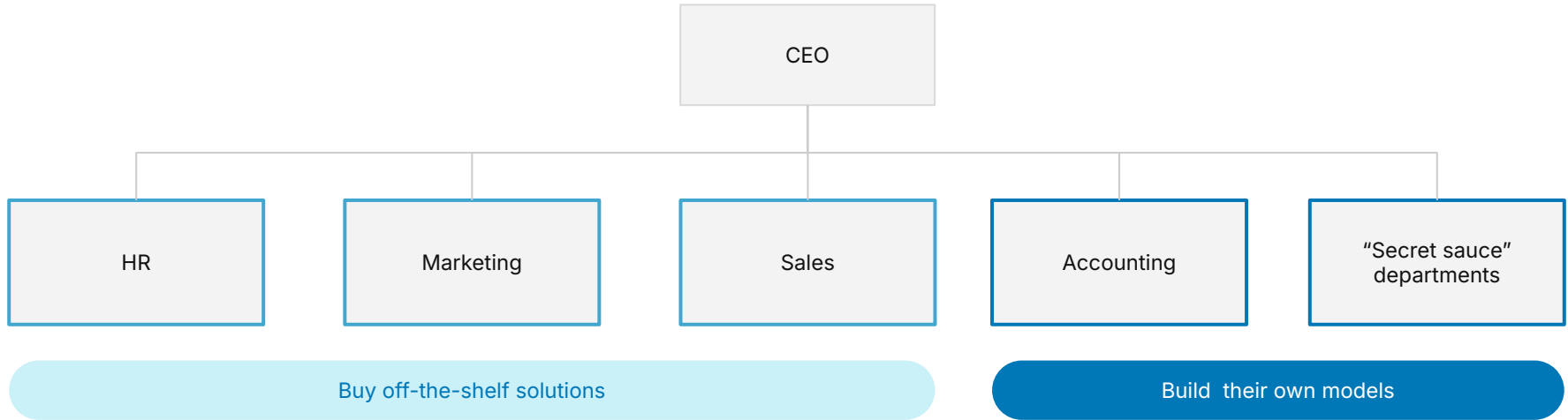
# Build many models - less single failure endpoint, easier to debug and fix 1-2 use cases at a time

GenAI served (hypothetical)





# Each organization should weigh their own pros and cons and arrive at independent decisions



**(hypothetical!)**

# LLM distillation - Lowering your Costs to Improve ROI

Dataset for Biomedical Research Question Answering

| pubid      | question   | context  | long_answer  | final_decision |
|------------|--|--|--|----------------|
| int32      | string   | sequence   | string   | string         |
| 25,429,730 | Are group 2 innate lymphoid cells ( ILC2s ) increased in chronic rhinosinusitis with nasal polyps or eosinophilia? | { "contexts": [ "Chronic rhinosinusitis (CRS) is a heterogeneous disease with an uncertain pathogenesis. Group 2 innate lymphoid cells (ILC2s) represent a recently discovered cell population which has been implicated in driving Th2 inflammation in CRS; however, their relationship with clinical disease characteristics has yet to be investigated.", "The aim of this study was to identify ILC2s in sinus mucosa in patients with | As ILC2s are elevated in patients with CRSwNP, they may drive nasal polyp formation in CRS. ILC2s are also linked with high tissue and blood eosinophilia and have a potential role in the activation and survival of eosinophils during the Th2 immune response. The association of innate lymphoid cells in CRS provides insights into its pathogenesis. | yes            |

Final decision:  
- yes  
- no

For LLM Distillation purposes, the dataset output is generated by **Llama 3.1 405B**.

# LLM distillation - Lowering your Costs to Improve ROI



| Model                         | Precision | Recall | F1-Score | Accuracy     |
|-------------------------------|-----------|--------|----------|--------------|
| Llama 3.1 405B                | 79.1%     | 84.5%  | 81.5%    | <b>76.4%</b> |
| Llama 3 8B before fine-tuning | 68.7%     | 78.4%  | 71.1%    | <b>67.3%</b> |
| Llama 3 8B after fine-tuning  | 80.7%     | 89.0%  | 84.6%    | <b>80.2%</b> |

# LLM distillation - Lowering your Costs to Improve ROI



|                          | <b>Llama 3.1 405B</b><br>(Amazon Bedrock) | <b>Llama 3 8B - (before fine-tuning)</b><br>(Amazon Bedrock) | <b>Fine-tuned Llama 3 8B</b><br>(Modal) |
|--------------------------|---|--|---|
| <b>Total time to run</b> | ~13.88 hours                              | ~ 9.47 hours   | ~ 7 hours                               |
| <b>Total cost</b>        | ~ \$76                                    | ~ \$5.8  | ~ \$24                                  |

# Results



## Classification performance improved

We achieved higher precision (+11%), recall (+6%), F1-score (+10%), and accuracy (+9.9%) with the Llama 3 8B model after fine-tuning it with 1,000 samples from Llama 3.1 405B.

## Reduce cost by 3x

The estimated cost shows that we can **reduce the cost of Llama 3.1 405B by 3x** by distilling its knowledge into a smaller model like Llama 3 8B, while still achieving comparable performance.

## Inference speed is 1.5x faster

The inference speed of Llama 3 8B is **1.5x faster** compared to Llama 3.1 405B. As a result, this model can be more practical for applying to specific tasks in real-world problems.

# Future proofing against next-generation models



- Most models are good enough
- Capture 60-70% improvements at this time
- When new technologies emerge, decide if it's worth revisiting and optimizing another 10-20%

# Model comparison

**Sandbox > Official Datasaur Demo** | 57 models running | Datasaur Datasaur Playground

**Applications** | Deployment

**Application** | 4 applications selected | + Add application

- Llama 3.1 AWS Bedrock | 1.829s | ⚙️ ⋮
- GPT 4o OpenAI | 2.165s | ⚙️ ⋮
- Claude 3.5 AWS Bedrock | 2.630s | ⚙️ ⋮
- Google Gemini 1.5 Flash | 0.964s | ⚙️ ⋮

**Prompt** | 1 prompt selected | + Add prompt | Run all (4 prompt runs) | Total predicted cost: \$0.018 Free

**Prompt 1** | 🗑️

What are the advantages of open-source vs. proprietary LLMs?

**Completion results** | ⏴ ⏵

- Llama 3.1 AWS Bedrock** | 1.829s | 💰 \$0.00008 Free (129 tokens) | ⋮  
The advantages of open-source LLMs include transparency, community-driven development, and flexibility, as well as the ability to modify and customize the model to suit specific needs. Open-source LLMs also often have a lower cost of entry and can be more accessible to researchers and developers who may not have the resources to access proprietary models. Additionally, open-source LLMs can foster innovation and collaboration, as developers can build upon and contribute to existing models.
- GPT 4o OpenAI** | 2.165s | 💰 \$0.002 Free (132 tokens) | ⋮  
Open-source large language models (LLMs) offer transparency, cost savings, and the flexibility to customize the model for specific needs, fostering innovation and collaboration within the community. Proprietary LLMs, on the other hand, often come with optimized performance, dedicated support, and better security guarantees, as they are usually developed and maintained by well-resourced organizations. The choice between them depends on the specific requirements such as budget, expertise, and the need for control versus out-of-the-box reliability.
- Claude 3.5 AWS Bedrock** | 2.630s | 💰 \$0.002 Free (106 tokens) | ⋮  
Open-source LLMs offer transparency, allowing users to inspect and modify the code, which can lead to faster improvements and customization. They often have lower costs and fewer usage restrictions compared to proprietary models. However, proprietary LLMs may have access to larger datasets and more resources for development, potentially resulting in better performance in some areas.

# Establish Ground Truth Datasets

The screenshot displays the Datasaur web interface for labeling a dataset. At the top, the breadcrumb navigation shows 'Demo project > demo.csv'. The main interface is divided into three primary sections:

- Prompt:** Contains the question 'Who is the first president of USA?' and a 'Prompt template' button.
- Completion:** Shows the model's response: 'The first president of the United States was George Washington. He served from 1789 to 1797.' Below the text is a five-star rating system with all stars currently unselected.
- Labeling Guidelines:** A sidebar on the right provides instructions for rating the model's output based on five star levels:
  - 5 stars: Excellent:** Choose this rating if the answer is highly relevant, accurate, and well-articulated.
  - 4 stars: Good:** Use this rating if the answer is mostly relevant and provides accurate information.
  - 3 stars: Average:** Assign this rating if the answer is somewhat relevant and contains some correct information. The answer is adequate but may lack depth or precision.
  - 2 stars: Poor:** Select this rating if the answer has significant errors or is partially relevant to the prompt. The answer demonstrates some understanding but lacks clarity and accuracy.
  - 1 star: Very Poor:** Choose this rating if the answer is completely incorrect or irrelevant to the prompt. The answer shows a lack of

At the bottom of the main panel, a 'Submit' button is visible. The interface also includes a progress indicator at the top center showing 'Labeling in progress' and a dropdown menu for the user 'D Datasaur' in 'Labeler Mode'. Navigation controls for the current item (3 / 20) and the overall dataset (1 / 100) are present.



# Regression testing

- Protect against model drift
- Ensure that everything that works in production continues working

▼ Summary

Average evaluator scores

|   | Evaluator   | Score |
|---|---|-------|
| 1 | Answer Correctness   LangChain<br>Evaluator model: gpt-4o-mini-2024-07-18 | 8.03  |

[Learn how the scores are calculated](#)

|   | Prompt   | Expected completion  | Completion   | Score |
|---|--|--|--|-------|
| 1 | question: Do mitochondria play a role in remodelling lace plant leaves during programmed cell death? \ncontext: {\ncontexts\": array([\nProgrammed cell death (PCD) is the regulated death of cells within an organism. The lace plant (Aponogeton madagascariensis) produces perforations in its leaves through PCD. The leaves of the plant consist of a latticework of longitudinal and transverse veins enclosing areoles. PCD occurs in the cells at the center of these areoles and progresses outwards, stopping <td>Results depicted mitochondrial dynamics in vivo as PCD progresses within the lace plant, and highlight the correlation of this organelle with other organelles during developmental PCD. To the best of our knowledge, this is the first report of mitochondria and chloroplasts moving on transvacuolar strands to form a ring structure surrounding the nucleus during developmental PCD. Also, for the first time, we have shown the feasibility for the use of CsA in a whole plant system. Overall, our</td> <td>The role of mitochondria during programmed cell death (PCD) has been recognized in animals, and this study aims to elucidate their role during developmentally regulated PCD in vivo in A. madagascariensis. The results show that mitochondrial dynamics change during PCD, and the treatment with cyclosporine A, which inhibits mitochondrial permeability transition pore formation, reduces the number of perforations in lace plant leaves. \n -</td> <td>8</td> | Results depicted mitochondrial dynamics in vivo as PCD progresses within the lace plant, and highlight the correlation of this organelle with other organelles during developmental PCD. To the best of our knowledge, this is the first report of mitochondria and chloroplasts moving on transvacuolar strands to form a ring structure surrounding the nucleus during developmental PCD. Also, for the first time, we have shown the feasibility for the use of CsA in a whole plant system. Overall, our | The role of mitochondria during programmed cell death (PCD) has been recognized in animals, and this study aims to elucidate their role during developmentally regulated PCD in vivo in A. madagascariensis. The results show that mitochondrial dynamics change during PCD, and the treatment with cyclosporine A, which inhibits mitochondrial permeability transition pore formation, reduces the number of perforations in lace plant leaves. \n - | 8     |

# Key Takeaways

## What's the Best Model™?

No single best model for all projects and workloads

## How do you calculate return on investment (ROI)?

- Evaluate your own parameters and ROIs
- Test multiple models to find the right fit
- Match the right model to the right project based on quality, speed, and cost


## Techniques at your disposal

- Don't rely on a single model
- LLM Distillation
- Establish Ground Truth Data
- Regression Testing



Thank you

Questions?

 [ivan@datasaur.ai](mailto:ivan@datasaur.ai)

 [linkedin.com/in/iylee/](https://www.linkedin.com/in/iylee/)