AI and Fintech frica 2025

https://www.techinafrica.com/crypto-fintech-agritech-ai-africas-most-promising-sectors-in-2025/

Lami Insurance Kenya

https://techcrunch.com/2022/08/02/kenyan-insurtech-lami-raises-3-7m-seed-extension-led-by-harlem-capital/

Untitled

Mckinsey & Co Insurance Report 2020

https://www.mckinsey.com/featured-insights/middle-east-and-africa/africas-insurance-market-is-set-for-takeoff

Curacel Nigeria Insuretech Embedded Insurance Raise 2023

https://techcrunch.com/2023/02/14/nigerias-curacel-raises-funding-to-power-insurance-offerings-and-expand-into-north-africa/

Poor Insurance Penetration in Nigeria

https://minetinsurancebrokersltd.com/poor-insurance-penetration-in-nigeria/

https://fsdafrica.org/for-insurance-industry-poor-awareness-slows-growth/

Nigerian Insurance Policy Mandate 2025

https://pavestoneslegal.com/regulatory-update-the-nigerian-insurance-industry-reform-act-2025-navigating-compliance/

https://pavestoneslegal.com/regulatory-update-2025-guidelines-for-insurtech-operations-in-nigeria-navigating-compliance/

https://www.reuters.com/world/africa/nigeria-enacts-sweeping-reforms-insurance-sector-2025-08-06/

The Act's focus on digitisation will be key to bridging access gaps, fostering the adoption of Insurtech solutions that streamline claims processing, combat fraud, and improve customer experience
https://punchng.com/niira-2025-new-dawn-for-nigerias-insurance-sector/

5 Biggest Challenges of Insurance in Nigeria

https://www.curacel.co/post/5-biggest-challenges-of-insurance-companies-in-nigeria

🔗 https://www.youtube.com/watch?v=Q-xQZyZtKbg

https://www.youtube.com/watch?v=_AVeb9hZcfE

Why Nigerian Business Owners Avoid Insurance

https://blueprint.ng/why-nigerian-business-owners-avoid-insurance-cover/

AIICO Insurance Case Study

https://nairaproject.com/projects/6517-an-analysis-of-insurance-penetration-in-nigeria-a-case-study-of-aiico-insurance.html

Fitch 2023 Nigerian Insurance Industry Report

https://your.fitch.group/rs/732-CKH-767/images/industry-profile-and-operating-environment-nigerian-insurance_10242430.pdf?utm_source=chatgpt.com

Fitch assesses the Nigerian insurance market as very competitive. We believe no single insurer or group of insurers have significant competitive advantage over the insurance market and offer similar, easily substitutable products. Insurance market concentration is limited; the top three life and non-life insurers controlled about 43% and 25% of the market, respectively, according to NAICOM data for 1Q23. Price undercuts have proven ineffective to increase company market shares and to increase overall insurance penetration. Fitch believes digitalisation efforts, knowledge transfer by foreign entrants from more developed insurance markets and confidence building to be the key drivers of competitive differentiation. The Nigerian insurance market consists of 57 registered insurance, and two registered reinsurance companies. Of this, 13 are life insurers.

Nigeria Insurance Sector Transformation

https://www.ainvest.com/news/nigeria-insurance-sector-transformation-high-yield-opportunity-post-niira-2025-2508/

2025 Nigeria Insurance Industry Report

https://www.agustoresearch.com/report/2025-insurance-industry-report/

8 AI Sales Assistants 2025

https://zapier.com/blog/ai-sales-assistant/

AI is Transforming Go-To-Market Strategies

https://www.forbes.com/councils/forbesbusinessdevelopmentcouncil/2024/08/09/how-ai-is-transforming-go-to-market-strategies/

General Catalyst Hemant Taneja al Investors Navigating Peak Ambiguity

https://www.ft.com/content/2757870b-dbbf-4603-9fc3-7ccd03ffc8da?utm_source=newsletter.strictlyvc.com&utm_medium=newsletter&utm_campaign=anthropic-issues-warning-about-vibe-coding&_bhlid=8c3a9fc1cd6dda2cdd013a14193197e44d9acb62

Vox AI for Fastfood

https://finance.yahoo.com/news/vox-ai-raises-8-7m-124600406.html?utm_source=newsletter.strictlyvc.com&utm_medium=newsletter&utm_campaign=anthropic-issues-warning-about-vibe-coding&_bhlid=3f7fc2f96a37dad5b98976e95dee45482d055b23&guccounter=1

Sola InsureTech Raises $8 Million

https://www.prnewswire.com/news-releases/sola-closes-8m-series-a-to-build-the-first-vertically-integrated-insurance-company-302537943.html?utm_source=newsletter.strictlyvc.com&utm_medium=newsletter&utm_campaign=anthropic-issues-warning-about-vibe-coding&_bhlid=f73ff639ee730a40e4d8043887b3a82c16c5dce1

Nauta Raises $7 Millions to Expand Global Import Logistics

https://www.pymnts.com/news/investment-tracker/2025/nauta-raises-7-million-to-expand-ai-powered-logistics-orchestration-platform/

AI Agent in Distributed Electic Grids

https://www.prnewswire.com/news-releases/aigent-raises-6m-in-funding-to-couple-ai-with-distributed-generation-to-deliver-grid-reliability-at-scale-302538264.html?utm_source=newsletter.strictlyvc.com&utm_medium=newsletter&utm_campaign=anthropic-issues-warning-about-vibe-coding&_bhlid=43950e4a57fcac60f4ab3da63de1f1fa14ef37e6

Researchers Are Already aLeaving Meta's New Superintelligence Lab

https://www.wired.com/story/researchers-leave-meta-superintelligence-labs-opena

Mark Cubans Disruption Formula

https://techcrunch.com/podcast/from-streaming-to-healthcare-to-ai-mark-cuban-reveals-his-disruption-formula/?utm_source=newsletter.strictlyvc.com&utm_medium=newsletter&utm_campaign=anthropic-issues-warning-about-vibe-coding&_bhlid=6663fd092019f4df83b3a5536cbaabafc96ff2fc

OnePipe Embedded Finance in Nigeria via Tech Crunch

https://techcrunch.com/2021/11/23/nigerias-onepipe-raises-3-5m-to-double-down-on-its-embedded-finance-offering/

https://growtrade.io/distributors/

https://docs.google.com/forms/d/e/1FAIpQLScN-0nct8E6Xdkmy5PmmqvSAjB4i3YRjn...

Maplerad BaaS Peter Theil Backed Nigerian Fintech

https://techcrunch.com/2022/10/17/nigerian-banking-as-a-service-platform-maplerad-raises-6m-led-by-peter-theils-valar-ventures/

Varo BaaS Tech Crunch

https://techcrunch.com/2022/09/16/fintech-varo-digital-bank/

BERT - To Do

https://arxiv.org/pdf/1810.04805

OpenAI Startup Growth Playbook in Asia

https://e27.co/from-bangkok-to-billions-inside-openais-startup-growth-playbook-20250824/

Finetuning LLMs

https://towardsdatascience.com/fine-tuning-large-language-models-llms-23473d763b91/

Multimodal Embeddings

https://towardsdatascience.com/multimodal-embeddings-an-introduction-5dc36975966f/

AI research is traditionally split into distinct fields: NLP, computer vision (CV), robotics, human-computer interface (HCI), etc. However, countless practical tasks require the **integration of these different research areas** e.g. autonomous vehicles (CV + robotics), AI agents (NLP + CV + HCI), personalized learning (NLP + HCI), etc.

Embeddings

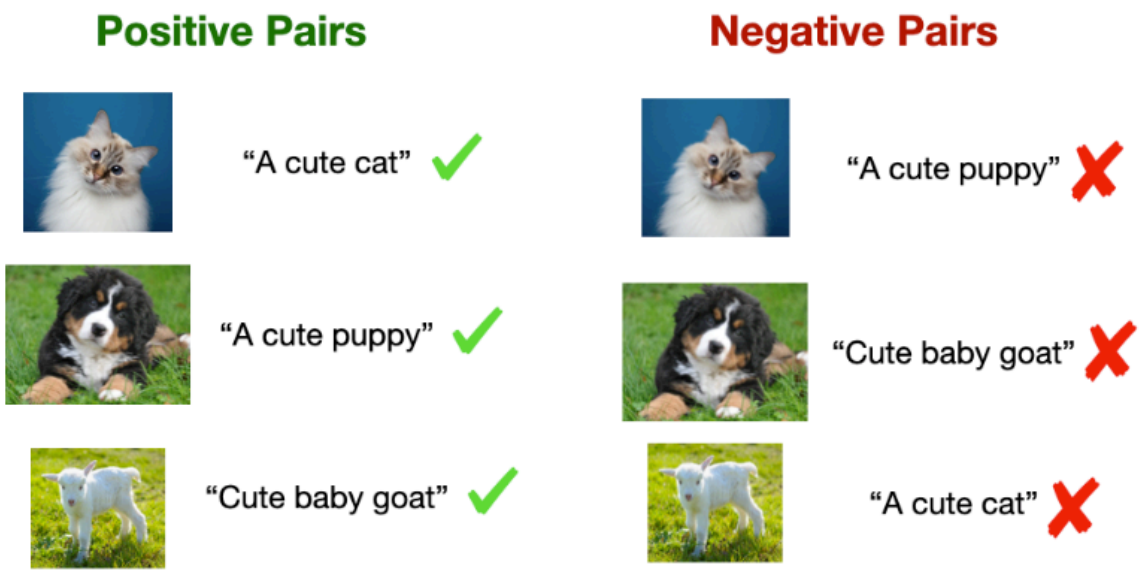
Embeddings are **(useful) numerical representations of data learned implicitly through model training**. For example, through learning how to predict text, BERT learned representations of text, which are helpful for many NLP tasks [1]. Another example is the Vision Transformer (ViT), trained for image classification on Image Net, which can be repurposed for other applications [2].

Multimodal Embeddings

Although text and images may look very different to us, in a neural network, these are **represented via the same mathematical object**, i.e., a vector. Therefore, in principle, text, images, or any other data modality can processed by a single model.
This fact underlies **multimodal embeddings**, which **represent multiple data modalities in the same vector space** such that similar concepts are co-located (independent of their original representations).

Contrastive Learning

The standard approach to aligning disparate embedding spaces is **contrastive learning (CL)**. A key intuition of CL is to **represent different views of the same *information* similarly** [5].
This consists of learning representations that **maximize the similarity between positive pairs** and **minimize the similarity of negative pairs**. In the case of an image-text model, a positive pair might be an image with an appropriate caption, while a negative pair would be an image with an irrelevant caption (as shown below).



Two key aspects of CL contribute to its effectiveness

1. Since positive and negative pairs can be curated from the data’s inherent structure (e.g., metadata from web images), CL training data **do not require manual labeling**, which unlocks larger-scale training and more powerful representations [3].
2. It simultaneously maximizes positive and minimizes negative pair similarity via a special loss function, as demonstrated by CLIP [3].

Contrastive Loss (CLIP)

Image term + Text term

$$L = -\frac{1}{2n}\left(\sum_{i=1}^{n}\log\frac{\exp(\text{logits}_{i,i})}{\sum_{j=1}^{n}\exp(\text{logits}_{i,j})} + \sum_{j=1}^{n}\log\frac{\exp(\text{logits}_{j,j})}{\sum_{i=1}^{n}\exp(\text{logits}_{i,j})}\right)$$

$\text{logits}_{i,j}$ = temperature-scaled similarity score between i^{th} image and j^{th} text

n = total number of positive pairs

*Note: logits matrix is generally **asymmetric***

Use Case

Use case 1: 0-shot Image Classification

The basic idea behind using CLIP for 0-shot image classification is to pass an image into the model along with a set of possible class labels. Then, a classification can be made by **evaluating which text input is most similar to the input image**.
We'll start by importing the Hugging Face Transformers library so that the CLIP model can be downloaded locally. Additionally, the PIL library is used to load images in Python.

```
from transformers import CLIPProcessor, CLIPModel
from PIL import Image
```

Next, we can import a version of the clip model and its associated data processor. *Note: the processor handles tokenizing input text and image preparation.*

```
# import model
model = CLIPModel.from_pretrained("openai/clip-vit-base-patch16")

# import processor (handles text tokenization and image preprocessing)
processor = CLIPProcessor.from_pretrained("openai/clip-vit-base-
patch16")
```

Multimodal Model LLM

https://towardsdatascience.com/multimodal-models-llms-that-can-see-and-hear-5c6737c981d3/

What is a Multimodal Model?

A **Multimodal Model (MM)** is an **AI system that can process multiple data modalities as input or output (or both)** [1]. Below are a few examples.

- **GPT-4o** – Input: text, images, and audio. Output: text.
- **FLUX** – Input: text. Output: images.
- **Suno** – Input: text. Output: audio.

One benefit of using existing LLM as a starting point for MMs is that they've **demonstrated a strong ability to acquire world knowledge through large-scale pre-training**, which can be leveraged to process concepts appearing in non-textual representations.

3 Paths to Multimodality

Here, I will focus on multimodal models developed from an LLM. Three popular approaches are described below.

1. **LLM + Tools**: Augment LLMs with pre-built components
2. **LLM + Adapters**: Augment LLMs with multi-modal encoders or decoders, which are aligned via adapter fine-tuning
3. **Unified Models**: Expand LLM architecture to fuse modalities at pre-training

Path 1: LLM + Tools

The simplest way to make an LLM multimodal is by **adding external modules that can readily translate between text and an arbitrary modality**. For example, a transcription model (e.g. Whisper) can be connected to an LLM to translate input speech into text, or a text-to-image model can generate images based on LLM outputs.

The key benefit of such an approach is **simplicity**. Tools can quickly be assembled without any additional model training. The downside, however, is that the quality of such a system may be limited. Just like when playing a game of telephone, messages mutate when passed from person to person. **Information may degrade going from one module to another via text descriptions only.**

Path 2: LLM + Adapters

One way to mitigate the "telephone problem" is by optimizing the representations of new modalities to align with the LLM's internal concept space. For example, ensuring an image of a dog and the description of one *look* similar to the LLM.

This is possible through the use of **adapters**, a relatively small set of **parameters that appropriately translate a dense vector representation for a downstream model** [2][4][5].

Adapters can be trained using, for example, image-caption pairs, where the adapter learns to translate an image encoding into a representation compatible with the LLM [2][4][6].

Path 3: Unified Models

The final way to make an LLM multimodal is by **incorporating multiple modalities at the pre-training stage**. This works by adding modality-specific tokenizers (rather than pre-trained encoder/decoder models) to the model architecture and expanding the embedding layer to accommodate new modalities [9].

While this approach comes with significantly greater technical challenges and computational requirements, it enables the **seamless integration of multiple modalities**

into a shared concept space, unlocking better reasoning capabilities and efficiencies [10].
The preeminent example of this unified approach is (presumably) GPT-4o, which processes text, image, and audio inputs to enable **expanded reasoning capabilities at faster inference times than previous versions of GPT-4**. Other models that follow this approach include Gemini, Emu3, BLIP, and Chameleon

Multimodal RAGS

https://medium.com/data-science/multimodal-rag-process-any-file-type-with-ai-e6921342c903

What is RAG?

RAG is an approach for**improving a model's response quality by dynamically providing the relevant context**for a given prompt. Here's an example of when this might be helpful.

Say, I forgot the name of a Python library a colleague mentioned in yesterday's meeting. This isn't something ChatGPT can help me with because it does not know the meeting's contents.

However, RAG could help with this by taking my question (e.g. "What was the name of that Python library that Rachel mentioned in yesterday's meeting?"), automatically pulling the meeting transcript, then providing my original query and the transcript to an LLM.

Multimodal RAG

Although improving LLMs with RAG unlocks several practical use cases, there are some situations where relevant information exists in non-text formats, e.g., images, videos, charts, and tables. In such cases, we can go one step further and build **multimodal RAG systems**, **AI systems capable of processing text and non-text data**.

Multimodal RAG enables more sophisticated inferences beyond what is conveyed by text alone. For example, it could analyze someone's facial expressions and speech tonality to give a richer context to a meeting's transcription.

3 Levels of MRAG

While there are several ways to implement a multimodal RAG (MRAG) system, here I will focus on three basic strategies at increasing levels of sophistication.

1. Translate modalities to text.
2. Text-only retrieval + MLLM
3. Multimodal retrieval + MLLM

A simple way to make a RAG system multimodal is by **translating new modalities to text before storing them in the knowledge base**. This could be as simple as converting meeting recordings into text transcripts, using an existing multimodal LLM (MLLM) to generate image captions, or converting tables to a readable text format (e.g., .csv or .json).

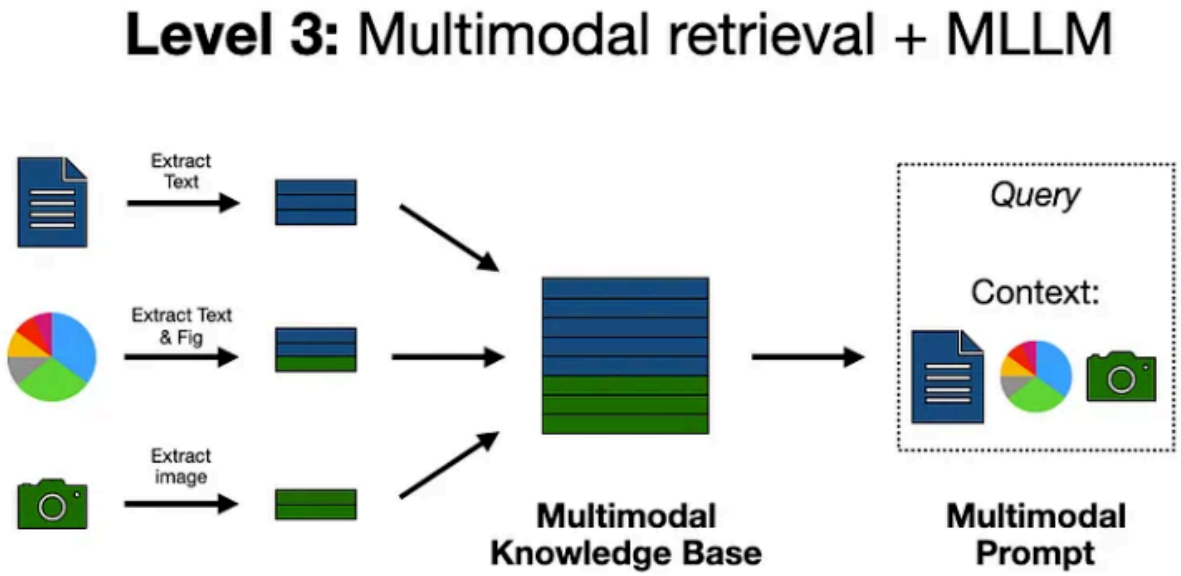
The key upside of this approach is that it **requires minimal changes to an existing RAG system**. Additionally, by explicitly generating text representations of non-text modalities, one has better control over the features of the data to extract. For instance, captions of analytical figures may include both a description and key insights.

Of course, the downside of this strategy is that the **model's responses cannot directly use non-textual data**, which means that the translation from, say, image to text can create a critical information bottleneck.

Text-only retrieval + MLLM

Another approach is to generate text representations of all items in the knowledge base, e.g., descriptions and meta-tags, for retrieval, but to **pass the original modality to a multimodal LLM (MLLM)**. For example, image metadata is used for the retrieval step, and the associated image is passed to a model for inference.
Multimodal RAG systems can synthesize knowledge stored in a variety of formats, expanding what's possible with AI. Here, we reviewed 3 simple strategies for developing such a system and then saw an example implementation of a multimodal blog QA assistant.
The key difference with this approach is that it requires an **MLLM**, which is **an LLM capable of processing non-text data**. This unlocks more advanced reasoning capabilities, as demonstrated by models like GPT-4o or LLaMA 3.2 Vision.
Although we could use keyword-based search in the retrieval processes for Level 1 and Level 2, it is a common practice to use so-called **vector search**. This consists of **generating vector representations (i.e., embeddings)** of items in the knowledge base

and then **performing a search by computing similarity scores** between an input query and each item in the knowledge base.



Although the example worked well enough for this demonstration, there are clear limitations to the search process. A few techniques that may improve this include using a **reranker to refine similarity search** results and to improve search quality via **fine-tuned multimodal embeddings**.

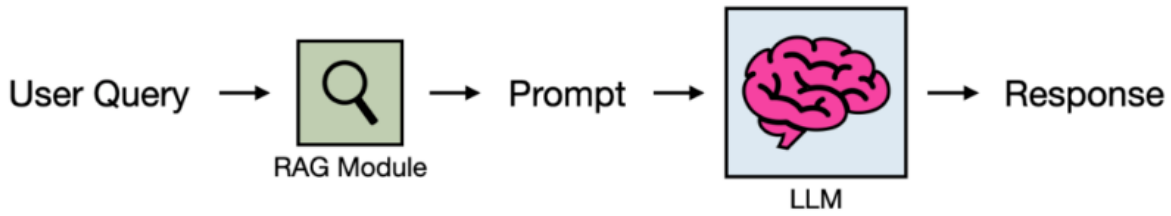
How to Improve LLMs with RAG - Retrieval Augmented Generation

https://towardsdatascience.com/how-to-improve-llms-with-rag-abdc132f76ac/?sk=d8d8ecfb1f6223539a54604c8f93d573

Large language models (LLMs) have demonstrated an impressive ability to store and deploy vast knowledge in response to user queries. While this has enabled the creation of powerful AI systems like ChatGPT, compressing world knowledge in this way has **two key limitations**.
First, an LLM's knowledge is static, i.e., not updated as new information becomes available. **Second**, LLMs may have an insufficient "understanding" of niche and specialized information that was not prominent in their training data. These limitations can result in undesirable (and even fictional) model responses to user queries.

One way we can mitigate these limitations is to **augment a model via a specialized and mutable knowledge base**, e.g., customer FAQs, software documentation, or product catalogs. This enables the creation of more robust and adaptable AI systems.
Retrieval augmented generation, or **RAG**, is one such approach. Here, I provide a high-level introduction to RAG and share example Python code for implementing a RAG system using LlamaIndex.

RAG works by adding a step to this basic process. Namely, a retrieval step is performed where, based on the user's prompt, the relevant information is extracted from an external knowledge base and injected into the prompt before being passed to the LLM.



How it works

There are 2 key elements of a RAG system: a **retriever** and a **knowledge base**.

Retriever

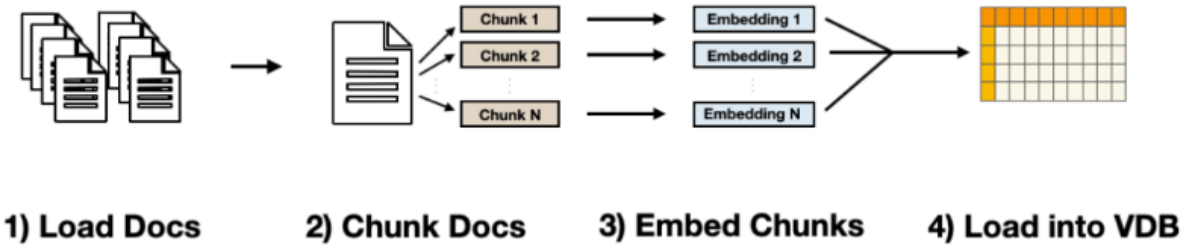
A retriever takes a user prompt and returns relevant items from a knowledge base. This typically works using so-called **text embeddings**, numerical representations of text in concept space. In other words, these are **numbers that represent the *meaning* of a given text**.
Text embeddings can be used to compute a similarity score between the user's query and each item in the knowledge base. The result of this process is a **ranking of each item's relevance to the input query**.
The retriever can then take the top k (say k=3) most relevant items and inject them into the user prompt. This augmented prompt is then passed into the LLM for generation.

Knowledge Base

The next key element of a RAG system is a knowledge base. This **houses all the information you want to make available to the LLM**. While there are countless ways to construct a knowledge base for RAG, here I'll focus on building one from a set of documents.
The process can be broken down into **4 key steps** [2,3].

1. **Load docs** – This consists of gathering a collection of documents and ensuring they are in a ready-to-parse format (more on this later).
2. **Chunk docs—**Since LLMs have limited context windows, documents must be split into smaller chunks **(e.g.,** 256 or 512 characters long).

3. **Embed chunks** – Translate each chunk into numbers using a text embedding model.
4. **Load into Vector DB**— Load text embeddings into a database (aka a vector database).



Accelerating Growth in Indonesia: An industrial policy for the rural informal sector

https://www.brookings.edu/articles/accelerating-growth-in-indonesia-an-industrial-policy-for-the-rural-informal-sector/#:~:text=They%20do%20not%20register%20to,selected%20locations%20as%20specified%20below.

Pintarnya Raises $16.7M to power jobs and financial services in Indonesia

- https://techcrunch.com/2025/08/24/pintarnya-raises-16-7m-to-power-jobs-and-financial-services-in-indonesia/
- https://pintarnya.com/

Pintarnya's go-to-market strategy was centered around acquiring and serving Indonesia's large base of blue-collar and informal workers by digitizing traditional job matching and lending processes, leveraging partnerships, and using AI-driven technology.

Key Go-To-Market Elements

Target Audience Focus

- Pintarnya specifically targeted Indonesia's 88 million informal workers, recognizing the pain points of searching for jobs offline and barriers to accessing formal financial services.
- The company focused on sectors with high demand and low tech penetration, such as food and beverage, hospitality, retail, and logistics.

Product and Platform Features

- Launched with an AI-powered app and web portal offering verified, curated job listings to improve trust and efficiency for job seekers and employers.
- Implemented features such as CV creation tools, job recommendations, and green-shield verification for postings to address fraud and improve user experience.

Partnerships and Ecosystem Integration

- Partnered with asset-backed lenders and local employers to provide not just jobs but also access to loans using alternative credit scoring (collateral such as gold or electronics), thus creating a full-service employment and finance solution.
- Built up a trusted, verified network by collaborating with both traditional and digital sector partners for outreach and user acquisition.

Bottom-Up and Localized Marketing

- Began go-to-market activities in sectors reopening after COVID (F&B, hospitality, logistics), utilizing localized branding and outreach campaigns to stimulate immediate job seeker sign-ups and employer engagement.
- Used digital marketing, social platforms, and community endorsements to rapidly build awareness, supplemented by pilot programs and branch launches in strategic regions.

Data-Driven Scaling

- Leveraged conversion data and user behavior analytics to refine recommendation algorithms and customize employer/jobseeker onboarding, increasing match rates and retention over time.
- Actively experimented with the platform, refining the matching process to maximize successful connections and build a track record of authentic, effective employment outcomes.

Pintarnya's go-to-market strategy effectively bridged offline-worker communities to online platforms, integrated financial and employment solutions, and scaled through partnerships, verification, and highly targeted marketing.

Pintarnya initially used digital channels to acquire users, focusing on social media marketing, partnerships, and direct outreach to local communities and employers serving blue-collar workers.

First User Acquisition Channels

Social Media and Digital Marketing

- Launched marketing campaigns across Facebook, Instagram, and WhatsApp to quickly reach blue-collar job seekers where they already spend time online.
- Digital promotions and targeted ads highlighted platform benefits and helped drive early sign-ups and engagement for job seekers.

Local Partnerships and Community Referrals

- Partnered with local employers and community-based organizations to reach workers through trusted channels.
- Leveraged referrals and informal word-of-mouth networks popular in Indonesia's blue-collar segments, accelerating user growth just months after launch.

Employer Onboarding and Job Listings

- Onboarded partner employers in target sectors (F&B, hospitality, retail, logistics) to provide curated job opportunities, attracting job seekers looking for credible listings.
- Used employer networks to amplify user acquisition, with employers actively helping recruit job seekers to the platform.

Pintarnya's rapid traction in its first months reflects the effectiveness of combining digital, social, and community channels tailored for Indonesia's blue-collar workforce

Partnerships significantly accelerated Pintarnya's market entry by enabling the company to quickly build trust, expand reach, and provide essential services that met both employment and financial needs of Indonesia's blue-collar workforce.

Key Ways Partnerships Accelerated Market Entry

Rapid Scaling Through Established Networks

- Partnering with asset-backed lendersallowed Pintarnya to offer secured loans backed by collateral (such as gold, electronics, or vehicles), overcoming credit approval barriers faced by informal workers and making the platform immediately useful to a large segment.
- By collaborating with local employers and financial institutions, Pintarnya plugged directly into pre-existing networks, enabling extensive user onboarding and job listing distribution without the time or cost required to build everything from scratch.

Building Trust and Reducing Friction

- Partnerships with recognized employers and trusted financial institutions lent credibility and legitimacy to Pintarnya's brand from day one, crucial for attracting users in a market wary of scams and ineffectual digital platforms.
- These alliances ensured that jobseekers and workers were matched with verified opportunities and could access loans safely and responsibly, enhancing user satisfaction and word-of-mouth growth.

Product Differentiation and User Retention

- Integrated partnerships enabled Pintarnya to combine employment and lending in a way that competitors—who focused on either jobs or credit but not both—could not match, making Pintarnya a "one-stop" solution for its target market.
- Ongoing collaborative innovation (such as developing micro-savings and investment products with partners) positioned Pintarnya for ongoing product leadership and user retention.

By leveraging partnerships, Pintarnya broke through the initial market adoption barrier, grew its user base quickly, and achieved distinct status in Indonesia's employment and financial inclusion space.

Pintarnya's pricing model was designed to encourage fast adoption and growth, especially among Indonesia's blue-collar and informal workforce who are price-sensitive and often left out by traditional job and financial platforms.

Support for Adoption and Growth

Low or No Barriers for Job Seekers

- Job seekers could register, create CVs, and access verified job postings for free or negligible cost, removing one of the main barriers to platform adoption in their target market.
- This approach significantly increased sign-ups and early engagement, as it matched the behavior and financial capacity of blue-collar workers.

Affordable and Transparent Lending

- Pintarnya partnered with asset-backed lenders and structured loans for affordability, using collateral like gold or electronics to secure better rates and more responsible lending terms.
- Loan approvals leveraged alternative data (employment history, job matches, collateral), helping users avoid predatory lending and debt traps while enabling access to essential financing.

Monetization via Employers and Financial Partners

- Employers interested in faster hiring and premium placement for jobs paid for listings, screening, or data-driven hiring solutions, while basic listings often remained accessible to maximize job seeker liquidity and platform growth.
- Pintarnya's partnerships with lenders and financial institutions enabled co-branded financial products—generating revenue without upfront costs for job seekers.

Gradual Introduction of Paid Value-Added Services

- While core platform access remained free for job seekers, Pintarnya planned to introduce micro-savings, investments, and upskilling content as optional, likely low-cost premium features as user trust and needs matured.

This pricing strategy made Pintarnya immediately attractive to workers and scalable for employers/partners, supporting both massive initial user acquisition and sustainable long-term monetization

Text Embeddings, Classification, and Semantic Search

https://towardsdatascience.com/text-embeddings-classification-and-semantic-search-8291746220be/?sk=03e4e68a420373a3525de8721f57c570

The Fastest Way to Install and Use Python with UV

https://shawhin.medium.com/the-fastest-way-to-install-and-use-python-uv-149e29ee9780

OpenAI Prompt Engineering

https://platform.openai.com/docs/guides/prompt-engineering

Anthropic Prompt Engineering

https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview#prompting-vs-finetuning

Prompt Engineering - How to Trick AI into Solving Your Problems

Building AI Apps with Prompt Engineering

The less easy way unlocks a **new paradigm of programming and software development**. No longer are developers required to define every inch of logic in their software systems. They now have the option to offload a non-trivial portion to LLMs. Let's look at a concrete example of what this might look like.

Suppose you want to create an **automatic grader for a high school history class**. The trouble, however, is that all the questions have written responses, so there often can be

multiple versions of a correct answer. For example, the following responses to "*Who was the 35th president of the United States of America?*" could be correct.

- John F. Kennedy
- JFK
- Jack Kennedy (a common nickname)
- John Fitzgerald Kennedy (probably trying to get extra credit)
- John F. Kenedy (misspelled last name)

In the **traditional programming paradigm**, it was on the developer to figure out how to account for all these variations. To do this, they might list all possible correct answers and use an exact string-matching algorithm or maybe even use fuzzy matching to help with misspelled words.

However, with this new **LLM-enabled paradigm**, **the problem can be solved through simple prompt engineering**. For instance, we could use the following prompt to evaluate student answers.

```
You are a high school history teacher grading homework assignments.
Based on the homework question indicated by "Q:" and the correct answer
indicated by "A:", your task is to determine whether the student's
answer is
correct.
Grading is binary; therefore, student answers can be correct or wrong.
Simple misspellings are okay.

Q: {question}
A: {correct_answer}

Student Answer: {student_answer}
```

We can think of this prompt as a function, where given a ***question***, **_correct_answer**_, and_ **student_answer_**, it generates the student's grade. This can then be integrated into a larger piece of software that implements the automatic grader.

In terms of time-saving, this prompt took me about 2 minutes to write, while if I were to try to develop an algorithm to do the same thing, it would take me hours (if not days) and probably have worse performance. **So the time savings for tasks like this are 100–1000x**.

Trick 1: Be Descriptive (More is Better)

A defining feature of LLMs is that they are trained on massive text corpora. This equips them with a vast knowledge of the world and the ability to perform an enormous variety of tasks. However, this impressive generality may hinder performance on a specific task if the proper context is not provided.
For example, let's compare two prompts for generating a birthday message for my dad.

Without Trick

```
Write me a birthday message for my dad.
```

With Trick

```
Write me a birthday message for my dad no longer than 200
characters. This is a big birthday because he is turning 50. To
celebrate,
I booked us a boys' trip to Cancun. Be sure to include some cheeky
humor, he
loves that.
```

Trick 2: Give Examples

The next trick is to give the LLM example responses to improve its performance on a particular task. The technical term for this is **few-shot learning,** and has been shown to improve LLM performance significantly [6].
Let's look at a specific example. Say we want to write a subtitle for a Towards Data Science article. We can use existing examples to help guide the LLM completion.

Without Trick

```
Given the title of a Towards Data Science blog article, write a subtitle
for it.

Title: Prompt Engineering—How to trick AI into solving your problems
Subtitle:
```

With Trick

```
Given the title of a Towards Data Science blog article, write a subtitle
for it.

Title: A Practical Introduction to LLMs
Subtitle: 3 levels of using LLMs in practice
```

```
Title: Cracking Open the OpenAI (Python) API
Subtitle: A complete beginner-friendly introduction with example code

Title: Prompt Engineering-How to trick AI into solving your problems
Subtitle:
```

Trick 3: Use Structured Text

Ensuring prompts follow an organized structure not only makes them easier to read and write, but also tends to help the model generate good completions. We employed this technique in the example for **Trick 2**, where we explicitly labeled the *title* and *subtitle* for each example.

However, there are countless ways we can give our prompts structure. Here are a handful of examples: use ALL CAPS for emphasis, use delimiters like “` to highlight a body of text, use markup languages like Markdown or HTML to format text, use JSON to organize information, etc.

Now, let’s see this in action.

Without Trick

```
Write me a recipe for chocolate chip cookies.
```

With Trick

```
Create a well-organized recipe for chocolate chip cookies. Use the
following
formatting elements:

**Title**: Classic Chocolate Chip Cookies
**Ingredients**: List the ingredients with precise measurements and
formatting.
**Instructions**: Provide step-by-step instructions in numbered format,
detailing the baking process.
**Tips**: Include a separate section with helpful baking tips and
possible variations.
```

Trick 4: Chain of Thought

This trick was proposed by Wei et al. [7]. The basic idea is to guide an LLM to think "step by step". This helps break down complex problems into manageable sub-problems, which gives the LLM "time to think" [3,5]. Zhang et al. showed that this could be as simple as including the text "*Let’s think step by step*" in the prompt [8].

This notion can be extended to any recipe-like process. For example, if I want to create a LinkedIn post based on my latest Medium blog, I can guide the LLM to mirror the step-by-step process I follow.

Without Trick

```
Write me a LinkedIn post based on the following Medium blog.

Medium blog: {Medium blog text}
```

With Trick

```
Write me a LinkedIn post based on the step-by-step process and Medium
blog
given below.

Step 1: Come up with a one line hook relevant to the blog.
Step 2: Extract 3 key points from the article
Step 3: Compress each point to less than 50 characters.
Step 4: Combine the hook, compressed key points from Step 3, and a call
to action
to generate the final output.

Medium blog: {Medium blog text}
```

Trick 5: Chatbot Personas

A somewhat surprising technique that tends to improve LLM performance is to prompt it to take on a particular persona e.g. "*you are an expert*". This is helpful because you may not know the best way to describe your problem to the LLM, but you may know who would help you solve that problem [1]. Here's what this might look like in practice.

Without Trick

```
Make me a travel itinerary for a weekend in New York City.
```

With Trick

```
Act as an NYC native and cabbie who knows everything about the city.
Please make me a travel itinerary for a weekend in New York City based
on
your experience. Don't forget to include your charming NY accent in your
response.
```

Trick 6: Flipped Approach

It can be difficult to optimally prompt an LLM when **we do not know what it knows or how it thinks**. That is where the "flipped approach" can be helpful. This is where you prompt the LLM to ask you questions until it has a sufficient understanding (i.e. context) of the problem you are trying to solve.

Without Trick

```
What is an idea for an LLM-based application?
```

With Trick

```
I want you to ask me questions to help me come up with an LLM-based
application idea. Ask me one question at a time to keep things
conversational.
```

Trick 7: Reflect, Review, and Refine

This final trick prompts the model to reflect on its past responses to improve them. Common use cases are having the model critically evaluate its own work by asking it if it "*completed the assignment*" or having it "*explain the reasoning and assumptions*" behind a response [1, 3].

Additionally, you can ask the LLM to refine not only its responses but **your prompts**. This is a simple way to automatically rewrite prompts so that they are easier for the model to "understand".

With Trick

```
Review your previous response, pinpoint areas for enhancement, and offer
an
improved version. Then explain your reasoning for how you improved the
response.
```

Practical Introduction to LLMs

https://towardsdatascience.com/a-practical-introduction-to-llms-65194dda1148/?sk=960e586f4fd6eae65db69e8f7254f13f

LLM is short for **Large Language Model**, which is a recent innovation in AI and machine learning. This powerful new type of AI went viral in Dec 2022 with the release of ChatGPT. Okay, so LLMs are a special type of language model, **but what makes them special?**

There are **2 key properties** that distinguish LLMs from other language models. One is quantitative, and the other is qualitative.

1. **Quantitatively**, what distinguishes an LLM is the number of parameters used in the model. Current LLMs have on the order of **10–100 billion parameters** [1].
2. **Qualitatively**, something remarkable happens when a language model becomes "large." It exhibits so-called ***emergent properties*** e.g. zero-shot learning [1]. These are **properties that seem to suddenly appear** when a language model reaches a sufficiently large size.

Zero-shot Learning

The major innovation of GPT-3 (and other LLMs) is that it is capable of **zero-shot learning** in a wide variety of contexts [2]. This means ChatGPT can **perform a task even if it has not been explicitly trained to do it**.

While this might be no big deal to us highly evolved humans, this zero-shot learning ability starkly contrasts the prior machine learning paradigm.

Previously, a model needed to be **explicitly trained on the task it aimed to do** in order to have good performance. This could require anywhere from 1k-1M pre-labeled training examples.

For instance, if you wanted a computer to do language translation, sentiment analysis, and identify grammatical errors. Each of these tasks would require a specialized model trained on a large set of labeled examples. Now, however, **LLMs can do all these things without explicit training**.

How do LLMs work?

The core task used to train most state-of-the-art LLMs is **word prediction**. In other words, given a sequence of words, **what is the probability distribution of the next word**?

For example, given the sequence "Listen to your _____," the most likely next words might be: heart, gut, body, parents, grandma, etc. This might look like the probability distribution shown below.

Interestingly, this is the same way many (non-large) language models have been trained in the past (e.g. GPT-1) [3]. However, for some reason, when language models get beyond a certain size (say ~10B parameters), these (emergent) abilities, such as zero-shot learning, can start to pop up [1].

Although there is no clear answer as to *why* this occurs (only speculations for now), it is clear that LLMs are a powerful technology with countless potential use cases.

Life insurance leads Nigeria’s insurance market with ₦276.8 billion in premiums

https://intelpoint.co/insights/life-insurance-leads-nigerias-insurance-market-with-%e2%82%a6276-8-billion-in-premiums/

Key Takeaways

- Life insurance dominates the market with ₦276.8 billion in gross premiums, more than any other sector.
- Oil and Gas insurance follows as the second-largest segment, generating ₦188.7 billion in Q1.
- Fire and Motor insurance sectors contributed ₦91.9 billion and ₦77.7 billion respectively, reflecting strong demand.
- Aviation insurance recorded the least income at ₦16.6 billion, likely due to the limited scope of operations.
- The top three segments (Life, Oil & Gas, and Fire) jointly account for over 75% of the total GPI in the quarter.

Nigeria Foreign Trade Q1 2025

https://intelpoint.co/insights/nigerias-foreign-trade-2/
https://intelpoint.co/insights/port-activity-in-nigeria/

Exports: 20.6T Naira
80% + Oil and Petroleum
Agriculture 8.3%
Raw Materials and Goods(e.g. leather, rubber, etc) - 5.1%

Imports: 15.4T Naira
75% - Oil and Petroleum
Raw Materials and Goods - 11.7%
Agricultural Goods - 6.71%

Trade Surplus of 5.17T Naira

- Apapa Port accounted for 71.6% of Nigeria’s total trade value in Q1 2025 and 82.12% of total exports
- Apapa Port handled ₦25.79 trillion worth of goods in Q1 2025, representing 71.6% of total trade. It remains the country’s primary trade hub, far surpassing all other ports combined.
- Apapa alone facilitated ₦17.74 trillion or 86.1% of Nigeria’s total exports, showing a high dependency on a single location for outbound goods.
- Tin Can Island is the only meaningful secondary hub With ₦3.44 trillion (9.5%) in total trade, ranking a distant second. It’s the only other port contributing more than ₦1 trillion each to imports and exports.
- Lekki has limited export impact, despite handling ₦1.70 trillion in imports. Lekki contributed only ₦0.30 trillion (1.5%) in exports, indicating underutilization for outbound trade.
- Murtala Muhammed International Airport processed just ₦647.91 billion (1.8%) of total trade, reinforcing that Nigeria’s international trade remains heavily maritime-focused.

Nigeria Non Oil Tax Accounted for 70% of Tax Revenue in 2024

https://intelpoint.co/insights/non-oil-company-income-tax-and-two-other-sources-accounted-for-over-70-of-nigerias-tax-revenue-in-2024/

AI in Africa 2025 (Mastercard)

archives demonstrate a step towards improving local data availability and quality
Building Stronger Data Ecosystems: Establishing stronger local data ecosystems and processes to collect, manage, and responsibly share high-quality data sets is essential to harness Africa’s diversity and provide contextually meaningful insights for decision-makers. Current challenges include fragmented data, reliance on imported algorithms trained on foreign datasets, and issues with data quality, timeliness, and accessibility.

Efforts like Nigeria's project to digitize national archives demonstrate a step towards improving local data availability and quality

Accelerating Cross-Border Payments: The accelerated development of AI-integrated cross-border payment systems can improve regional trade and economic integration, lower transaction barriers for small businesses and informal traders, and bridge financial inclusion gaps for millions of unbanked individuals. Africa has already led in mobile money adoption, which revolutionized financial inclusion and laid the groundwork for integrating more sophisticated technologies like AI through fintech innovation. However, a fragmented policy and regulatory landscape across 54 countries remains a key barrier to seamless cross-border operations

Celebrating Success Stories: Amplifying the success stories of entrepreneurs and researchers can create relatable role models, inspiring broader AI adoption and innovation. Highlighting these achievements can empower local talent, resolve continent-specific perception challenges, and drive meaningful global participation. The document itself showcases numerous successes in countries like Kenya, South Africa, and Nigeria, and through initiatives like Farmerline and Rising Academies

Finance & Banking: The financial services sector has seen the greatest impact of AI-backed tools, promoting financial inclusion, enhancing security, and streamlining operations

Financial Inclusion and Credit Scoring: AI-driven microfinance and digital payments are expanding financial access. AI analyzes mobile money data to assess loan risk more accurately than conventional methods, opening up funding for those previously excluded. Companies like Kenya-based Tala analyze mobile phone usage and payment behavior to approve micro-loans. Jumo uses AI and ML to create tailored financial products for the underbanked across several African countries. Kenya's Hustler Fund uses AI-driven credit scoring for low-interest microloans to small businesses. M-KOPA uses AI-integrated IoT technologies for its digital micropayments platform, processing up to 500 payments per minute for over 3 million customers

Fraud Detection and Risk Management: AI enhances security and regulatory compliance through better fraud detection and faster Know Your Customer (KYC) checks. AI-powered systems detect identity theft, account takeovers, and false transactions. Mastercard's fraud detection systems, built and trained in African cities like Lagos, Nairobi, and

Johannesburg, process transactions in real time. Mastercard uses generative AI to double the speed at which it can flag potentially compromised cards. Paystack employs machine learning for real-time transaction analysis, flagging suspicious behavior

Driving Economic Growth and Job Creation: AI is expected to contribute approximately USD 15.7 trillion to the global economy by 2030, and Africa is well-placed to tap into a substantial portion of this surge. The estimated market size for AI in Africa is USD 4.51 billion in 2025, projected to grow to USD 16.53 billion by 2030, with a Compound Annual Growth Rate (CAGR) of 27.42%. This growth trajectory is forecast to create as many as 230 million digital jobs in Sub-Saharan Africa by 2030

Addressing Financial Inclusion: AI has the potential to revolutionize financial inclusion, reaching underserved communities and promoting broader access to financial services. Africa's early adoption of 'mobile money' (e.g., M-Pesa) has already laid a strong foundation, enabling millions of unbanked individuals to access financial services, which in turn facilitates the integration of more sophisticated AI technologies through fintech innovation. AI can analyze alternative data, like mobile money usage, to assess loan risk more accurately than traditional methods, opening up funding for those previously excluded. This is particularly critical given that over 400 million people in Sub-Saharan Africa are financially unserved or underserved

Existing Digital Infrastructure: Africa has already shown leadership in technology adoption, particularly with mobile money, which created a widespread platform for payments, savings, and credit, laying the groundwork for AI integration. Rapid gains from AI are possible if this infrastructure continues to power electricity access and digitization

Opportunities in Data Ecosystems: Africa's emerging digital ecosystems are generating vast data sets. Building stronger local data ecosystems and processes to collect, manage, and responsibly share high-quality, contextually diverse data is crucial. This data, reflective of African realities, can empower decision-makers with meaningful insights

In essence, AI is seen as an unparalleled chance for Africa to "leapfrog traditional development barriers" and achieve growth that is both transformative and inclusive, ultimately improving lives and expanding access to essential services

National Programs: Countries are launching large-scale talent development programs. Nigeria's 3 Million Technical Talent (3MTT) program, for instance, aims to train 150,000

people in AI and machine learning, positioning the country as a leading hub for digital skills.
Detailed Timeline
2012 (March 22):

Memorial Sloan Kettering Cancer Center and IBM announce collaboration to apply Watson technology to oncology, demonstrating an early application of AI in healthcare.
2014:

Rising Academies is established, expanding to over 300,000 students in Sierra Leone, Liberia, Ghana, and Rwanda, and later developing the AI-powered math tutor 'Rori'.
2017:

Deep Learning Indaba is launched in South Africa, aiming to strengthen Africa's AI and Machine Learning community by equipping researchers and practitioners with skills and resources.
FruitPunch AI is founded, beginning with the 'AI for Wildlife' project to detect poachers in South African wildlife reserves using autonomous drones.
2018 (October 1):

M-Pesa and its "market-led" approach to financial inclusion are discussed, highlighting Kenya's leadership in mobile money.
2019:

The African Union (AU) establishes a Working Group on AI to develop a "common African stance on AI" and a continent-wide capacity-building framework.
Huawei Cloud establishes a local data center in South Africa, leading to significant growth in cloud services.
OECD Principles on AI are offered as general principles to guide member countries in crafting their AI policies.
2020 (February):

Masakhane, a grassroots organization dedicated to democratizing AI for African languages, has published over 49 translation results for 38-plus African languages on GitHub.
2021 (December 6):

Babylon launches its AI-powered triage tool in Rwanda, improving healthcare access.
2022 (November 1):

Unicef highlights Afrilearn's efforts in providing accessible, adaptive, and affordable learning for Africans.
South Africa's Minister Khumbudzo Ntshavheni launches the Artificial Intelligence Institute of South Africa and AI hubs.
Kenya launches the Hustler Fund, which uses AI-driven credit scoring for microloans.
2023:

The Nigeria Securities and Exchange Commission issues Robo Advisory Services Rules to regulate the sector.
VC investments in AI reach USD 610 million in South Africa, USD 218 million in Nigeria, and USD 15 million in Kenya.
The Malabo Convention (a pan-African instrument for data protection and cybersecurity) comes into force in June.
MoroccoAI Annual Conference is held, discussing recommendations towards a national AI strategy.
The State of AI in Africa Report 2023 is published by the Centre for Intellectual Property and Information Technology Law.
2024 (January):

Microsoft discusses governing AI in Africa, emphasizing policy frameworks.
Nigeria's National Artificial Intelligence Strategy (NAIS) is launched, outlining the country's vision to become a global leader in AI.
Oxford Insights releases the Government AI Readiness Index 2024, ranking African countries.
Aurora Health Systems, founded in 2022, is highlighted for developing AI-based software for early detection of heart and lung diseases.
Huawei Cloud in South Africa experiences significant growth, more than 16 times since its establishment in 2019.
2024 (March):

Google's The Keyword highlights how AI supports early disease detection in India, serving as a model for Africa.
Farmerline launches Darli AI, an AI-powered agricultural tool accessible via WhatsApp chatbots and IVR system, supporting over 27 languages.
Amb. Philip Thigo is appointed Kenya's inaugural Special Envoy on Technology.
Jacaranda Health expands UlizaLlama, its open-source LLM, to provide support in five African languages.
2024 (April):

Moroccan proposes establishment of national agency for AI governance.
Boston Consulting Group publishes findings on consumer awareness of AI.
The Kigali Declaration on Responsible AI in Africa is adopted, underscoring the continent's collective commitment to ethical AI.
2024 (May 13):

Mauritius hosts its inaugural AI Summit under UNESCO.
Mastercard accelerates card fraud detection with generative AI technology.
2024 (July):

Logidoo, a Pan-African logistics start-up, receives a USD 50,000 grant to develop AI solutions for Africa's logistics market.
The AU's Continental Artificial Intelligence Strategy is released.
2024 (August):

South Africa's Department of Communications and Digital Technologies (DoCDT) publishes the National Artificial Intelligence Policy Framework.
2024 (September):

Botswana implements AI-powered drones to monitor wildlife populations and detect poachers.
Uganda's Murchison Falls National Park equips African white-backed vultures with AI-enabled trackers.
The Guardian reports on Kenyan farmers deploying AI to increase productivity.
Trends from 20 years of AI in financial services in Africa are published by Moela et al.
2024 (October):

Morocco launches its Digital 2030 strategy.
TIME magazine recognizes Darli AI as one of the Best Inventions of 2024.
Business Insider Africa lists 10 African countries with the highest number of AI organizations.
Nigeria's Ministry of Communications, Innovation & Digital Economy announces NGN 2.8 billion Google support for AI talent development.
2024 (November):

OJTA discusses AI as a game-changer for just energy transition in Africa.
CBS News reports on Kenya becoming the "Silicon Savannah."

Reuters reports on South Africa's MTN teaming up with China Telecom and Huawei on 5G and AI.
Reuters reports on Huawei Cloud's fast business growth in South Africa.
Walden Dissertations and Doctoral Studies Collection includes a study on fraud detection and prevention in the Nigerian financial industry.
African Development Bank Group discusses skills for employability and productivity in Africa.
2024 (December):

IT News Africa discusses how AI can help combat fraud in Africa.
Morocco World News reports on AI boosting Moroccan bank performance.
MobiHealthNews discusses Babylon's AI-powered triage tool in Rwanda.
Arab Founders describes AI in Egypt as a booming market.
The Guardian reports on AI monitoring cutting stillbirths and neonatal deaths in a Malawi clinic.
2025 (January):

Morocco World News reports on AI boosting Moroccan bank performance and cutting operation time.
The Republic Of Kenya Ministry Of Information, Communications And The Digital Economy releases a draft of the National Artificial Intelligence Strategy 2025–2030.
Middle East AI News reports on the Egyptian president launching the 2025-2030 National AI Strategy.
Vanguard reports Nigeria surpassing the global average with a 70% AI adoption rate.
The Wall Street Journal reports on vultures helping fight poachers in Uganda's Murchison Falls National Park.
2025 (February):

Nigeria's Federal Executive Council approves the National AI Trust to mobilize resources and guide AI development.
verivAfrica discusses AI's impact on Nigeria's job market.
2025 (April):

Kenya's KBC discusses the country's AI strategy.
The Borgen Project explores the impact of AI on agriculture in Kenya and Nigeria.
2025 (May):

Official statements from Kenyan, Nigerian, Moroccan, and South African governments are received by Mastercard, providing updates on AI initiatives.
The Declaration on Responsible AI in Africa is signed.
2025 (September-October):

South Africa anticipates completing the process for its National AI Policy Framework before the G20 sitting.
2026 (April):

South Africa hopes to adopt its finalized National AI Policy before the end of the next financial year.
2026:

Financial losses from credit card fraud are projected to hit USD 43 billion.
2027:

Kenya's National Optic Fiber Network Backhaul Initiative (NOFBI) aims to deploy over 100,000 km of optical fiber.
2028:

The value of digital transactions in Morocco is expected to rise to USD 8.47 billion.
2030:

AI is estimated to contribute about USD 15.7 trillion to the global economy.
The AI market size in Africa is projected to reach USD 16.53 billion.
230 million digital jobs are forecast in Sub-Saharan Africa due to AI growth.
Morocco's Maroc Digital 2030 strategy envisages USD 1.1 billion in investment and targets 240,000 digital jobs.
Egypt's National AI Strategy 2025–2030 aims to raise the ICT sector’s contribution to GDP to 7.7%, establish over 250 AI companies, and train 30,000 AI professionals.
South Africa aims to develop up to 300 AI start-ups and train 5,000 AI experts.
South Africa has a targeted investment in AI forecast of USD 3.7 billion.
Africa’s youth population is projected to become the largest in the world.
Africa's population is expected to grow by 800 million.

Shaping Africa's Inclusive and Trustworthy Digital Future

https://www.brookings.edu/articles/shaping-africas-inclusive-and-trustworthy-digital-future-how-kenya-is-reimagining-technology-leadership/

Jacaranda Health Open Source LLM for African Languages

https://jacarandahealth.org/jacaranda-launches-open-source-llm-in-five-african-languages/

Darli AI Africa FarmTech in Time Magazine

https://farmerline.co/farmerlines-darli-ai-recognized-on-times-list-of-the-best-inventions-of-2024/

Google in Africa Fund

https://blog.google/around-the-globe/google-africa/google-for-africa/

Cassava Technologies Executive Reorganization 2025

https://www.cassavatechnologies.com/cassava-technologies-announces-leadership-changes-to-accelerate-growth-and-strengthen-its-future-competitiveness/

Cassava Technologies and Zindi(Data Scientist Africa)

https://www.cassavatechnologies.com/cassava-technologies-collaborates-with-zindi-to-showcase-african-ai-innovation/

Cassave Technologies and Sand AI Technologies AI Data Centers

https://www.cassavatechnologies.com/cassava-technologies-and-sand-technologies-partner-to-boost-ai-capabilities-and-accessibility-for-african-enterprises/

Cassava Technologies and South Africa GPU as a Service

https://www.cassavatechnologies.com/cassava-technologies-partners-with-the-south-african-artificial-intelligence-association-to-boost-local-access-to-ai-compute-services/

Cassava’s AI-enabled data centres will help Africa develop domestic AI technologies

https://www.cassavatechnologies.com/cassava-to-upgrade-its-data-centres-with-nvidia-supercomputers-to-drive-africas-ai-future/#:~:text=Cassava's%20AI%20Factory%20will%20leverage,to%20power%20AI%20computing%20workloads.

From mobile money to machine learning: Africa’s $16 billion AI opportunity takes shape

https://afridigest.com/mobile-money-machine-learning-africas-16-billion-ai-opportunity-takes-shape/

Unstructured: Preserving Table Structure for Better Retrieval

https://unstructured.io/blog/preserving-table-structure-for-better-retrieval

https://colab.research.google.com/drive/1__axq0MRDR9i1M_uEW-pR8aKYH_Qk1hj?usp=sharing

Unstrcutured: Getting Started with AWS S3 and Redis

https://unstructured.io/blog/getting-started-with-unstructured-and-redis

Unstructured: How We Got Started

https://unstructured.io/blog/how-we-got-started

Speeding Up Text Generation with Non-Autoregressive Language Models

https://unstructured.io/blog/speeding-up-text-generation-with-non-autoregressive-language-models

Intro to Vision Transformers for Document Understanding

https://unstructured.io/blog/an-introduction-to-vision-transformers-for-document-understanding

Documentation in Cursor

https://docs.cursor.com/en/guides/advanced/working-with-documentation
Working with Documentation
How to leverage documentation effectively in Cursor through prompting, external sources, and internal context

Why documentation matters
Documentation provides current, accurate context. Without it, models use outdated or incomplete training data. Documentation helps models understand things like:
Current APIs and parameters
Best practices
Organization conventions
Domain terminology
And much more. Read on to learn how to use documentation right in Cursor without having to context switch.

Model knowledge cutoff
Large language models are trained on data up to a specific point in time, called a "knowledge cutoff." This means:
Recent library updates might not be reflected
New frameworks or tools may be unknown
API changes after the cutoff date are missed
Best practices may have evolved since training
For example, if a model's knowledge cutoff is early 2024, it won't know about features released in late 2024, even for popular frameworks.

Public documentation

External documentation covers publicly available information that models might have limited or outdated knowledge about. Cursor provides two primary ways to access this information.

Using @Docs

@Docsconnects Cursor to official documentation from popular tools and frameworks. Use it when you need current, authoritative information about:

- API references: Function signatures, parameters, return types

- Getting started guides: Setup, configuration, basic usage
- Best practices: Recommended patterns from the source
- Framework-specific debugging: Official troubleshooting guides

Using @Web

`@Web` searches the live internet for current information, blog posts, and community discussions. Use it when you need:

- Recent tutorials: Community-generated content and examples
- Comparisons: Articles comparing different approaches
- Recent updates: Very recent updates or announcements
- Multiple perspectives: Different approaches to problems

Internal documentation

Internal documentation includes information specific to your organization that AI models have never encountered during training. This might be:

- Internal APIs: Custom services and microservices
- Company standards: Coding conventions, architecture patterns
- Proprietary systems: Custom tools, databases, workflows
- Domain knowledge: Business logic, compliance requirements

Accessing internal docs with MCP

Model Context Protocol (MCP) provides a standardized way to bring your private documentation and systems into Cursor. MCP acts as a thin layer between Cursor and your internal resources.Why MCP matters:

- Models can’t guess your internal conventions
- API documentation for custom services isn’t publicly available
- Business logic and domain knowledge is unique to your organization
- Compliance and security requirements vary by company

Common MCP integrations

Integration	Access	Examples
Confluence	Company Confluence spaces	Architecture documentation, API specifications for internal services, coding standards and guidelines, process documentation
Google Drive	Shared documents and folders	Specification documents, meeting notes and decision records, design documents and requirements, team knowledge bases
Notion	Workspace databases and pages	Project documentation, team wikis, knowledge bases, product requirements, technical specifications
Custom	Internal systems and databases	Proprietary APIs, legacy documentation systems, custom knowledge bases, specialized tools and workflows

Custom solutions

For unique needs, you can build custom MCP servers that:

- Scrape internal websites or portals
- Connect to proprietary databases
- Access custom documentation systems
- Pull from internal wikis or knowledge bases

Example custom MCP server for scraping internal docs:

TypeScript

Python

Copy

Ask AI import { McpServer, ResourceTemplate } from "@modelcontextprotocol/sdk/server/mcp.js"; import { StdioServerTransport } from "@modelcontextprotocol/sdk/server/stdio.js"; import { z } from "zod"; import TurndownService from "turndown";

// Create an MCP server for scraping internal docs const server = new McpServer({ name: "internal-docs", version: "1.0.0" });

const turndownService = new TurndownService();

// Add tool to scrape internal documentation server.tool("get_doc", { url: z.string() }, async ({ url }) => { try { const response = await fetch(url); const html = await response.text();

```
  // Convert HTML to markdown   const markdown =
turndownService.turndown(html);     return {     content: [{ type: "text",
text: markdown }]   }; } catch (error) {   return {     content: [{ type:
"text", text: `Error scraping ${url}: ${error.message}` }]   }; }
});
// Start receiving messages on stdin and sending messages on stdout const transport =
new StdioServerTransport(); await server.connect(transport);
```

Keeping docs up to date Documentation becomes stale quickly. Cursor can help you maintain current, useful documentation by generating and updating it based on your actual code and development conversations.

Takeaways

- Documentation as context makes Cursor more accurate and current
- Use `@Docs` for official documentation and `@Web` for community knowledge
- MCP bridges the gap between Cursor and your internal systems
- Generate documentation from code and conversations to keep knowledge current
- Combine external and internal documentation sources for comprehensive understanding

Large Codebases with Cursor

https://docs.cursor.com/en/guides/advanced/large-codebases

For larger changes, spending an above-average amount of thought to create a precise, well-scoped plan can significantly improve Cursor's output. If you find that you're not getting the result you want after a few different variations of the same prompt, consider zooming out and creating a more detailed plan from scratch, as if you were creating a PRD for a coworker. Oftentimes the hard part is figuring out what change should be made, a task suited well for humans. With the right instructions, we can delegate some parts of the implementation to Cursor. One way to use AI to augment the plan-creation process is to use Ask mode. To create a plan, turn on Ask mode in Cursor and dump whatever context you have from your project management systems, internal docs, or loose thoughts. Think about what files and dependencies you have in the codebase that you already know you want to include. This can be a file that includes pieces of code you want to integrate with, or perhaps a whole folder.
Here's an example prompt: Planning prompt

Copy
Ask AI

- create a plan for how we shoud create a new feature (just like @existingfeature.ts)
- ask me questions (max 3) if anything is unclear
- make sure to search the codebase

@Past Chats (my earlier exploration prompts)
here's some more context from [project management tool]: [pasted ticket description]
We're asking the model to create a plan and gather context by asking the human questions, referencing any earlier exploration prompts and also the ticket descriptions. Using a thinking model like claude-3.7-sonnet, gemini-2.5-pro, or o3 is recommended as they can understand the intent of the change and better synthesize a plan.

Pick the right tool for the job
One of the most important skills in using Cursor effectively is choosing the right tool for the job. Think about what you're trying to accomplish and pick the approach that will keep you in flow.
Tool Use case Strength Limitation
Tab Quick, manual changes Full control, fast Single-file
Inline Edit Scoped changes in one file Focused edits Single-file
Chat Larger, multi-file changes Auto-gathers context, deep edits Slower, context-heavy
Each tool has its sweet spot:
Tab is your go-to for quick edits where you want to be in the driver's seat
Inline Edit shines when you need to make focused changes to a specific section of code
Chat is perfect for those bigger changes where you need Cursor to understand the broader context
When you're using Chat mode (which can feel a bit slower but is incredibly powerful), help it help you by providing good context. Use @files to point to similar code you want to emulate, or @folder to give it a better understanding of your project structure. And don't be afraid to break bigger changes into smaller chunks - starting fresh chats helps keep things focused and efficient.

Takeaways
Scope down changes and don't try to do too much at once
Include relevant context when you can
Use Chat, Inline Edit & Tab for what they're best at
Create new chats often
Plan with Ask mode, implement with Agent mode

Cursor Architecture Diagrams

https://docs.cursor.com/en/guides/tutorials/architectural-diagrams

Why diagrams matter

Diagrams clarify how data flows and how components interact. They're useful when you:

- Want to understand flow control in your codebase
- Need to trace data lineage from input to output
- Are onboarding others or documenting your system

They're also great for debugging and asking smarter questions. Visuals help you (and the model) see the bigger picture.

Two dimensions to consider

There are a few different angles to think about:

- Purpose: Are you mapping logic, data flow, infrastructure, or something else?
- Format: Do you want something quick (like a Mermaid diagram) or formal (like UML)?

How to prompt

Start with a clear goal. Here are some common ways to ask:

- Flow control: "Show me how requests go from the controller to the database."
- Data lineage: "Trace this variable from where it enters to where it ends up."
- Structure: "Give me a component-level view of this service."

You can include start and end points, or ask Cursor to find the full path.

Working with Mermaid

Mermaid is simple to learn and renders directly in Markdown (with the right extension). Cursor can generate diagrams like:

- flowchartfor logic and sequences
- sequenceDiagramfor interactions

- classDiagramfor object structure
- graph TDfor simple directional maps

Diagram strategy

Start small. Don’t aim to map everything at once.

- Pick one function, route, or process
- Ask Cursor to diagram that part using Mermaid
- Once you have a few, ask it to combine them

This mirrors the C4 model– where you start at a low level (code or components) and work upward to higher-level overviews.

Recommended flow

1. Start with a detailed, low-level diagram
2. Summarize it into a mid-level view
3. Repeat until you reach the level of abstraction you want
4. Ask Cursor to merge them into a single diagram or system map

Takeaways

- Use diagrams to understand flow, logic, and data
- Start with small prompts and grow your diagram from there
- Mermaid is the easiest format to work with in Cursor
- Start low-level and abstract upward, just like in the C4 model
- Cursor can help you generate, refine, and combine diagrams with ease
-

Web Development in Cursor

https://docs.cursor.com/en/guides/tutorials/web-development

Model Context Protocols

https://modelcontextprotocol.io/docs/learn/architecture#data-layer-protocol
https://modelcontextprotocol.io/docs/learn/architecture

https://modelcontextprotocol.io/docs/learn/server-concepts

Primitives

MCP primitives are the most important concept within MCP. They define what clients and servers can offer each other. These primitives specify the types of contextual information that can be shared with AI applications and the range of actions that can be performed.MCP defines three core primitives that*server*scan expose:

- Tools: Executable functions that AI applications can invoke to perform actions (e.g., file operations, API calls, database queries)
- Resources: Data sources that provide contextual information to AI applications (e.g., file contents, database records, API responses)
- Prompts: Reusable templates that help structure interactions with language models (e.g., system prompts, few-shot examples)

Each primitive type has associated methods for discovery (`*/list`), retrieval (`*/get`), and in some cases, execution (`tools/call`). MCP clients will use the `*/list` methods to discover available primitives. For example, a client can first list all available tools (`tools/list`) and then execute them. This design allows listings to be dynamic.As a concrete example, consider an MCP server that provides context about a database. It can expose tools for querying the database, a resource that contains the schema of the database, and a prompt that includes few-shot examples for interacting with the tools.

Building MCPs with LLMs

https://modelcontextprotocol.io/tutorials/building-mcp-with-llms

https://modelcontextprotocol.io/llms-full.txt

https://github.com/modelcontextprotocol/python-sdk

https://github.com/modelcontextprotocol/typescript-sdk

Preparing the documentation

Before starting, gather the necessary documentation to help Claude understand MCP:
Visit https://modelcontextprotocol.io/llms-full.txt and copy the full documentation text

Navigate to either the MCP TypeScript SDK or Python SDK repository
Copy the README files and other relevant documentation
Paste these documents into your conversation with Claude

Describing your server

Once you've provided the documentation, clearly describe to Claude what kind of server you want to build. Be specific about: What resources your server will expose What tools it will provide Any prompts it should offer What external systems it needs to interact with
For example:
Copy Build an MCP server that:

- Connects to my company's PostgreSQL database
- Exposes table schemas as resources
- Provides tools for running read-only SQL queries
- Includes prompts for common data analysis tasks

Working with Claude

When working with Claude on MCP servers:
Start with the core functionality first, then iterate to add more features
Ask Claude to explain any parts of the code you don't understand
Request modifications or improvements as needed
Have Claude help you test the server and handle edge cases
Claude can help implement all the key MCP features:
Resource management and exposure
Tool definitions and implementations
Prompt templates and handlers
Error handling and logging
Connection and transport setup

Best practices

When building MCP servers with Claude:
Break down complex servers into smaller pieces
Test each component thoroughly before moving on
Keep security in mind - validate inputs and limit access appropriately
Document your code well for future maintenance
Follow MCP protocol specifications carefully

Next steps

After Claude helps you build your server:
Review the generated code carefully
Test the server with the MCP Inspector tool
Connect it to Claude.app or other MCP clients
Iterate based on real usage and feedback
Remember that Claude can help you modify and improve your server as requirements change over time.

How to Use Cursor

https://docs.cursor.com/en/guides/working-with-context

Takeaways

- Context is the foundation of effective AI coding, consisting of intent (what you want) and state (what exists). Providing both helps Cursor make accurate predictions.
- Use surgical context with @-symbols (@code, @file, @folder) to guide Cursor precisely, rather than relying solely on automatic context gathering.
- Capture repeatable knowledge in rules for team-wide reuse, and extend Cursor's capabilities with Model Context Protocol to connect external systems.
- Insufficient context leads to hallucinations or inefficiency, while too much irrelevant context dilutes the signal. Strike the right balance for optimal results.

Worlds Largest Hackathon

https://worldslargesthackathon.devpost.com/

https://worldslargesthackathon.devpost.com/resources#challenge-resources

AI Agents Payments Stack

https://jonturow.substack.com/p/why-agents-need-a-new-payment-stack?mkt_tok=MzgyLUpaQi03OTgAAAGau490x8TH-1obmCgFQQAqdEJTOG7ZlDKdiS7XfND0KsxPDLxxP9MAgdt-QNb8vFFSe6El6VAJ3WGpu8_gpGygYG1O-Inv5qh3a-I-fKG9rALT

https://nekuda.ai/

https://paymanai.com/?mkt_tok=MzgyLUpaQi03OTgAAAGau490x3D2D8g9V9_nZTa3Y7upCgKqMATKtB5hYfzeZB0DnJIgLN3i_poP06aklk_JBocwwME726LVfOxuMnIz2TmnsC3_1VwralP0Ih21hrHQ

Visa recently reported a 1,200% year-over-year increase in traffic to retail sites from AI agents. That scale of growth is reminiscent of early internet days — and it's only the beginning. Today, when ChatGPT recommends a pair of sneakers, it links you to a checkout page. But that's just a workaround. A transitional interface. What if I want an agent to not just suggest a product, but to search, refine, and purchase on my behalf?

That's when today's systems break down.

Agents that try to transact like humans risk getting flagged as bots. Because… they are bots. The current financial infrastructure isn't built for "card not present, human not present" transactions. That's the opportunity.

Nekuda is building the payment stack for AI agents. Their platform enables agents to transact with built-in authentication, guardrails, and compliance, enforced all the way down to the payment networks.

This creates a critical new layer of trust:

Trust for consumers, who can delegate authority to agents with confidence that those agents will act only within clearly defined limits — "buy this if it drops below $500," or "book Delta flight 446 but nothing else."

Trust for developers, who can build powerful agentic workflows without getting flagged as fraud or risking catastrophic misfires — no 10,000 sneaker orders.

Trust for networks, which now have the primitives they need to recognize, process, and authorize agent-driven transactions with confidence.

This isn't just a technical breakthrough. It's an unlock for the agent economy's GDP. Because agents must become trustworthy before they can become trusted. And when they are, they'll be able to take on real, valuable work — buying, booking, sourcing, and transacting — at scale.

Nekuda's vision is bold, but it's not speculative. We were drawn to the urgency and clarity of the founding team, and the speed at which they are building. Since we invested, Nekuda has added Amex Ventures and Visa Ventures as investors, along with key angels like Paul Klein (Browserbase), Sahar Mor (ex-Stripe), and Shyamal Anandkat (OpenAI). That's deep credibility across agent infra and payments.

Most importantly, Nekuda are a launch partner for Visa Intelligent Commerce, Visa's new initiative to extend their network to AI-driven transactions. We expect more collaborations like this in the near future.

Shopify Makes AI Commerce Agents MCPs

https://nekuda.substack.com/p/shopify-makes-commerce-agents-first

Moving Money in the 21st Century

https://a16z.com/how-to-move-money-in-the-21st-century/

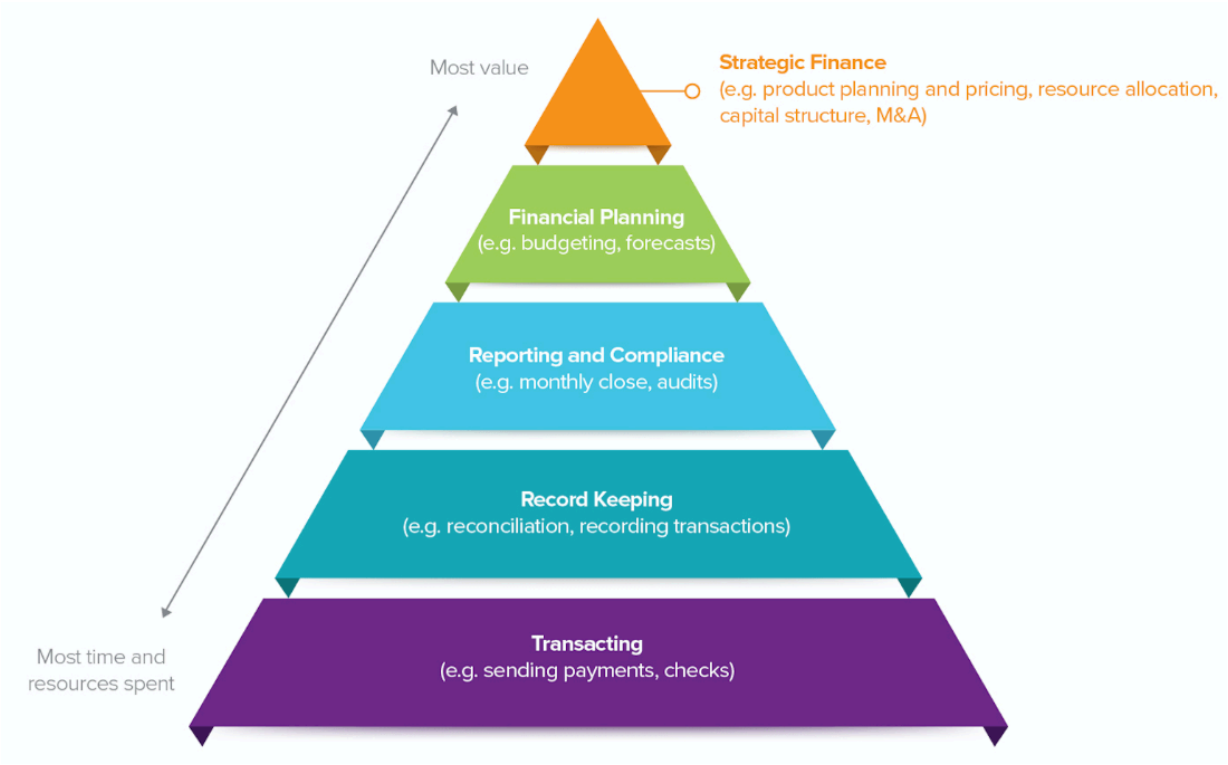
https://a16z.com/the-cfo-in-crisis-mode-modern-times-call-for-new-tools/

Data is stale, limited, and hard to access. An enterprise resource planning (ERP) system, the central repository for financial data, is primarily designed around accounting. As a result, it only gives finance teams a backward-facing view, typically two or more weeks after the month's end. With that lag, the CFO often has difficulty assessing cash burn, revenue, or expenses in real-time. It goes without saying that, particularly now, most companies can't afford to wait six weeks to figure out when they are going to run out of cash.

These building block tools also provide little in the way of predictive forecasting and benchmarking against competing companies. Moreover, retrieving that data, cleaning it, and turning it into a user-friendly format often requires SQL experience, which the finance team may or may not have.

Manual entry and reconciliation never end. The CFO role is plagued by manual, repetitive tasks, from initiating bank transfers to recording checks into the ledger. In preparation for the month-end closing, cash movement (wires, credit cards, bank account balances, etc.) needs to be matched against the total product sold and the invoices paid. All of this information is then painstakingly pulled into reports and investor presentations. Though closing software like FloQast exists to create checklists and flag irregularities, data entry and categorizing is largely still performed by hand.

To make matters worse, none of these products have particularly intuitive interfaces and can take months to train users on. It's why so many finance teams ultimately stick with a familiar workhorse: Excel.


Most value
Strategic Finance
(e.g. product planning and pricing, resource allocation, capital structure, M&A)
Financial Planning
(e.g. budgeting, forecasts)
Reporting and Compliance
(e.g. monthly close, audits)
Record Keeping
(e.g. reconciliation, recording transactions)
Transacting
(e.g. sending payments, checks)
Most time and resources spent

These are some of the areas where we see opportunity:

1. Intelligent Building Blocks. It's time to rebuild tools for point solutions like expense management that can also offer predictive analytics and forecasting. ***Imagine if your expense software told you who usually paid on time and who didn't, sent reminders, and set payment terms based on past performance.*** Already, Brex, Divvy, Airbase, Ramp, and others have replaced clunky legacy systems with products that are intuitive, easily integrated with other systems, and that yield greater control. What's more, many of these companies issue physical and virtual cards to provide an accurate record of all transactions. This saves countless backoffice hours typically spent correcting human error. ***A bigger challenge will be to design a new general ledger, one that pulls in operational data and tracks customer lifetime value in real time, rather than just accounting statements.***
2. Integrated Data Layer. Another approach is to build connective tissue—software that sits on top of existing tools, extracts data, and provides intelligence to help with dynamic planning. It could also provide benchmarking to help companies better understand how their metrics (from compensation to cash burn to days payable) compares to their peers. That might mean a workforce planning tool that assesses your payroll, cap table management, and budget to enable better headcount decisions. That planning tool would also automatically map headcount to revenue growth and planning for product launches. Or this could be a collaborative cash flow management tool that sits on top of procurement and expense management and tracks spending.
3. Banking Operations. Software can also help companies better manage their banking operations. With s***oftware that connects to their bank accounts, companies could understand their cash position, debt, and money movements in real-time. In addition, such software could send and receive payments on an automated basis (better than having to manually fill out a form to initiate wire transfers) and assist with treasury management to reconcile payments.***
4. ***Automating Data Entry and Reporting. Finally, software can help automatically extract and review data. Receipts and checks are not going to disappear overnight, but tech tools like optical character recognition and machine learning can eliminate the need to manually enter invoice information, collect receipts, and match the information against checks received on a company's bank statement. Similarly, reports and board updates can be automatically pre-populated by creating connections into the right data sources.***

Increasingly, companies in this space also have the ability to monetize with fintech, through transaction revenues, lending, or even insurance. This shift has been driven by a few trends. First, companies are eager to transact online—it's viewed as more efficient and

trustworthy than cutting paper checks to pay invoices or payroll. Second, ***we've witnessed the creation of a fintech infrastructure that enables companies to take payments and connect with banking partners. Expense management companies, for example, can easily spin up virtual cards to collect an interchange fee on all expenses filed. Accounts payable (AP) and accounts receivable (AR) software companies can collect transaction fees when customers use their platforms to facilitate payments. Finally, software companies are providing better data to underwrite risk for financial services. Cash flow lending and invoice factoring, for example, are possible based on that AP and AR data***. These financial products provide finance software startups with an additional source of revenue. And because those tools are often tied to volume, the market grows—and the product becomes stickier—as customers process more business.

If money is the lifeblood of business, the inflows and outflows of revenue and expenses are like the heartbeat of a company. And while tracking and managing these financial operations are critical to keeping a business healthy, the tools that exist today are incredibly antiquated. Meanwhile, payment complexity is increasing as businesses become more global and seek to add new payment methods every day.

To keep up, many of the largest companies have invested in homegrown solutions and rely on an army of engineers and operations specialists. Companies like Airbnb and Uber have teams of more than 200 simply to keep track of their real-time financial health, both cash on hand as well as payables/receivables.

Worse yet, breakdowns in these payment operations can have massive consequences. Last year, Matt Levine famously wrote about the payments debacle where **Citibank wired $900 million** to a hedge fund—by mistake. The cause? A set of email approvals and legacy software that didn't catch the error. And this isn't an isolated instance—companies spend millions of dollars per year on audits and many more on fines and lost revenue for not tracking payments correctly.

Companies need (and deserve) better software to manage day-to-day finance workflows and **make CFOs more strategic**. We believe that **"financial operations" (FinOps) is an emerging software category with a massive opportunity to rethink the way businesses manage their money** by streamlining, automating, and optimizing existing work while giving the entire organization a real-time view of a business's financial health.

At its core, there are two money movement workflows, the majority of which are manual today.

- **Payouts** send money outside of the business, including sending money via bank wires and ACH to pay vendor invoices or make disbursements to customers, generating invoices, creating transaction logs, managing pooling vs. custody account flows, and reconciling to the accounting systems/ERP.
- **Pay-ins** bring money into the business by accepting and adding different payment methods; orchestrating and optimizing payment routes and authorization rates; identifying cases of fraud and failure; allowing for rewards, gift cards, and payment splitting; and reconciling to the accounting systems/ERP.

For pay-ins, payment acceptance methods—credit cards, ewallets, crypto, gift cards, merchant acquirers, bank transfers—differ by an individual company's needs as well as by geography. Adding new methods can take weeks or months of engineering time to implement. For example, a consumer marketplace may want to accept a new local payment type or e-wallet as they expand into a new market or geography (e.g., Ideal in the Netherlands, or Grab's consumer wallet in Southeast Asia). This isn't as simple as flipping a switch in Stripe or Adyen, and it often requires a team of engineers to integrate a payment acquirer, embed it within the checkout flow, and tie it to the ERP.

For payouts, initiating a bank transfer can often mean a complex series of manual approvals sent via emails or CSV files to the bank and then tracked in spreadsheets or software developed 20 years ago (at best). The typical workflow for a fintech lender might include a finance team verifying a loan amount, batching all disbursements into a single daily request list, and then sharing that list with a bank via CSV upload or email/fax. Someone in finance then needs to check with the bank to verify that the payment went through, and manually log that payment disbursal into an internal spreadsheet.

For both pay-ins and payouts, transactions are often reconciled manually, with stale data and no single source of truth. While the ERP is a system of record for the company, it doesn't have—and isn't going to build—the capabilities to track real-time financial health. With very limited intelligence around payments, companies are losing out on revenue from failed payments and declines, but they don't know how to maximize success rates across processors or how they should configure their re-try strategies, when one fails, or to identify insufficient funds for ACH payments for a particular bank account. Nor can they benchmark performance to see how authorization rates compare against the market.

The opportunity to build payments software is deceptively large. This isn't just one company—we expect to see a number of companies build solutions, and potentially even systems of record, for payins and payouts as well as regional approaches across geographies.

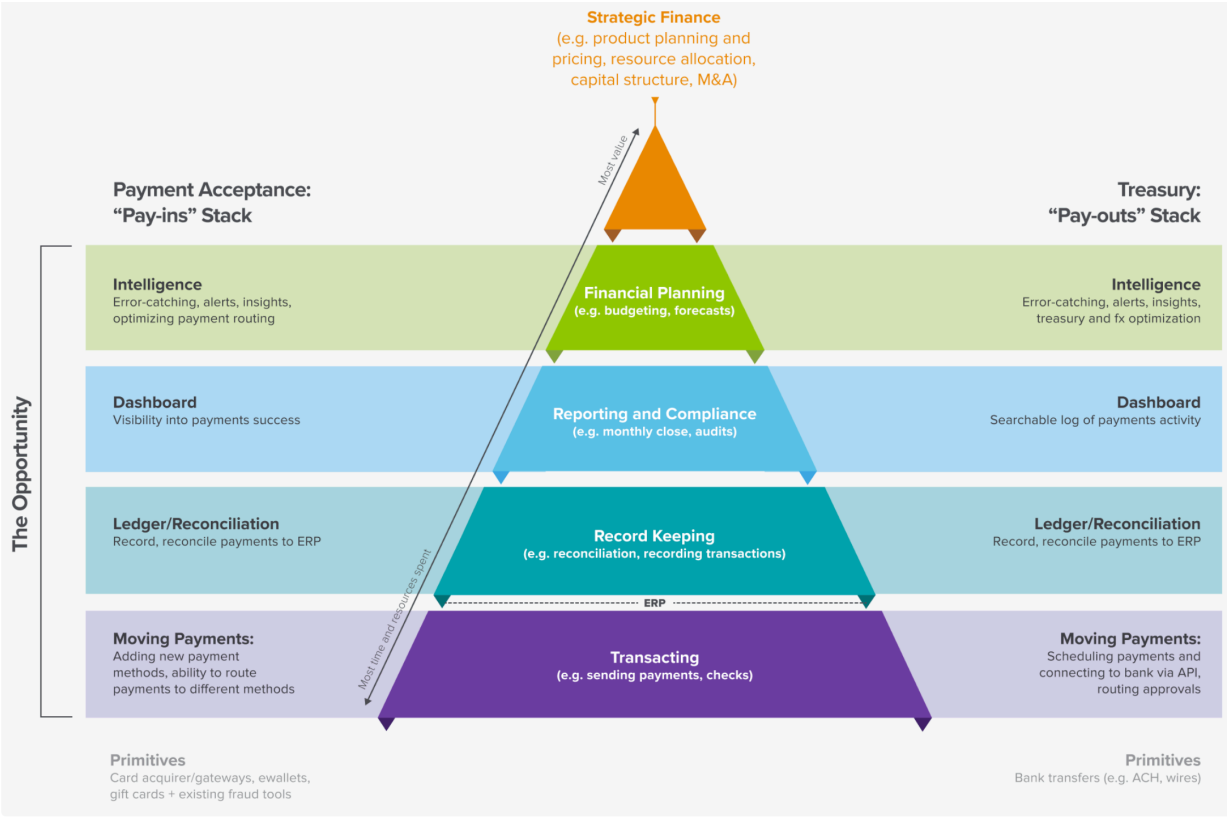
Fundamentally, the payin stack and the payout stack focus on different buyers and partners. The payouts stack connects to the company's banks, and the buyer is typically on the treasury side. On the payins side, the software needs to integrate into and maintain connections into a number of gateways, and the buyer is more likely from the product management side than the finance side.

For both payins and payouts, the current workflows around payments can be organized into layers on top of the existing "primitives," which are payment methods like bank transfers (e.g., wire, ACH), card payments, e-wallets, as well as payments-adjacent tools (e.g., fraud detection tools, compliance, authentication, tokenization) that influence whether payments proceed.

hese layers are:

- **Moving payments:** Enables programmatic movement of money
- **Ledger/reconciliation:** The recording of the payment entries and connection to the ERP
- **Dashboard:** Enables the ability see and search payments traffic (e.g. by country, processor/network/bank, customer type) and
- **Intelligence:** Offers the ability to set alerts and catch errors, and provide analytics and insights around performance, how to optimize payments, and maximize success rates. This is likely combined with better data (e.g. BIN lists) and benchmarking to understand how authorization rates compare.

Each layer builds on the layer below it, and the data generated and actions taken there. Together, they create the opportunity for new software to replace existing manual work.



The opportunity to build here also differs by geography. Different local payment methods, regional ERP systems, and banking systems and regulations surface different product requirements. In Latin America, fraud rates are high, therefore improving payment authentication (routing to the payment provider with the highest acceptance and accurate fraud risk assessment for a particular transaction) is most attractive. In Europe and Asia, managing a larger number of payment methods is core functionality.

If the technology is involved in the payments flow, the company likely needs a money transmitter license (which are also granted at the local and regional level and can take months to procure). This all translates into different products for Latin America, North America, Europe, Asia Pacific, Africa, etc.

Building trust and reducing friction in onboarding are critical to building successfully in this category. The product needs to be reliable given this is critical infrastructure (and trust can only be earned once). Relatedly, the product wedge also can't create a single point of failure for the customer and likely has to be additive on top of existing products to start, not a replacement of existing infrastructure.

The opportunity to build here also differs by geography. Different local payment methods, regional ERP systems, and banking systems and regulations surface different product requirements. In Latin America, fraud rates are high, therefore improving payment authentication (routing to the payment provider with the highest acceptance and accurate fraud risk assessment for a particular transaction) is most attractive. In Europe and Asia, managing a larger number of payment methods is core functionality.
If the technology is involved in the payments flow, the company likely needs a money transmitter license (which are also granted at the local and regional level and can take months to procure). This all translates into different products for Latin America, North America, Europe, Asia Pacific, Africa, etc.

To get to market, early design partners are critical, and lighthouse logos can signal trust to prospective customers. Product marketing is also critical to get right. This is not a set of problems that everyone across an organization is familiar with. Product marketing should be communicated in simple terms and as a revenue and business enabler (e.g., onboard more customers with a new payment method). And that then should be backed by a product that makes it simple for leadership across the company to understand, and get insights into the real-time financial health of the company, and potentially even replace the company's ERP.

The need for a new generation of FinOps software is only increasing as companies continue to expand globally and new payment methods are created. This infrastructure will be critical to understanding a company's real-time financial health and empowering the CFO (and their organization) to become more strategic. It's time to build new payments infrastructure—one that delivers a real-time system of record, can automate manual workflows, and deliver unique insights and organizational intelligence.

CFO Crisis in AI Era

https://a16z.com/the-cfo-in-crisis-mode-modern-times-call-for-new-tools/

Financial Opportunity of AI

https://a16z.com/financial-opportunity-of-ai/

In the background, there was also a profoundly impactful technological revolution called the spreadsheet. Released in 1979, VisiCalc was the first "killer app" of finance (on the Apple IIe!), and one of the things that allowed KKR and other early firms to model outcomes and make so much money. With this faster method of calculating, what might have taken weeks could now take seconds. Milken himself is said to have credited (or blamed) VisiCalc and spreadsheets for the growth of the Private Equity (PE) industry, since cash flows vs. debt payments could be easily monitored, and a formerly complex net present value calculation now just involved a formula for a cell. Early KKR executive Donald Herdrich is said to have bought an Apple IIe in 1980 for his children, got a demo of VisiCalc at the electronics store, and that turned into a decisive advantage for KKR going forward. Eventually, this all became table stakes, with every PE firm employing the same analytical tools and analytical minds to find potentially upgradable or fixable companies. PE is now a giant industry—from its humble beginnings with spreadsheets and junk debt to almost $5 trillion of assets.

Generative AI is likely going to usher in a far more profound method of company transformation. Instead of financial engineering and the **improved management techniques that PE promotes**, we'll start seeing AI cut costs and make existing companies vastly more profitable...while also enabling new business models to emerge.

It's important to recognize that while the impact that generative AI can have on "bits" is substantial—since generative AI can "manipulate" those bits very easily—we're probably too early to see a massive opportunity in "atoms" businesses. Lockheed Martin makes F-35s by assembling atoms, and has a 13% gross margin; Salesforce makes software by assembling bits, and has a 74% gross margin.

Known knowns are companies/products/ideas that exist today, and that have clear demand from customers. Can costs be reduced, customer support improved, NPS increased, ***new, previously unprofitable sales opportunities be unlocked? As technology continues to improve***, the answer is an unequivocal "yes." United Airlines can't simply hire and train 10,000 new call center reps in 12 hours when terrible weather shows up, but dynamic compute can solve this. What about complicated corner cases? **Delta has a bereavement policy for discounted fares**, but requires interaction with a representative, ostensibly to prevent abuse. Apple will sometimes escalate something to a "Level 2" technician if the front-line can't figure it out, but this can sometimes take hours or days. Verification, validation, obscure corner cases, automation of tedious tasks—all of these can be done by AI.

There are really three investable opportunities in the category of "known knowns":

- Sell software to incumbents
- Compete with incumbents, with generative AI at the core
- ***Buy incumbents and remake them with AI—what a "generative AI" KKR would be doing***

Think about Rocket Mortgage, with thousands of mortgage brokers and 2022 net revenues of $5.8 billion, against almost $2.8 billion of "salaries, commissions, and team member benefits." Somebody will likely start a company/build a product to either supercharge Rocket's existing workforce or replace more of their workforce with software. It's clear Rocket could pay a lot for it, as would other companies that compete with Rocket. Somebody will likely start a net new company that does mortgage origination and refinance with virtually no human interaction—think vertical integration of the prior software example. And lastly, somebody might even acquire Rocket Mortgage—a $20 billion company as it is today—if its average EBITDA margin of 40% from 2019-2022 could be changed to 60% or more through generative AI.

Known Knowns: Three *Obvious* Areas of Investment

"Sell Software to Rocket Mortgage"	"Compete with Rocket Mortgage"	"Take Rocket Mortgage Private"
SaaS for Incumbents	**New AI First Full-Stack Players**	**Private Equity Opportunity**
• AWS for bit manipulation • Identify companies where high percentage of COGS are sales & support • Replace current offshoring initiatives with full or partial automation	• Incumbents may not be able to fully adopt AI as a newco with AI at its core • Full-stack entrants may have lower long-term cost structures and be able to sustainably counter position vs. incumbents	• Motherlode of PE profits to be unleashed by buying companies and aggressively replacing bit manipulation opex/ COGS/SG&A with AI • Replaces and/or augments traditional model of cost-cutting and outsourcing

Known Unknowns: Known unknowns can be best illustrated with a simple Economics 101 Supply/Demand graph. There are some products where there's massive supply at a very high price point, and massive demand at a very low price point…but there is no intersection. The curves simply do not meet.

Upwork and Fiverr both have extensive marketplaces for custom images and artwork, but it seems Midjourney has demonstrably more revenue than both of their graphics categories combined. Why? Because $20/month unlocks a tremendous amount of demand that simply did not exist at $500/image. It's not always about cost—it's also about speed. A Midjourney image takes less than 30 seconds to create, unlocking demand that was literally impossible with a human artist as a bottleneck—no matter what the cost.

LVMH likely spends tens of millions of dollars a year fighting counterfeit goods, sending cease and desist letters, cooperating with law enforcement, etc. How many small Shopify merchants might want the exact same service? All of them! How many could spend $50M/year? None of them. How many might spend $1,000/year? Maybe all of them?

Getting Started with AI JavaScript Tech Stack

https://a16z.com/the-getting-started-with-ai-stack-for-javascript/

Who Owns Gen AI Platform

https://a16z.com/who-owns-the-generative-ai-platform/

Over the last year, we've met with dozens of startup founders and operators in large companies who deal directly with generative AI. We've observed that **infrastructure vendors** are likely the biggest winners in this market so far, capturing the majority of dollars flowing through the stack. **Application companies** are growing topline revenues very quickly but often struggle with retention, product differentiation, and gross margins. And most **model providers**, though responsible for the very existence of this market, haven't yet achieved large commercial scale.
In other words, the companies creating the most value — i.e. training generative AI models and applying them in new apps — haven't captured most of it. Predicting what will happen next is much harder. But we think the key thing to understand is which parts of the stack are truly differentiated and defensible. This will have a major impact on market structure ***(i.e. horizontal vs. vertical company development) and the drivers of long-term value (e.g. margins and retention).***
The stack can be divided into three layers:

- **Applications** that integrate generative AI models into a user-facing product, either running their own model pipelines ("end-to-end apps") or relying on a third-party API
- **Models** that power AI products, made available either as proprietary APIs or as open-source checkpoints (which, in turn, require a hosting solution)
- **Infrastructure** vendors (i.e. cloud platforms and hardware manufacturers) that run training and inference workloads for generative AI models

In prior technology cycles, the conventional wisdom was that to build a large, independent company, you must own the end-customer — whether that meant individual consumers or B2B buyers. It's tempting to believe that the biggest companies in generative AI will also be end-user applications. So far, it's not clear that's the case.

To be sure, the growth of generative AI applications has been staggering, propelled by sheer novelty and a plethora of use cases. In fact, we're aware of at least three product categories that have already exceeded $100 million of annualized revenue: image generation, copywriting, and code writing.

However, growth alone is not enough to build durable software companies. Critically, growth must be profitable — in the sense that users and customers, once they sign up, generate profits (high gross margins) and stick around for a long time (high retention). In the absence of strong technical differentiation, B2B and B2C apps drive long-term customer value through network effects, holding onto data, or building increasingly complex workflows.

In generative AI, those assumptions don't necessarily hold true. Across app companies we've spoken with, there's a wide range of gross margins — as high as 90% in a few cases but more often as low as 50-60%, driven largely by the cost of model inference. Top-of-funnel growth has been amazing, but it's unclear if current customer acquisition strategies will be scalable — we're already seeing paid acquisition efficacy and retention start to tail off. Many apps are also relatively undifferentiated, since they rely on similar underlying AI models and haven't discovered obvious network effects, or data/workflows, that are hard for competitors to duplicate.

Margins should improve as competition and efficiency in language models increases (more on this below). Retention should increase as AI tourists leave the market. And ***there's a strong argument to be made that vertically integrated apps have an advantage in driving differentiation.***

Looking ahead, some of the big questions facing generative AI app companies include:

- **Vertical integration ("model + app").** Consuming AI models as a service allows app developers to iterate quickly with a small team and swap model providers as technology advances. On the flip side, some devs argue that the product *is* the model, and that training from scratch is the only way to create defensibility — i.e. by continually re-training on proprietary product data. But it comes at the cost of much higher capital requirements and a less nimble product team.
- **Building features vs. apps.** Generative AI products take a number of different forms: desktop apps, mobile apps, Figma/Photoshop plugins, Chrome extensions, even Discord bots. It's easy to integrate AI products where users already work, since the UI is generally just a text box. Which of these will become standalone companies — and which will be absorbed by incumbents, like Microsoft or Google, already incorporating AI into their product lines?
- **Managing through the hype cycle.** It's not yet clear whether churn is inherent in the current batch of generative AI products, or if it's an artifact of an early market. Or if the surge of interest in generative AI will fall off as the hype subsides. These questions have important implications for app companies, including when to hit the gas pedal on fundraising; how aggressively to invest in customer acquisition; which user segments to prioritize; and when to declare product-market fit.

Perhaps the clearest takeaway for model providers, so far, is that commercialization is likely tied to hosting. Demand for proprietary APIs (e.g. from OpenAI) is growing rapidly. Hosting services for open-source models (e.g. Hugging Face and Replicate) are emerging as useful hubs to easily share and integrate models — and even have some indirect network effects between model producers and consumers. There's also a strong hypothesis that it's possible to monetize through fine-tuning and hosting agreements with enterprise customers.

On top of this, startups training their own models have raised billions of dollars in venture capital — the majority of which (up to 80-90% in early rounds) is typically also spent with the cloud providers. Many public tech companies spend hundreds of millions per year on model training, either with external cloud providers or directly with hardware manufacturers.

This is what we'd call, in technical terms, "a lot of money" — especially for a nascent market. Most of it is spent at the *Big 3* clouds: Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. These cloud providers collectively **spend more than** *$100 billion per year* in capex to ensure they have the most comprehensive, reliable, and cost-competitive platforms. In generative AI, in particular, they also benefit from supply constraints because they have preferential access to scarce hardware (e.g. Nvidia A100 and H100 GPUs).

Interestingly, though, we are starting to see credible competition emerge. Challengers like Oracle have made inroads with big capex expenditures and sales incentives. And a few startups, like Coreweave and Lambda Labs, have grown rapidly with solutions targeted specifically at large model developers. They compete on cost, availability, and personalized support. They also expose more granular resource abstractions (i.e. containers), while the large clouds offer only VM instances due to GPU virtualization limits.

Behind the scenes, running the vast majority of AI workloads, is perhaps the biggest winner in generative AI so far: Nvidia. The company **reported $3.8** *billion* of data center GPU revenue in the third quarter of its fiscal year 2023*,* including a meaningful portion for generative AI use cases. And they've built strong moats around this business via decades of investment in the GPU architecture, a robust software ecosystem, and deep usage in the academic community. **One recent analysis** found that Nvidia GPUs are cited in research papers 90 times more than the top AI chip startups combined.

Other hardware options do exist, including Google Tensor Processing Units (TPUs); AMD Instinct GPUs; AWS Inferentia and Trainium chips; and AI accelerators from startups like Cerebras, Sambanova, and Graphcore. Intel, late to the game, is also entering the market with their high-end Habana chips and Ponte Vecchio GPUs. But so far, few of these new chips have taken significant market share. The two exceptions to watch are Google, whose TPUs have gained traction in the Stable Diffusion community and in some large GCP deals, and TSMC, who is believed to manufacture *all* of the chips listed here, including Nvidia GPUs (Intel uses a mix of its own fabs and TSMC to make its chips).

Infrastructure is, in other words, a lucrative, durable, and seemingly defensible layer in the stack. The big questions to answer for infra companies include:

- **Holding onto stateless workloads.** Nvidia GPUs are the same wherever you rent them. Most AI workloads are stateless, in the sense that model inference does not require attached databases or storage (other than for the model weights themselves). This means that AI workloads may be more portable across clouds than traditional application workloads. How, in this context, can cloud providers create stickiness and prevent customers from jumping to the cheapest option?

- **Surviving the end of chip scarcity.** Pricing for cloud providers, and for Nvidia itself, has been supported by scarce supplies of the most desirable GPUs. One provider told us that the list price for A100s has actually *increased* since launch, which is highly unusual for compute hardware. When this supply constraint is eventually removed, through increased production and/or adoption of new hardware platforms, how will this impact cloud providers?
- **Can a challenger cloud break through?** We are strong believers that **vertical clouds** will take market share from the Big 3 with more specialized offerings. In AI so far, challengers have carved out meaningful traction through moderate technical differentiation and the support of Nvidia — for whom the incumbent cloud providers are both the biggest customers and emerging competitors. The long term question is, will this be enough to overcome the scale advantages of the Big 3?

There don't appear, today, to be any systemic moats in generative AI. As a first-order approximation, applications lack strong product differentiation because they use similar models; models face unclear long-term differentiation because they are trained on similar datasets with similar architectures; cloud providers lack deep technical differentiation because they run the same GPUs; and even the hardware companies manufacture their chips at the same fabs.

There are, of course, ***the standard moats: scale moats ("I have or can raise more money than you!"), supply-chain moats ("I have the GPUs, you don't!"), ecosystem moats ("Everyone uses my software already!"), algorithmic moats ("We're more clever than you!"), distribution moats ("I already have a sales team and more customers than you!") and data pipeline moats ("I've crawled more of the internet than you!"). But none of these moats tend to be durable over the long term.*** And it's too early to tell if strong, direct network effects are taking hold in any layer of the stack.

We also expect both horizontal and vertical companies to succeed, with the best approach dictated by end-markets and end-users. For example, if the primary differentiation in the end-product is the AI itself, it's likely that ***verticalization (i.e. tightly coupling the user-facing app to the home-grown model) will win out. Whereas if the AI is part of a larger, long-tail feature set, then it's more likely***

horizontalization will occur. Of course, we should also see the building of more traditional moats over time — and we may even see new types of moats take hold.
Thoughts:
It is clear that selling the picks and shovels for enterprise companies looking to transform their organizations for a AI forward world will lead to AI companies that focus on enabling enterprises will win BIG!

Reducto - AI Data Cleaning + Prep for Enterprise

🔗 https://www.youtube.com/watch?v=QBC_cViA7j8

https://reducto.ai/

https://reducto.ai/edit

Big Bank Fees Could Kill FinTech 2025

https://a16z.com/newsletter/big-bank-fees-could-kill-competition/

https://x.com/_MomentHQ

https://moment.com/

Untitled

Every Company Will be a Fintech Company

https://a16z.com/every-company-will-be-a-fintech-company/

Accounting Vertical in AI Era

https://a16z.com/newsletter/the-rise-of-vertical-ai-in-accounting/

AI can be used where BPOs(Business Process Outsourcing) are currently being used by the enterprise e.g.

Nearly every accounting firm, large or small, to whom we spoke brought up one specific growth vector: Client Advisory Services, or CAS. CAS comprises a mix of outsourced CFO and controller services, making it a particularly strategic department for three reasons.

CAS creates sticky, recurring relationships with clients that can serve as the basis for cross-selling engagements into other parts of the firm (namely tax and audit).
CAS revenue growth is outpacing broader accounting-firm revenue growth. Firms with a CAS practice are reporting 30% median revenue growth year over year, while the industry at large reports ~9% annual growth in net-client fees.
Qualitatively, CAS broadens the surface area for advisory revenue, which can often be more predictable and less seasonal than traditional accounting and tax engagements.
But with scaled CAS, comes great labor needs! Many of the activities that constitute outsourced controller services (e.g., helping with month-end closes, transaction reconciliation, expense management, etc.) require a person to repetitively perform rote tasks; we've previously referred to these "jobs to be done" as data collection and ingestion. Additionally, this manual extraction of data (from invoices, contracts, emails, general ledgers, and the like) is work that's usually completed by a junior CPA or by an offshore laborer. The promise made by AI for accounting is to cut time spent on these

activities down from hours to minutes, delivering immediate ROI in the form of freed-up labor hours.

While this sounds ideal, it's much harder to pull off than one might think, even with the powerful new technologies that underpin our current wave of AI-native accounting applications. For starters, accounting is a field in which accuracy is of paramount importance. For software to actually free up firms to either repurpose the skilled labor currently performing this work or end relationships with third-party BPOs, it has to actually work. Not only that, it needs to work "horizontally" across industries, many of which are riddled with unique data sources (industry-specific ERPs) and processes (how jobs get billed, how money flows, etc.).
It should come as no surprise, then, that CAS teams are already typically organized by vertical. Much like investment banks have coverage groups across specialties — healthcare, financial institutions, consumer retail, and so on — CAS has dedicated staff that understand the nuances of the customers they serve.

So what does this all mean for startups? While we are optimistic that as models improve and their associated costs continue to come down, the most prudent approach for technology companies hoping to serve CAS practices might be to really lean into specific verticals, around which they can build a brand and subject matter expertise. If they can demonstrate an immediate and clear ROI to one vertical in CAS with highly accurate results, they will deepen customer trust in their product. As a result, they will be able to dramatically shorten their sales cycles for firmwide expansion.

Example:
Construction Accounting - https://www.adaptive.build/

BPOs

Third-party Business Process Outsourcing (BPO) involves a company contracting out specific business operations, which are often non-core to its primary activities, to an external service provider.
Here's a breakdown:
Focus on Core Competencies: By outsourcing non-essential tasks, businesses can dedicate their resources and attention to the activities that are central to their mission and

objectives, enhancing efficiency, productivity, and innovation in core areas.
Cost Efficiency: Outsourcing can reduce operational costs, as businesses don't need to invest in developing and maintaining the infrastructure and technology required for these outsourced functions. They can also benefit from lower labor costs by engaging providers in other regions or countries with lower wages.
Access to Expertise: BPO providers often specialize in particular areas and possess specific expertise, technology, and advanced tools that the contracting company may not have in-house.
Scalability and Flexibility: Businesses can quickly scale up or down their outsourced services based on demand, which is especially beneficial for companies with fluctuating workloads or those planning for growth.
Examples of Outsourced Tasks: Common tasks outsourced include customer support, IT services, human resources (like payroll and benefits administration), accounting, marketing, and data entry.
Types of third-party BPO:
Front-Office BPO: Handles customer-facing tasks like customer service, technical support, and sales.
Back-Office BPO: Involves internal operations like payroll, accounting, data entry, and IT support.
Offshore BPO: Involves outsourcing to a company located in a different country, often chosen for lower costs and access to a wider talent pool.
Nearshore BPO: Outsourcing to a vendor in a neighboring country, often for similar time zones and cultural compatibility.
Onshore BPO: Outsourcing to a company within the same country, potentially in a different city or region, offering local expertise and easier collaboration.
Considerations when using a third-party BPO:
Data Security and Privacy: Sharing sensitive data with a third party requires ensuring robust data security measures and compliance with relevant regulations (like HIPAA or GDPR).
Communication Challenges: Time zone differences, language barriers, and cultural variations can affect communication and collaboration with offshore or nearshore providers.
Potential for Quality Issues: Maintaining quality standards and ensuring the provider consistently meets performance expectations is crucial.
Loss of Control: Businesses need to be comfortable relinquishing some control over the outsourced functions and trust the provider's processes.
The third-party BPO market is expanding and becoming increasingly important in a globally connected economy, according to Grand View Research.

Regulation in AI Era

https://a16z.com/newsletter/the-rise-of-vertical-ai-in-accounting/#regulation-becomes-code

Global Payments Infrastructure

https://a16z.com/global-payments-infrastructure/

The opportunities: global fintech infrastructure

Money **moves clumsily across borders**, but increased consumer and business demand for better experiences is attracting top entrepreneurs to solve these problems. There are many areas of opportunity here:

Creating multi-country rails

It continues to be challenging to move money across a *single* border. Businesses are often left waiting for days for a payment to go through, without knowing the exact foreign exchange fee. Moving money between multiple countries multiplies this problem. Several existing companies have already integrated with various local rails to help companies orchestrate money movement, but opportunity still exists for new players to create seamless and transparent money movement experiences between countries.

Building embeddable payment infrastructure

Increasingly, companies aim to monetize through financial services, but in many geographies, it is still difficult to find modern card-issuing partners or white label payment acceptance. Furthermore, global software companies are frequently compelled to partner with several different infrastructure providers to cover the necessary geographies, a complex process that requires maintaining multiple vendors.

Enabling borderless business banking

Businesses operating in multiple countries often open several bank accounts per country—correspondent banks to facilitate international transfers, local banks to take advantage of the best local banking services, additional investment or treasury accounts, and more. This process is slow, hinders cash visibility across the company, incurs high expenses when moving money (even within a single company), and complicates end-of-month reconciliation. The increasing prevalence of open banking in many countries offers new companies the opportunity to create an application layer that offers multi-country account visibility and other services.

Satisfying global compliance

Know Your Customer or Know Your Business (KYC/KYB) compliance in a single country is often complicated. It not only requires integrating the right data sources, but also creating a process that feels frictionless to the customer while satisfying all compliance requirements. Outside of customer onboarding, complying with local regulations around aspects such as data storage or reporting requirements can be challenging, especially when operating in multiple countries. Though there are best-in-class "**as a service**" companies that solve this problem in the U.S., globally these challenges remain unsolved, especially for businesses. There is potential to abstract this complexity with software.

Combatting fraud

The advent of real-time payments in many countries will solve some frustrating payment delays. However, this magnifies another problem: fraud. As generative AI tools become more widespread, the cost to fraudsters of iterating malicious content drops to near zero: they can write and test thousands of phishing attack emails in minutes and continuously tweak the ones that work best. In this new landscape, effective fraud solutions for cross-border payments will become increasingly important.

For all these reasons, there's a huge opportunity to develop software that makes moving money across borders seamless and transparent—and for making those services trusted and secure.

Market Research in AI Era

https://a16z.com/ai-market-research/

The Global Finance Stack

https://a16z.com/the-company-of-the-future-is-default-global/

Selling AI Product Around the World is Still Too Hard

https://a16z.com/more-countries-more-problems-selling-ai-products-around-the-world-is-still-too-hard/

Existing e-commerce platforms focus almost exclusively on physical products and usually serve customers selling in one primary geography. They lack the complete feature set these digital-first products require:

Global vs. single geography: Selling in multiple geographies dramatically increases complexity and can require setting up merchant accounts in multiple geographies, pay-in and pay-out orchestration in local currencies, tax remittance, and more.
Digital vs. physical products: Selling digital products requires features that physical products do not including metered billing, subscriptions, license key management, and retention and upselling tools.
Thus, similar to how Shopify and Stripe enabled the D2C, physical e-commerce product wave, we believe there is an opportunity for new platform companies to emerge and empower this wave of digital-first, AI-enabled, default global companies.

The landscape today

There are currently two primary paths for entrepreneurs to sell their digital-first products globally:

Stitch together best-in-class point solutions for maximum flexibility from a technical and pricing perspective

Use a Merchant of Record that abstracts away the complexity of stitching together tools with an all-in-one platform

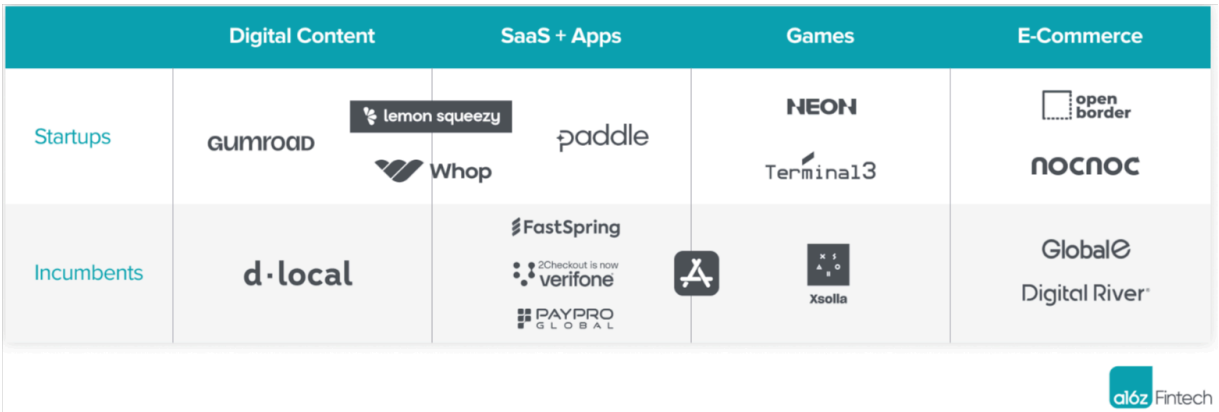
To most effectively sell software/digital products globally, founders will require:

- Pay-In + Payout + Orchestration: To accept payments from customers around the world in their preferred (often local) payment method. This can require setting up merchant accounts in different countries, plugging into payment gateways, and orchestrating between them.
- Metered Billing & Subscriptions: To maximize revenue potential by charging customers based on their usage via metered/usage-based billing platforms and managing customer subscriptions.
- Global Tax Calculation + Remittance: To calculate, file, and remit sales tax in jurisdictions where customers are located and where threshold amounts have been reached (i.e. when revenue exceeds a specific dollar amount in a given geography a company is selling in) in compliance with tax laws.
- Fraud Prevention: To prevent fraud globally, by either plugging into best-in-class point solutions or fraud orchestration platforms, without increasing customer friction.
- Marketing + Retention + Support: To attract, retain, and upsell customers from around the world, while providing product and payment support to minimize churn.

While larger companies can build processes internally for this, for new startups, the integration, coordination, and upkeep of a variety of vendors can take developers away from their highest value use of time: building their product.

Enter the Merchant of Record

A Merchant of Record (MOR) is an entity that accepts payments on behalf of a business such as the Apple App Store for mobile apps, as we have written about previously. The MOR is authorized and held liable for these transactions, which includes processing payments, managing payment processor fees, managing refunds and chargebacks, providing billing-related customer support, and ensuring businesses remain compliant with global tax regulations. The Merchant of Record gains a direct, standalone relationship with the end customer–the MOR's name is what a customer sees on their bank statement for instance.



Beyond making the selling experience for merchants simpler, MOR platforms also have an interesting opportunity to create network effects. The MOR controls the final customer experience at the point of checkout, and thus the MOR can cross-sell products from other merchants on its platform. For example, Digital River was successful in creating network effects in the '90s, not only by cross-selling SKUs from other merchants at the point of checkout, but also a broader set of financial services products (e.g., insurance for electronics).

Potential for developer-first distribution

Unlike in the physical-product wave, this digital-product wave is mostly being led by technical founders—opening the opportunity for a developer-first platform (i.e., APIs with great documentation) and community-driven developer distribution. We believe there is an opportunity for point solutions or MORs to become the standard for a "Selling" function in the emerging **javascript stack to build AI-based applications**. (Github repository **here**!)

RIP to RPA in AI Era

https://a16z.com/rip-to-rpa-the-rise-of-intelligent-automation/

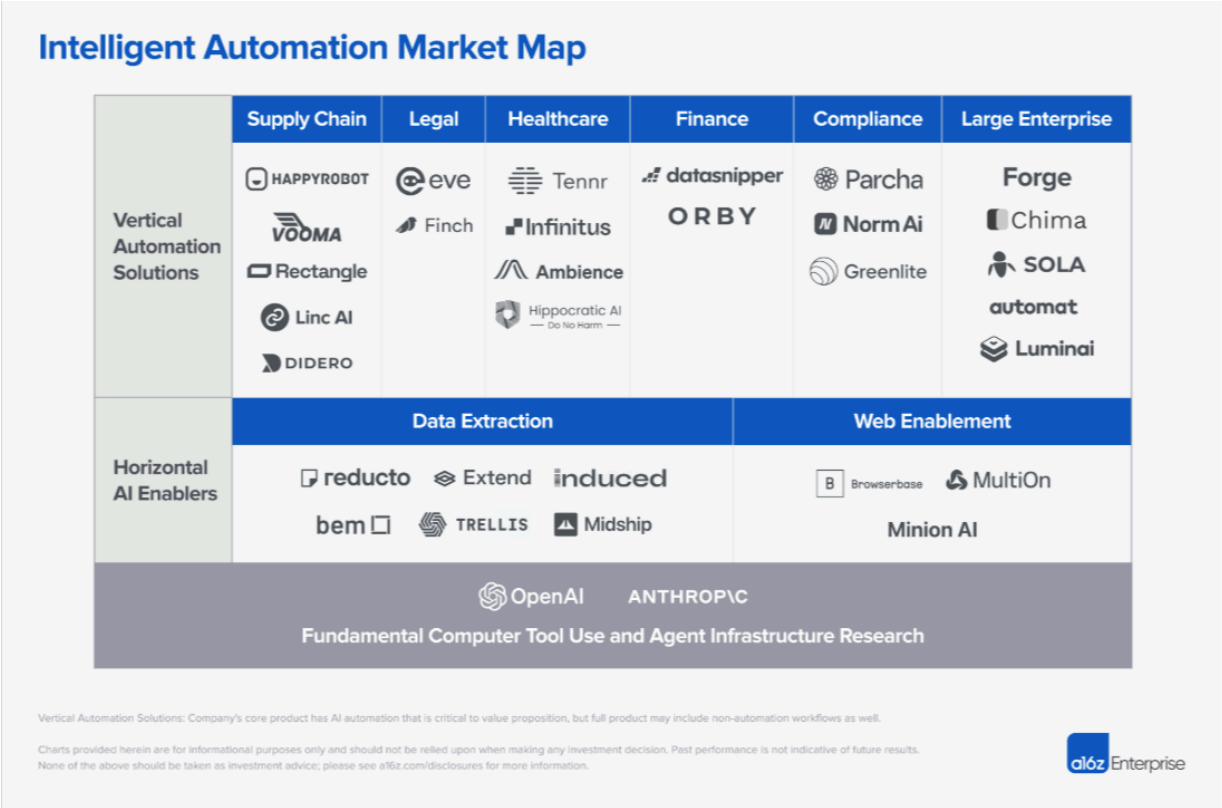
Bureau of Labor and Statistics(BLS) for industries + domains that are large but have been underserved due to lack on innovation of companies serving these industries or they didn't have Fortune 500 budgets.

AI Voice Agents to capture data/business and revenue opportunities.
"The potential market is enormous. For all the work that current software can handle, there are orders of magnitude more work that it cannot: work that is being done via pen and paper, spreadsheets, phone calls and fax. Intelligent automation can address the current

labor costs associated with this work — comprising over 8 million operations / information clerk roles according to the Bureau of Labor Statistics — as well as the spend associated with outsourcing this work, representing a meaningful portion of the $250 billion business process outsourcing (BPO) market."

"Startups largely have a greenfield opportunity in this space. There is often no existing software product for these workflows given their bespoke nature: the people were the product. As a result, these roles never developed "systems of record" in the way other roles did (e.g., Salesforce for sales, Workday for HR), meaning there is no software incumbent to "add AI" into their existing product suite. This market is wide open for startups."

"Specifically, we view the market opportunity as focused on two main areas: horizontal AI enablers that execute a specific function for a broad range of industries, and vertical automation solutions that build end-to-end workflows tailored to specific industries."



Horizontal AI enablers

Today, every intelligent automation company is building a similar set of capabilities and internal tooling. This creates a perfect opportunity for startups to simplify the process by focusing on one, specific foundational component..

For example: almost every intelligent automation company has to parse unstructured data and output contextualized, structured data. Many companies have built this out internally, and companies like Reducto and Extend are working to be the horizontal enabler to solve this specific need.

We think there are many other core building blocks needed for complex intelligent automation — including but not limited to building web data crawlers, structuring data from unstructured sources, or writing data back to legacy systems.

End-to-end vertical automation

We've previously written about our excitement for investing in vertical software (software that sells to one particular industry). We think this is a particularly good fit for intelligent automation, since operational agents will need to have the narrower context and deep integrations to achieve the accuracy and consistency customers expect.

Every industry has back office operations that could be automated, and we've already seen startups use LLMs to automate one flow as a strategic wedge to build deeper for specific industry needs.

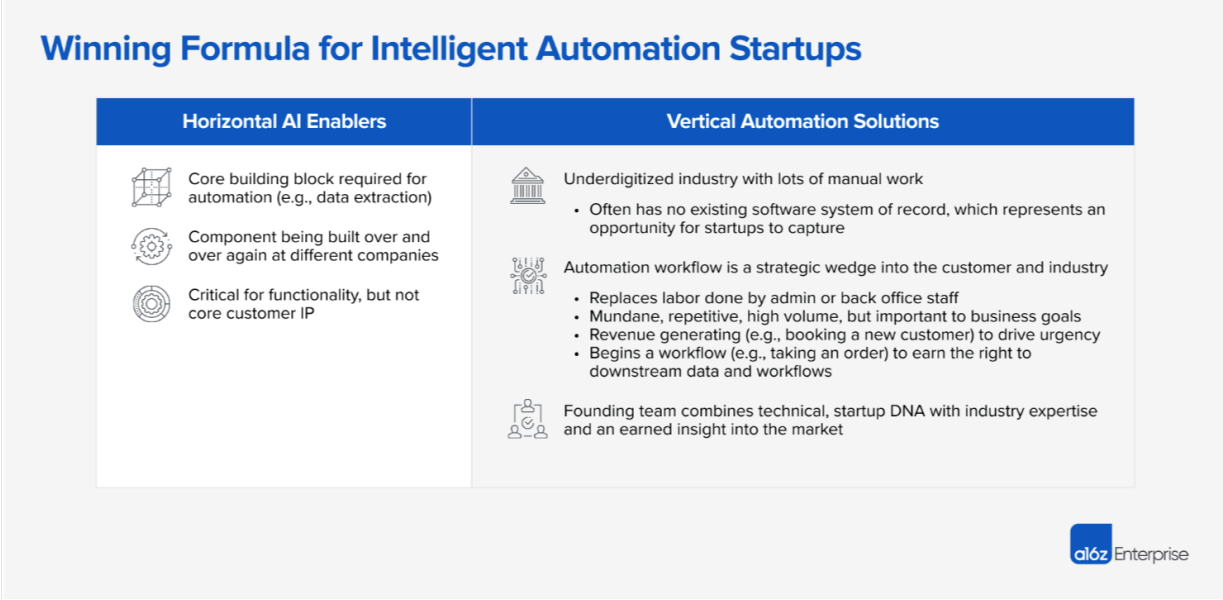
Examples:

In healthcare, for example, Tennr has automated the referral management flow. Referrals are the lifeblood of any growing healthcare practice, but accepting a referral used to require a lot of manual labor (e.g., receiving a fax, having the front desk pull the information from the fax, and manually inputting that patient information into their system). Tennr has built intelligent automation to solve this information transfer problem – using LLMs to extract unstructured data from PDFs and faxes, run validations on the information, and then write that information back into the system of record (EHR) automatically. This dramatically reduces the time it takes to accept a referral, which allows customers to secure new business more quickly.

In logistics, trucking brokers spend an enormous amount of time processing inbound orders and tracking loads. Now, using intelligent automation, companies like Happyrobot can automatically check on load status and updates via AI-powered voice assistants, and companies like Vooma are able to ingest unstructured email data to automate price quoting and order entry into the trucking management system (TMS).

These companies often focus on automating a very narrow, but very common and important workflow in their respective industries, often involving data and information transfer. They do not seek to be the "system of record" — at least initially — and can thus bypass the difficult rip-and-replace problems of going after legacy systems. They also start by automating revenue-generating workflows making themselves top priorities for their customers. And because these automations start at the beginning of a workflow, these startups earn the right to the upfront data and downstream workflows.

We believe this approach is a winning formula for intelligent automation startups, and we're eager to partner with those going after this opportunity across different industries.



What "Working" Means in the Era of AI Apps

https://a16z.com/revenue-benchmarks-ai-apps/

Pre-AI era a common benchmark for a best-in-class enterprise startup is $1M ARR in 12 months.

"Given the rapid growth both AI-native B2B and B2C companies are achieving between Seed and Series A, startups looking to raise venture capital need a strong velocity story. If not yet in live commercial traction, then certainly in shipping speed and product iteration. Speed is becoming a moat."

"It's not just about revenue — other metrics still matter. When evaluating companies at the Series A stage, we often have no more than 12 months of usage and retention data. Later-stage financing rounds will likely rely more heavily on traditional software metrics; rapid top-line growth won't be enough to compensate for low engagement or high churn"
*"Somewhat surprisingly, the revenue benchmarks for B2C are outpacing those for B2B. This is partially because consumer companies have a different "shape" now. One-third of the consumer companies in our sample raised significant funding to train their own models — and many see a massive revenue jump following new model releases. These spikes often resemble step function growth, which can later plateau until the next release. While conversion to paid may be lower for generative AI B2C businesses compared to their pre-AI counterparts, **our data** suggests that once users do convert, they retain just as well. "*

The Smartest Consumer Apps Now Cost $200 a Month

https://a16z.com/narrow-startups/

The math is compelling. Traditional ad-supported apps need hundreds of millions of users to build real businesses. But at $200 per month, you only need 41,000 customers to build a $100 million RR company
This changes everything about building consumer software. No more growth at all costs. No more engagement hacking. No more selling your users' attention to advertisers. Just go insanely deep on something a very specific group of people loves or needs, and charge accordingly.

This should also be a much more satisfying experience for builders. The winners will go the deepest on product instead of being the best at growth marketing. In an era of

abundance, specificity wins.

My prediction is that the mass market consumer startup will be increasingly rare. The future belongs to Narrow Startups that go deep, not wide. And, what the smartest apps charge today is what every valuable app will charge in three years.

5 Principles for Product Manager in the AI Era

https://a16z.com/stay-relevant-in-ai/

To find the value in the unexpected, PMs also need to become skilled at writing evals: structured tests that help you see where your model performs well and where it's falling short. Evals aren't just about measuring accuracy, they're identifying and assessing emergent capabilities to inform your product design.
Network effects remain the gold standard of software moats. However in the competitive landscape of AI, where the volume of products being spun up is so enormous, many of the traditional frameworks for establishing moats may not apply. For example, systems of record can now be indexed via vision models + RPA, potentially rendering the strength of this moat less powerful. As builders gain access to the same models and infrastructure, "soft" moats — like mindshare and momentum — that once seemed too weak to sustain a competitive advantage are becoming increasingly important in consumer AI. A playbook we're seeing more of: first and fast. Leading founders are the first to build a product, then stay at the front of the pack by continually shipping new features and capabilities.

You can't productize a system you don't understand. That means it's not enough to dabble in ChatGPT, you need to understand the difference between a language model and a reasoning model. Have you tried Deep Research, Operator, Gemini Flash, custom GPTs, and GPT-4o in multimodal mode? Have you read up on chain of thought, or observed it when using DeepSeek or any of the other reasoning models that expose it? The single most important intuition-builder for PMs is **reflexively using** AI products every day, in

every part of their job. This view is quickly tipping into consensus, as the CEOs of **Shopify**, **Duolingo**, **Box** and many more declare their companies pivoting to AI-first in all efforts.

From Demos to Deals: Insights for Building in Enterprise AI

https://a16z.com/insights-for-enterprise-ai-builders/

"AI companies have become highly sophisticated at both maximizing the capabilities of state-of-the-art models and constraining them for enterprise-grade reliability. They rigorously run evals on the latest models, orchestrate sequences of actions across different models, build substantial scaffolding on top of the base models, and set a product roadmap that threads the needle between what is possible today and what will be possible tomorrow. Teams often toggle between models based on how well they perform against a specific task, and have to make trade-offs based on model quality, cost, speed and scalability. In many cases they also fine-tune their own smaller models that live alongside the larger models in production. The result is a robust product experience that no single API call can deliver."

"In addition, for any AI product to be valuable, it needs to understand the context and logic of the business it's serving. Models do not do that out-of-the-box. As a result, AI companies are investing meaningful engineering and implementation resources into getting their products to work within the unique policies, culture, and systems of each individual customer instance. It's gritty, in-the-weeds customer work that needs to get done, but that the horizontal model companies don't and likely won't do themselves."

In the fiercely competitive AI market, companies still need to be building enduring products. Moats still matter. They can do so in a few ways.

Become the system of record: The most classic moat in enterprise software is becoming an organization's core system of record, or the source of truth for critical data. AI has opened up exciting new opportunities in different markets (especially vertical markets), because it's a great way to get customer velocity and demonstrate clear value in a short timeframe. But ultimately the system of record is still a dominant business model to ensure enduring value.

That doesn't mean it's not worth pursuing AI wedges. Several AI companies, such as Eve, Salient, and Toma are using AI wedges to capture data at the point of creation (through voice calls or ingesting unstructured data) in ways that traditional software

couldn't. They are then building downstream workflows from that point of creation, with the ultimate goal of maturing into the core system of record within their industries.

Create workflow lock-in: Once a company is able to get users to embed a product into their daily workflows, it creates operational and behavioral muscle memory that makes switching psychologically and culturally disruptive. While it's common to say that AI software is doing the work autonomously and therefore lacks classic GUI-based user interfaces, there's still a heavy human-to-AI interaction loop, as humans are still often overseeing and auditing the work being executed by AI today. Decagon, for example, builds AI agents that autonomously deflects support tickets, but still has robust product workflows for humans to monitor, tweak, or analyze the work these AI agents are doing, and clever ways to escalate to human agents when needed. Getting users comfortable with their product's UI/UX has created powerful workflow moats, making customers unlikely to switch to new tools.

Build deep vertical integrations: Many enterprise customers live in and depend on a complicated web of business software systems, many of which have limited APIs and don't speak to each other. As a result, AI companies hoping to serve these enterprises must often deeply connect and integrate into these bespoke systems to drive value. Investing in these integrations as a first-class effort helps build durability in the customer base by embedding the product into a customer's core operational workflow, making it difficult to rip out without disrupting other systems.

In healthcare, for instance, Tennr has invested heavily into integrating with a long tail of legacy fax and healthcare systems to streamline pre-patient referrals across providers. In logistics, HappyRobot builds connections into homegrown TMS (trucking management system) platforms to deploy its AI voice agents for freight call operations. And Glean drives much of its value from critical integrations across core enterprise tools. These deep, often customer-specific integrations are key to making AI solutions work in the real world and can serve as powerful competitive moats.

Entrench customer relationships: Enterprise buyers are still human. Trusted relationships often matter more than any specific feature or vendor price, especially as AI vendors increasingly act as strategic AI thought partners to the customers they serve. Many AI companies now have the buyer's ear in ways traditional software vendors rarely did, helping shape customer roadmaps and AI strategy, not just tool purchases.

Shift to AI Platform Era: Lessons from the Cloud Era

https://a16z.com/cloud-lessons-for-the-ai-era/

"From mainframes to PCs, desktop to mobile, unnetworked to internet, and on-prem to cloud and SaaS, technical progress in software tends to follow Schumpteter's "creative destruction", new winners emerge with new eras."

Across each transition from on-prem to SaaS to Cloud the B2B software companies revenue has grown 5.9x over the past 25 years. From $99B to $589B in 2023.

AI's value is akin to the value the internet created i.e. opening entirely new ways of doing things, not just better ways of doing old things.

Incumbents that are weakest are the companies that have infrastructure that requires significant investment in current system architecture in order to provide AI enabled capabilities.

"For SaaS incumbents, the real winners at the application layer are likely to be those who can figure out how to evolve a system of record into a system of prediction and eventually, execution. Enterprise SaaS CRMs beat on-prem incumbents because they were more useful systems of records, with features that made it possible to better track and forecast sales. Now as incumbents, these SaaS era sales tools have to compete with AI sales intelligence products that, for instance, already have the ability to tell sales reps, "this is the next most promising customer to talk to, and here's what to say" or even replicate a convincing phone conversation directly with customers"

When a platform shifts a new infrastructure layer emerges.

"Where the value of AI accrues in the long run is ultimately a question of defensibility. In the SaaS era, the biggest sources of defensibility were typically hard-to-copy technology that won developer mindshare (e.g., Databricks), platform systems of record that served as

the foundation for enterprise workflows and downstream applications (e.g., Salesforce), network effects that were embedded directly in the product experience (e.g., Slack), and go-to-market dominance unlocking a flywheel of customer feedback that informed product expansion (e.g., Workday)"

AI Era Defensibility

Data Corpus:

Commercial Lock-in - commercial agreements that guarantee exclusive data access e.g. KoBold Metals
Data Lifecycle Control - exclusive control of data lifecycle, often by controlling data generated by an underlying product e.g. Flock Safety
Data Scale Effect - sufficiently big data corpus discourages new entrants from replicating the same corpus e.g.Waymo
Regulated Data - regulatory frameworks and government procurement processes limit access to sensitive data e.g. Anduril

Product + Technology:

Product Network Effect - More users -> better experiences -> more users, data scale effect often the causal mechanism e.g. Character.ai
Internally-developed Foundation Models- model performance provides improved fidelity and accuracy e.g. Midjourney
Legacy Product Workflow Depth - natural language interface exposes legacy product capabilities e.g. ClickUp

Go-To-Market:

Legacy Customer Base - go-to-market head start via customer base, often quickly adopting a thin GPT-4 wrapper e.g. GitHub
Security Posture & Privacy Concerns - enterprise buyers emphasize data security and AI privacy for highly sensitive data corpus e.g. Anthropic

"AI will completely reinvent the workflow andUI for software, as increasingly AI software acts as a system of prediction and execution, not just a system of record. The speed with which we already see this shift happening means the nimble companies who can attract AI talent and move fast with some distribution are positioned to win"

"A paper back in 2019 found that consumers value "free" products in shockingly big dollar terms, estimating a willingness to pay as high as $17.5K for search engines, $8.4K for email, and $1.2K for streaming services. Given all the interesting consumer applications already emerging—virtual therapists accessible 24/7, physicians informed by the entire corpus of medical knowledge, automated secretarial assistants that deal with all the monotonous details of daily life—the scale of consumer surplus to come from AI is sure to spur a wave of incredible innovation."

Selling is Hard Right Now. Here's How to Win Business in the Gen AI Era

https://a16z.com/selling-winning-new-business-genai/

Notes:

All Companies across the board use Gen AI for the following business needs:

1. Productivity and information summary
2. Customer Support
3. Software Development

Immediate Term Impact: Companies are building Gen AI into customer-facing applications and work with LLMs to directly to build their own internal capabilities, like enhancing data analysis or building a recommendation engine to improve their products. These companies often build their own custom LLM solutions and need help implementing and scaling the solution. Due to the Gen AI threat to their business they're less focused on ROI in the near term for these investments.

Medium Term Impact - Levering third party tool or building internal tools to improve productivity or customer facing experience e.g. website chatbot

Long-term Impact - Companies using Gen AI to deliver measurable ROI and less concerned with developing an innovation advantage. Some of the early Gen AI apps in this segment

fall under sales and marketing use cases like creating marketing copy or personalizing emails.

Percentage of Respondents with LLM Use Cases in Production:

Info Search/Summary: 64-67% e.g. Contract Review, Enterprise Search, Text Summarization
Customer Support: 48%-50% e.g. Customer Service, Chatbot
Software Development: 25-28% e.g. Coding copilot
Data Science: 44%-58% e.g. Data Labeling, Data Analysis, NLP to SQL
Recommendation Algorithm: 6%-25%

"There's a massive opportunity across all buckets at the Gen AI app layer. Most of the GEN AI software we're seeing today is the same "app-building" or infrastructure software pitched from different angles, likely because we just have Gen AI primitives right now. In the long term however theres a big opportunity to displace entrenched enterprise software, both in horizontal applications(like CRMs) and vertical applications(like electronic health records). If you're building these solutions, consider what you're doing to win business at the app layer, whether that's by hand-holding customers through implementation, organizing their data, integrating with other apps, having customer who are familiar with your workflow, or building/nurturing a user community."

Thoughts:
Focus on opportunities where the cost savings in operations has high ROI. Finding opportunities where the margins are wide in the current solution but can be optimized or automated with AI. This is where you'll find your revenue. Their Cost Savings is Your Revenue!

Momentum as a Moat for AI

https://a16z.com/momentum-as-ai-moat/

Fintech in Colombia

https://a16z.com/global-payments-colombia/

Latin America's FinTech Boom

https://a16z.com/latin-americas-fintech-boom/

A16z Enterprise AI Buying 2025

https://a16z.com/ai-enterprise-2025/

How Will My Agent Pay for Things

https://a16z.com/newsletter/agent-payments-stack/

Export Your Brain with AI

https://a16z.com/export-your-brain-with-ai/

From Prompt to Product

https://a16z.com/ai-web-app-builders/