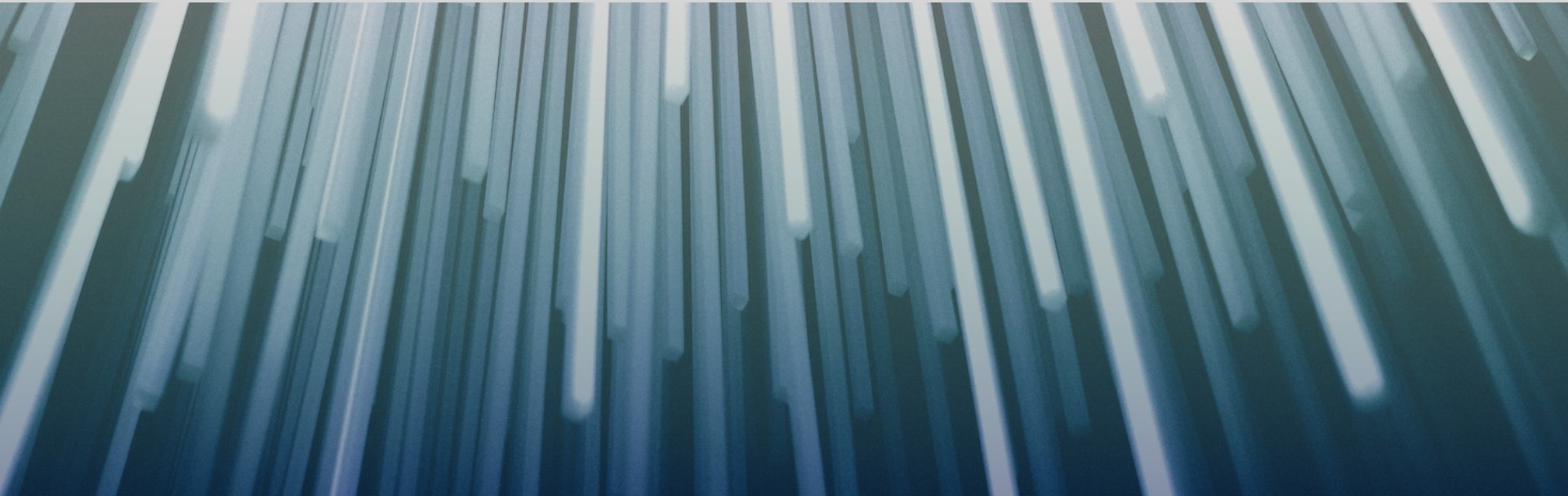
quantiphi
Solving What Matters

NVIDIA Elite Partner

March 19, 2024

Architecting Enterprise AI success with RAG and LLMs

Lessons from the First 12 Months of Building Generative AI Solutions



quantiphi
Solving What Matters

NVIDIA Elite Partner

AGENDA

01 Generative AI - A Phase Transition in Computing

02 Enterprise Application Patterns

03 Retrieval Augmented Generation

04 Case Studies

05 Lessons Learnt Architecting LLM Deployments

01

Busy executive’s guide to LLMs

How did we get here?

quantiphi
Solving What Matters

GENERATIVE AI - A PHASE TRANSITION IN COMPUTING

- Jensen Huang, NVIDIA

Generative AI Renaissance

OpenAI GPT-4 · PaLM 2 · cohere · ANTHROPIC · Jasper · Expedia · Rephrase.ai · Stable Diffusion · MISTRAL AI_ · BLOOM · synthesia · Adobe Firefly · character.ai · Claude 2 · Midjourney · DALL·E · Bloomberg GPT · Bard · Khan Academy · GitHub Copilot · copy.ai

The Mobile Application Gold Rush

DOORDASH · UBER · airbnb · tinder

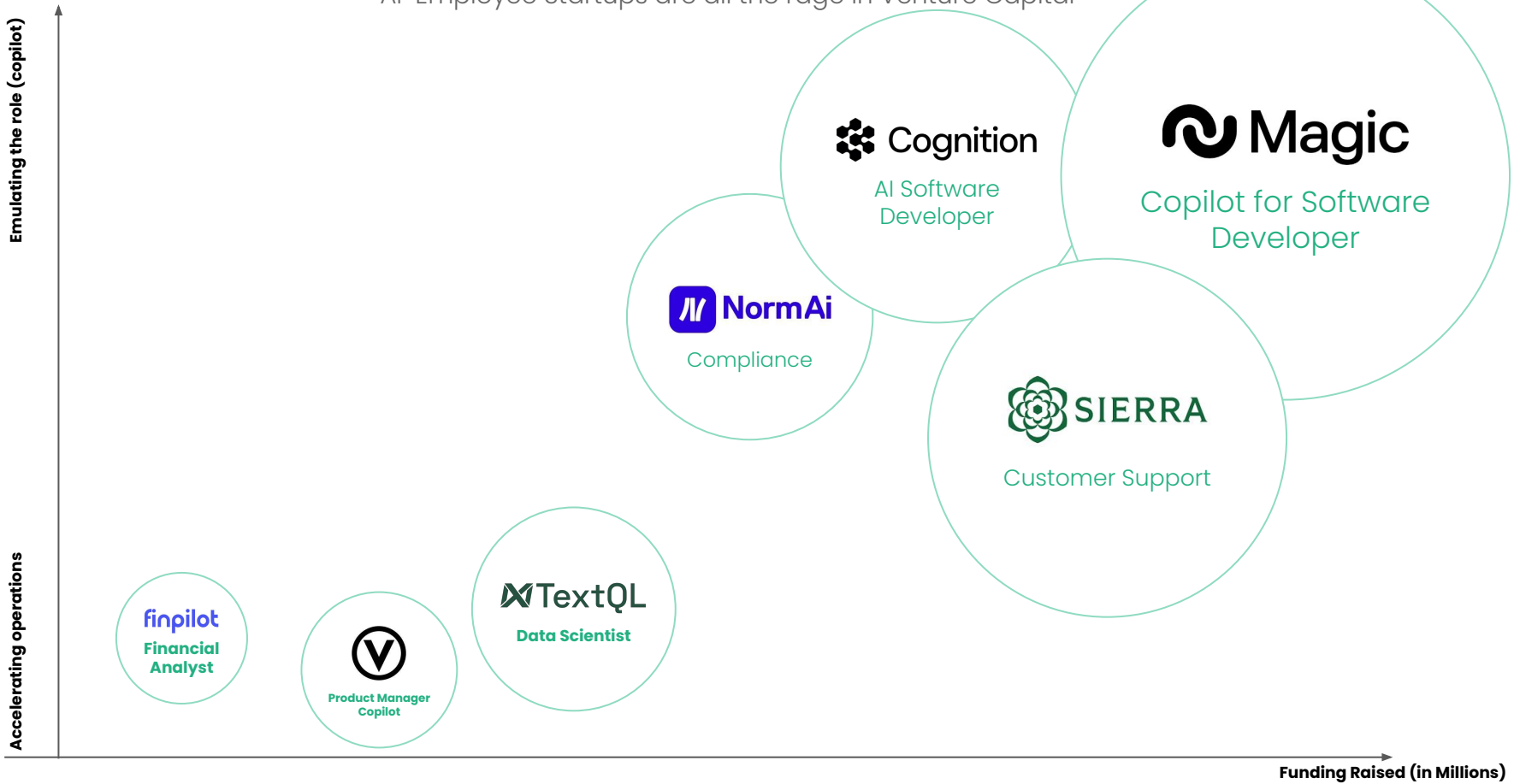
Foundation of the Digital Revolution

amazon.com · Google · Microsoft Hotmail · facebook · YouTube · Y! · ebay · napster

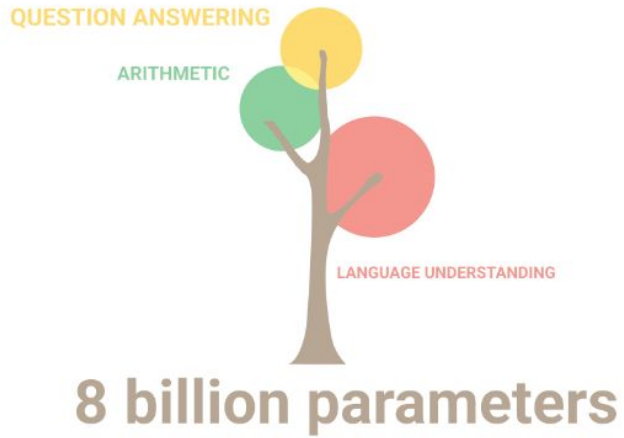
quantiphi
Solving What Matters

VALUE CREATION WITH NEW APPLICATIONS

AI-Employee Startups are all the rage in Venture Capital

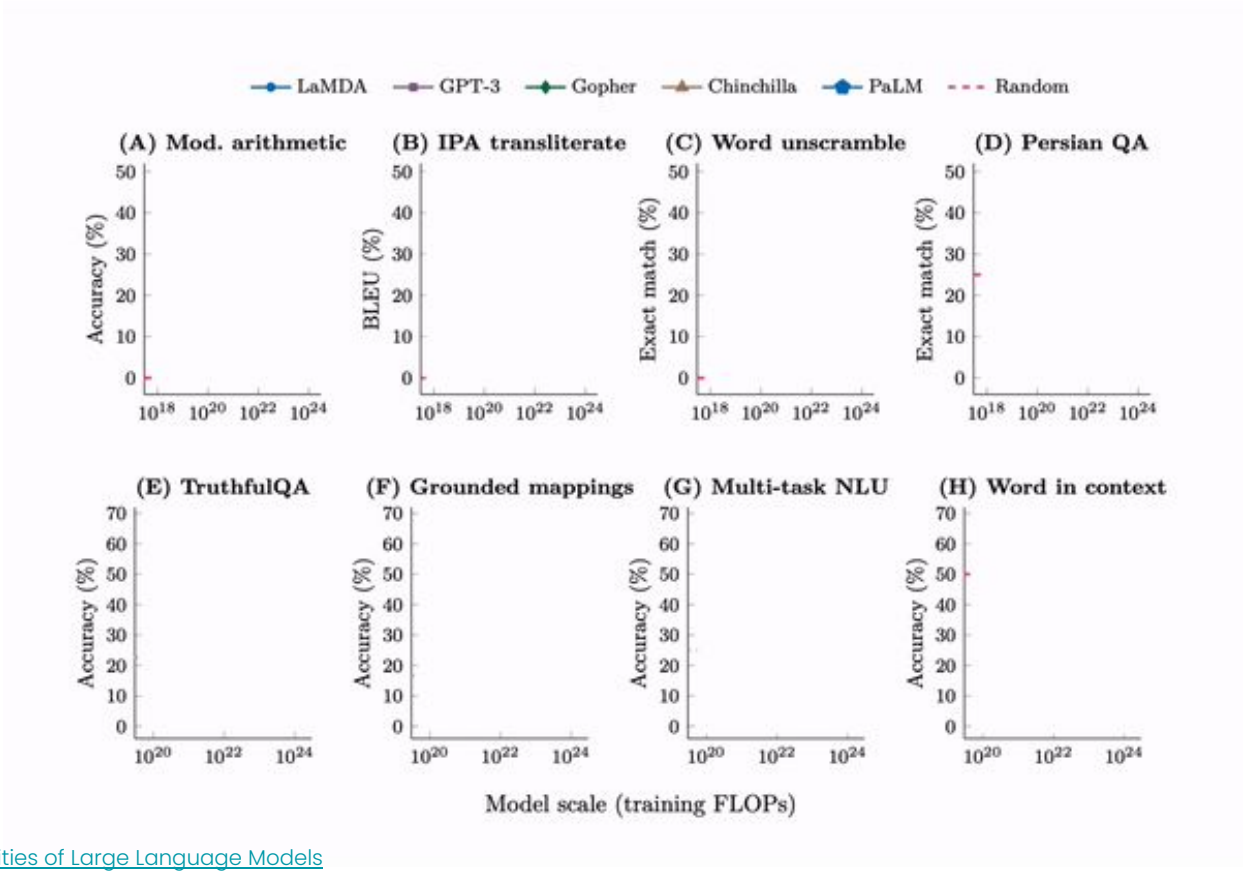


EMERGENT CAPABILITIES + IN-CONTEXT LEARNING IS … MARVELOUS



Reference: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance

EMERGENT CAPABILITIES + IN-CONTEXT LEARNING IS … MARVELOUS



Reference: [Emergent Abilities of Large Language Models]

02

Enterprise Applications

How do Enterprises want to use Generative AI?

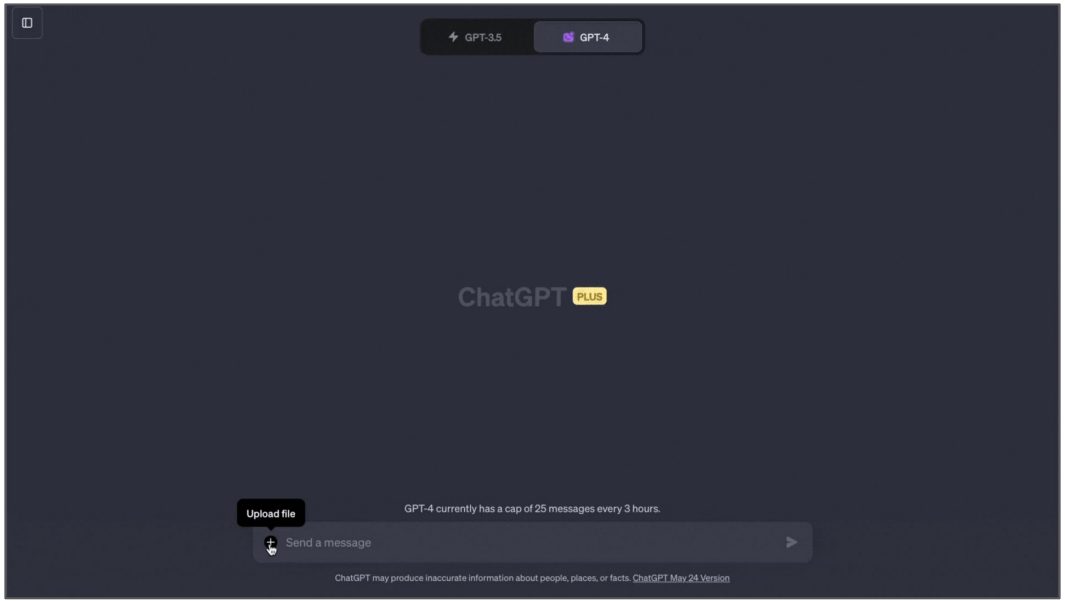
We conducted 300 workshops

This is what we have learnt

We conducted 300 workshops

This is what we have learnt

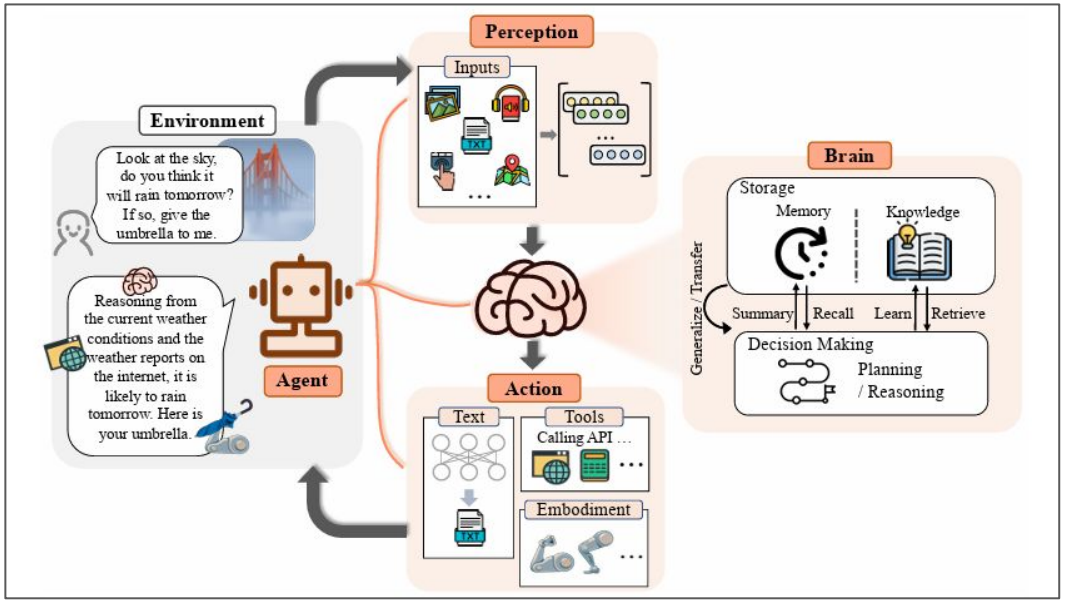
1. Help me talk to my data



We conducted 300 workshops

This is what we have learnt

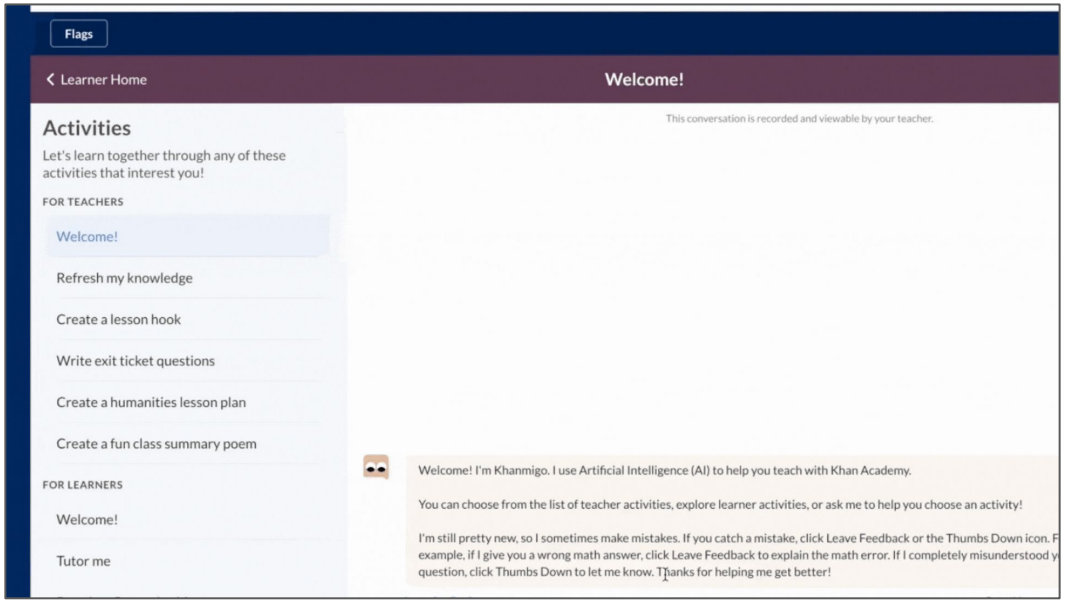
1. Help me talk to my data
2. Help me Automate and Plan Tasks



We conducted 300 workshops

This is what we have learnt

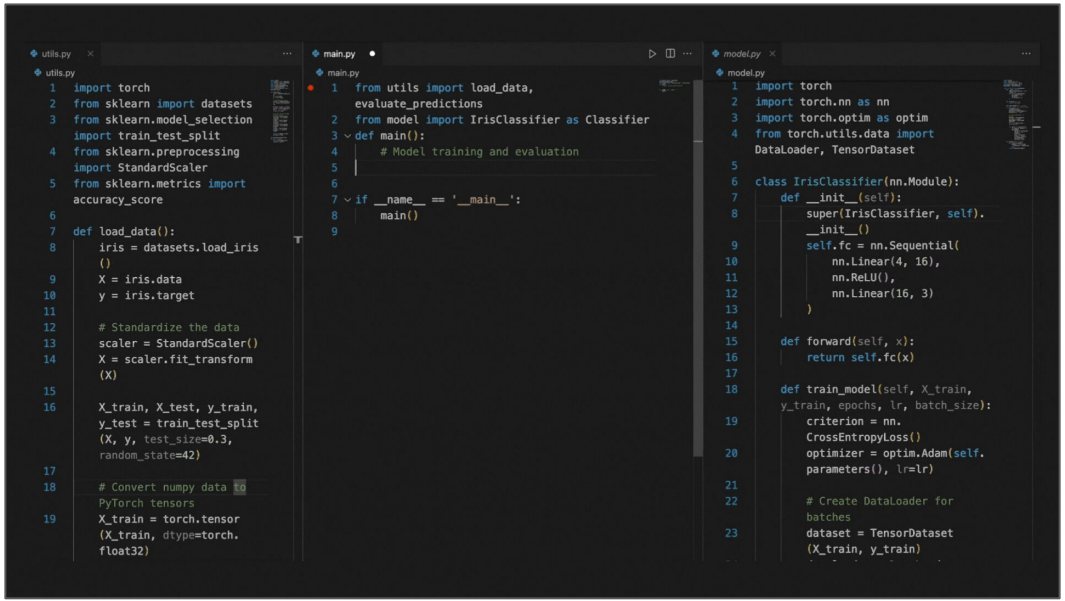
1. Help me talk to my data
2. Help me Automate and Plan Tasks
3. Elevate Conversations



We conducted 300 workshops

This is what we have learnt

1. Help me talk to my data
2. Help me Automate and Plan Tasks
3. Elevating Conversations
4. Help me get productive and ship code quicker



utils.py

```python
import torch
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score

def load_data():
    iris = datasets.load_iris()
    X = iris.data
    y = iris.target

    # Standardize the data
    scaler = StandardScaler()
    X = scaler.fit_transform(X)

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

    # Convert numpy data to PyTorch tensors
    X_train = torch.tensor(X_train, dtype=torch.float32)
```

main.py

```python
from utils import load_data, evaluate_predictions
from model import IrisClassifier as Classifier
def main():
    # Model training and evaluation

if __name__ == '__main__':
    main()
```

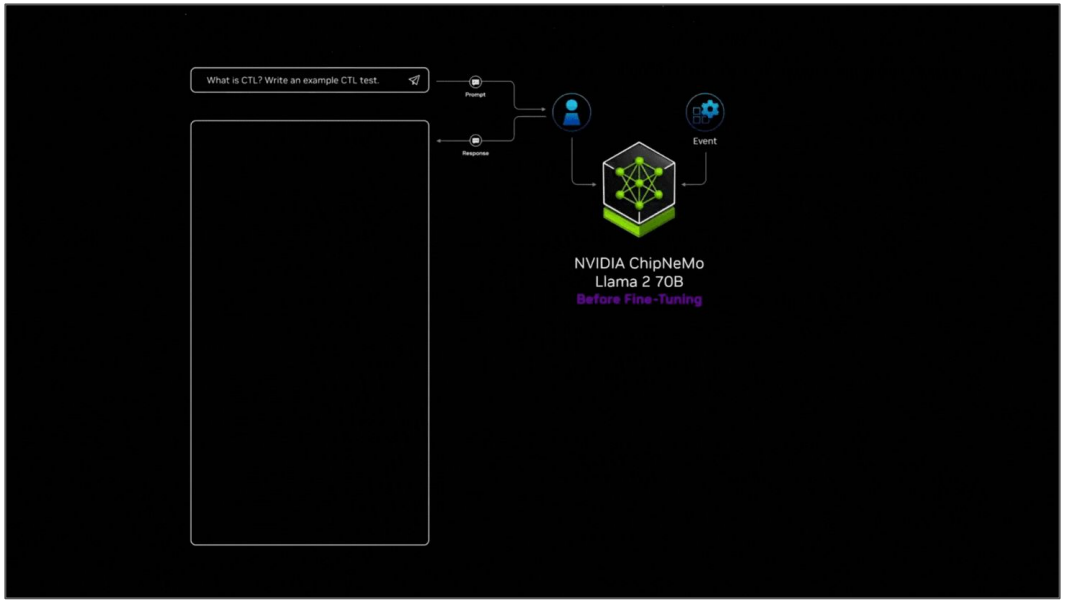
model.py

```python
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import DataLoader, TensorDataset

class IrisClassifier(nn.Module):
    def __init__(self):
        super(IrisClassifier, self).__init__()
        self.fc = nn.Sequential(
            nn.Linear(4, 16),
            nn.ReLU(),
            nn.Linear(16, 3)
        )

    def forward(self, x):
        return self.fc(x)

    def train_model(self, X_train, y_train, epochs, lr, batch_size):
        criterion = nn.CrossEntropyLoss()
        optimizer = optim.Adam(self.parameters(), lr=lr)

        # Create DataLoader for batches
        dataset = TensorDataset(X_train, y_train)
```

We conducted 300 workshops

This is what we have learnt

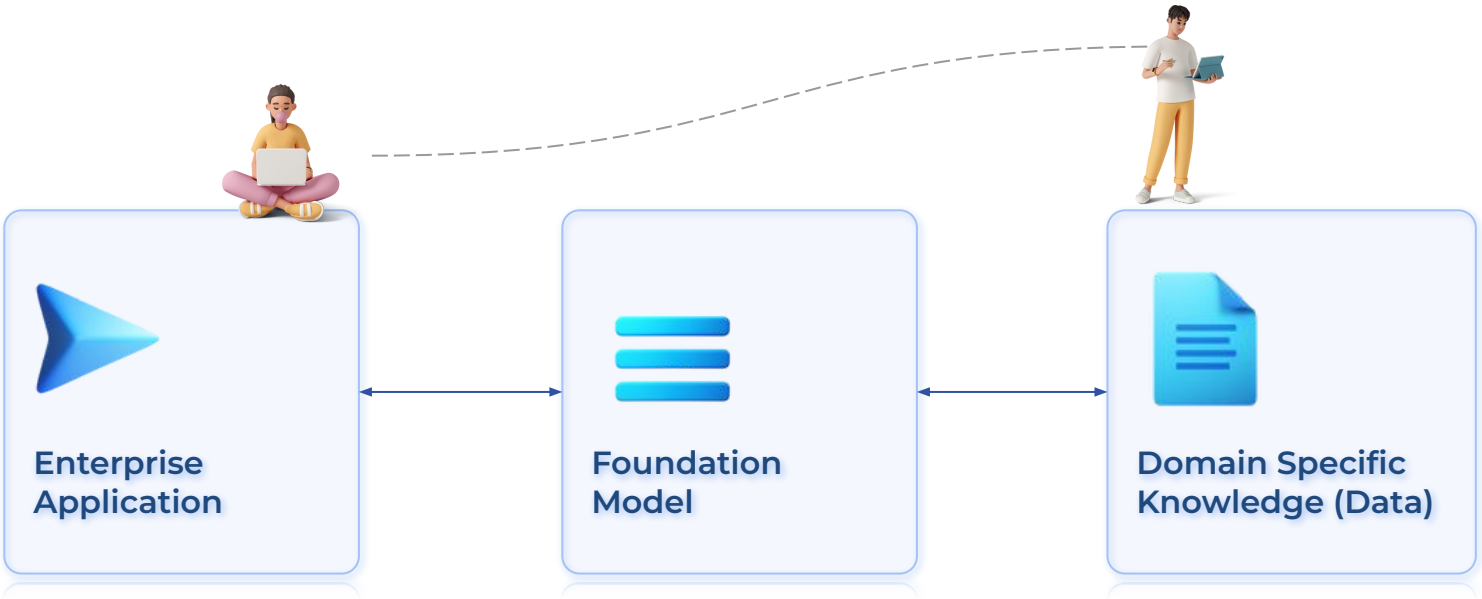
1. Help me talk to my data
2. Help me Automate and Plan Tasks
3. Elevating Conversations
4. Help me get productive and ship code quicker
5. Help me catalyze growth with my Data Moat



quantiphi
Solving What Matters

AUGMENTING LLMS WITH RETRIEVAL: *THE OPEN BOOK EXAM*

Enhancing Task-Specific Performance



Enhanced Context Understanding

Tailored Responses

Reduced Irrelevance

03

CASE STUDIES: Early success with boosting knowledge worker productivity

quantiphi
Solving What Matters

RAG SUCCESS: *CASE STUDIES*

Dialog Agent for Field Technicians

Intelligent RAG-based agent to assist Telecom Field Operators in configuration, installation, and troubleshooting network and customer devices and minimize supervisor involvement

90% Fewer Supervisor Calls, quicker turnaround time and cost reduction

LLM Agents for Banking IT Operations Management

LLM-powered Q&A and insights from SQL logs, and IT product marketplace using knowledge graphs, to monitor and manage IT Ops based on defined IT access rules and protocols

5X Improvement in IT Incident Prioritization and Resolution*

Clinical Trial Protocol Matching

RAG-powered system to automate matching process of trial-eligible patients using EHR data, based on alignment with trial protocols and providing eligibility rationale

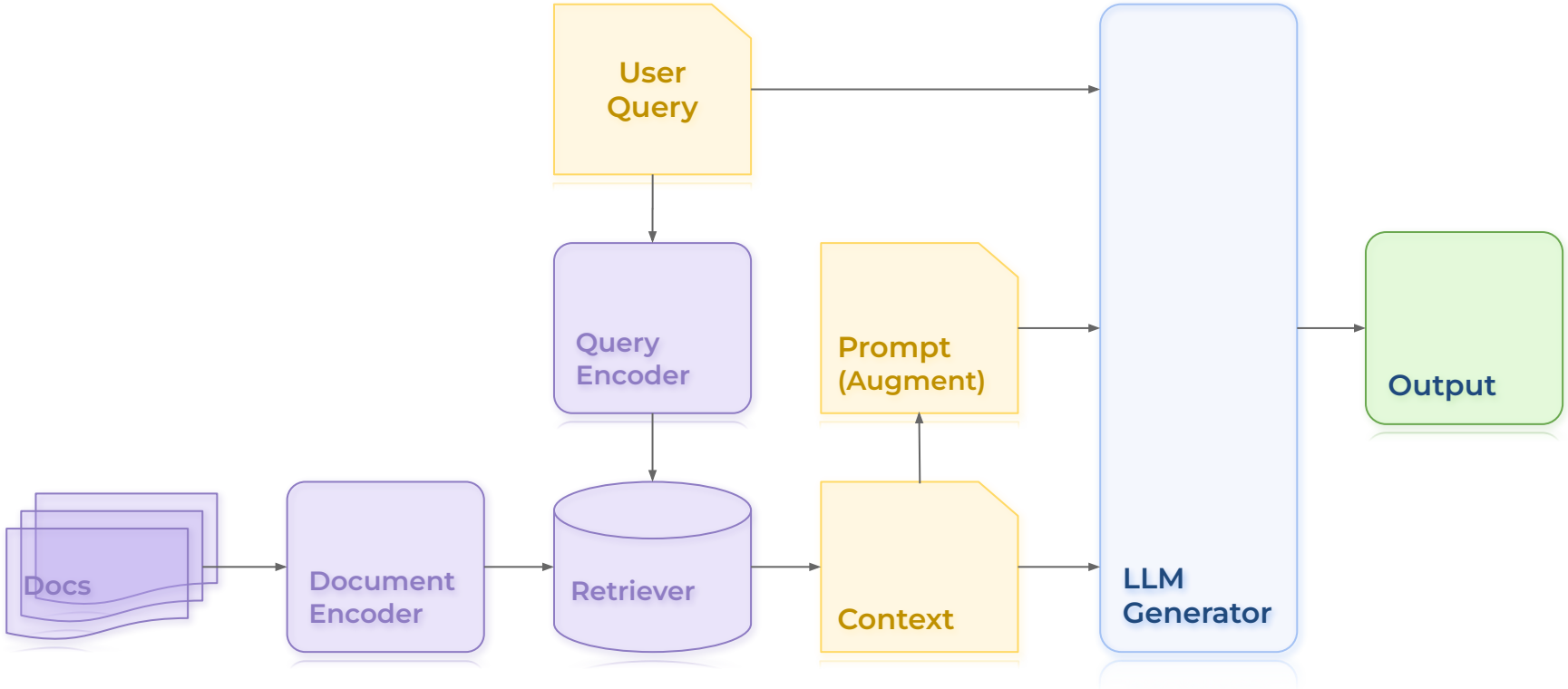
90% Reduction in Patient Matching Turnaround Time (TAT)

Powered by NVIDIA AI Enterprise

04

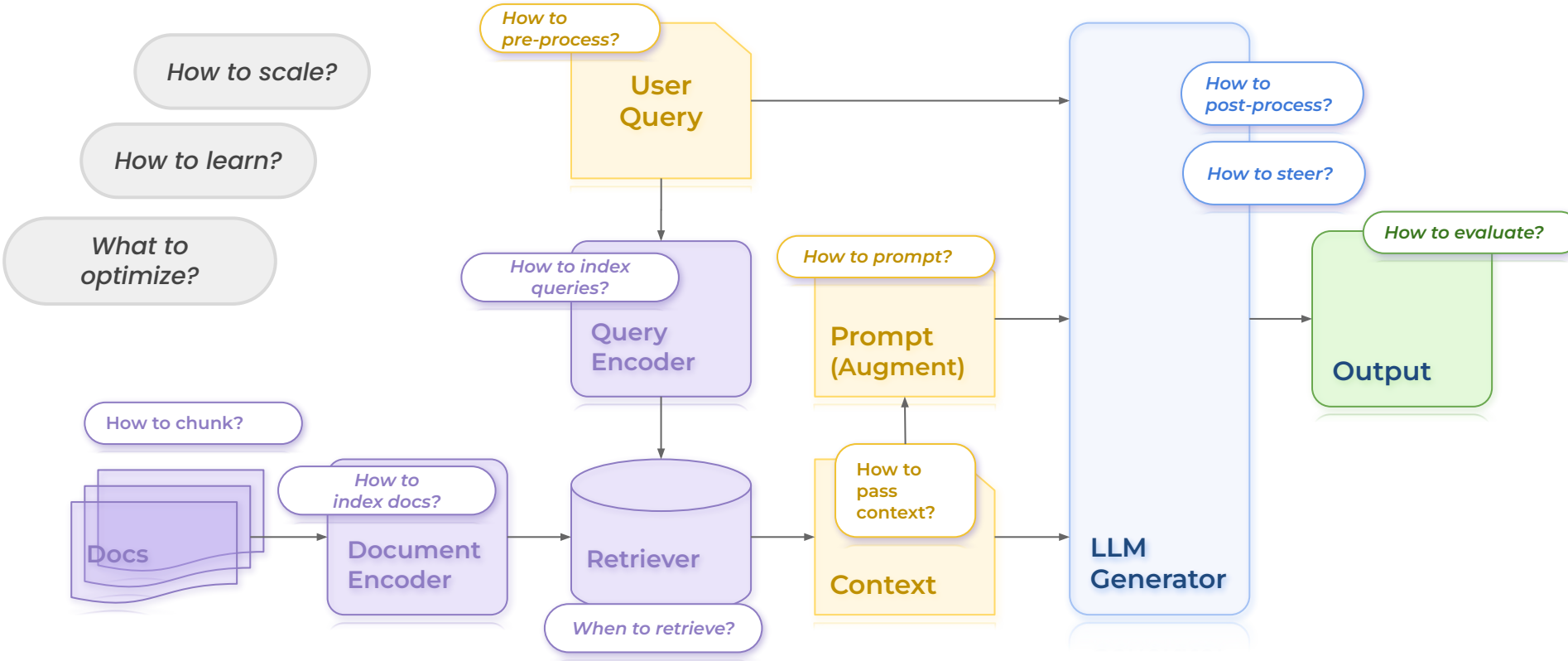
No Free Lunch: The Hidden Technical Debt of Building RAG Applications

THE EASY: *BUILDING A RAG APPLICATION, GREAT TOOLING!*



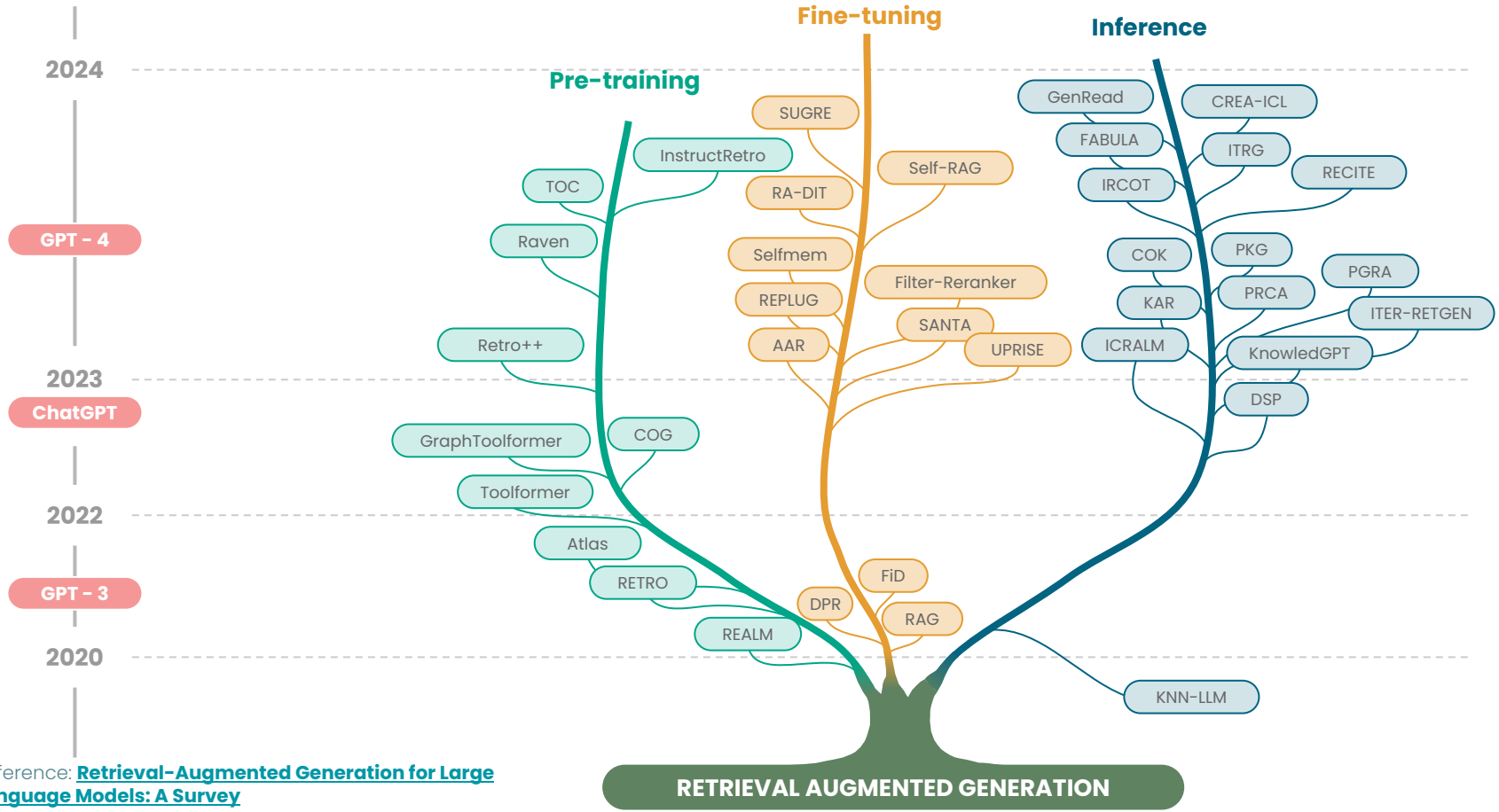
quantiphi
Solving What Matters

THE COMPLEX: *SYSTEM DESIGN FOR RAG APPLICATIONS*



quantiphi
Solving What Matters

THE UNKNOWN: *EVER EVOLVING FRONTIER OF RAG*



Reference: [Retrieval-Augmented Generation for Large Language Models: A Survey]

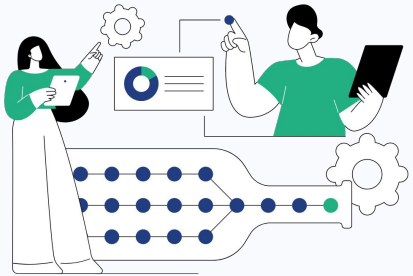
06

Conclusion

LESSONS LEARNT FROM PRACTICAL DEPLOYMENTS

PRACTICAL GENERATIVE AI SYSTEMS ARE COMPOUND AI SYSTEMS

MAP DOMAIN REQUIREMENTS TO SYSTEM CONSTRAINTS EARLY



BUILD WITH NVIDIA AI ENTERPRISE SOFTWARE

RAGS are a Full Stack Accelerated Computing Challenge

….. Each Prompt is an Inference Request

A Call to Multiple Components …..

Each Component Needs Acceleration …..

quantiphi
Solving What Matters

3x

Americas Service Delivery Partner of the year

Join us at:

- **Booth #1513**
  AI CoE Pavillion
- **Booth #G129**
  Generative AI Pavillion

Scan to accelerate your AI journey

Thank you