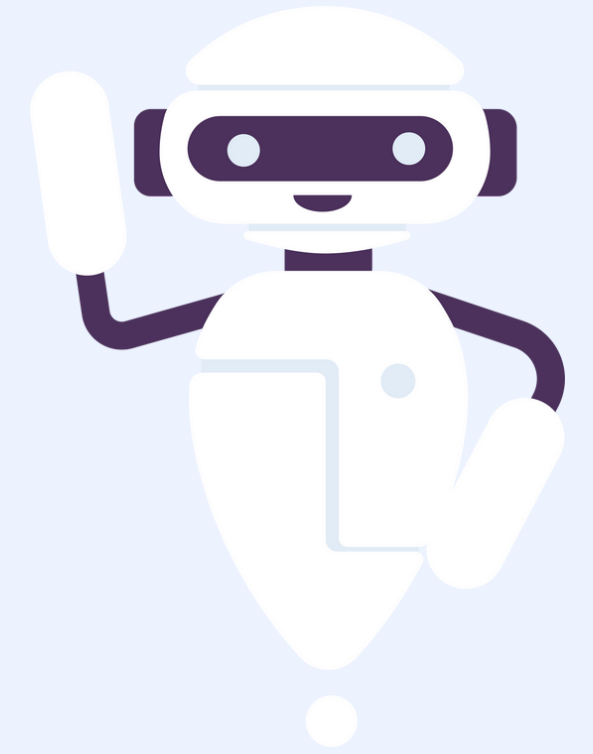
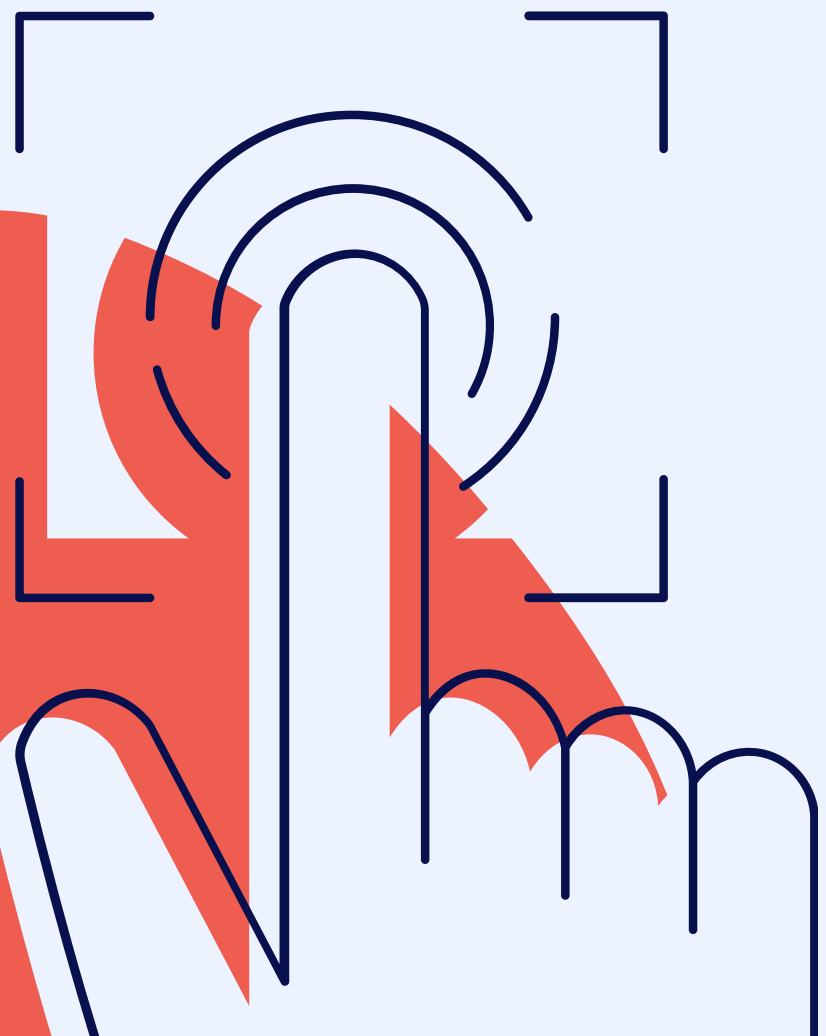




school of ai
Algiers



Build GPT from Scratch (smaller version)



About me

- **5th Year Student at "ESI Alger" Computer Systems Option.**
- **MLOps intern at Namla**
- **Technical Department Co-manager at School of AI Algiers.** ●



LLMs

Why ?

How ?

**Why people are
afraid of them ?**

**What are
they ?**

LLMs

From when ?

Do they think ?

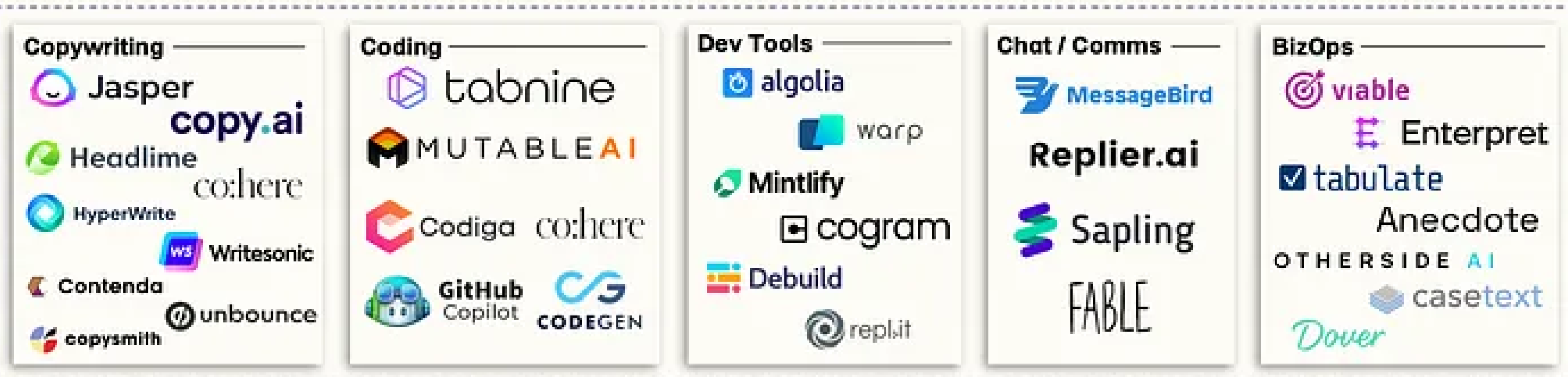
They are everywhere !

LLMs

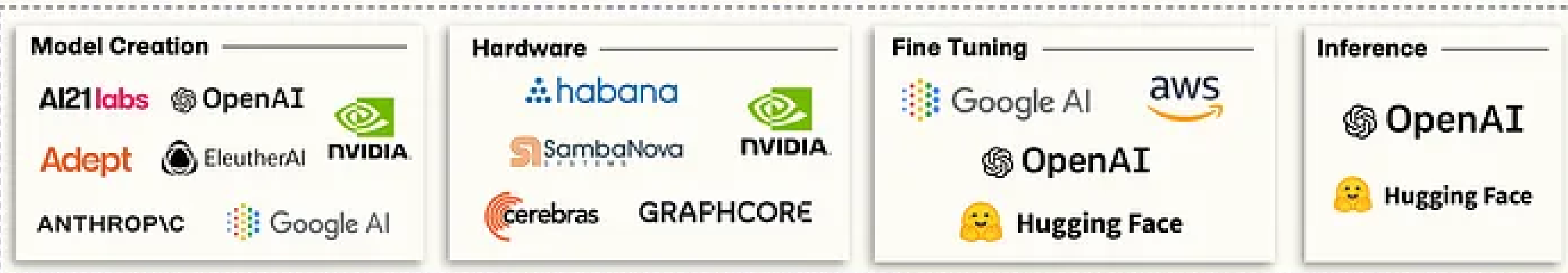
Large Language Models

BCV

Application Layer



Infrastructure Layer



From BainCapitalVentures.com

Large Language Models

BCV

Application Layer

Copywriting



Coding



Dev Tools



Chat / Comms



BizOps



Infrastructure Layer

Model Creation



Hardware



Fine Tuning



Inference



Large Language Models

BCV

the forefront of
designing and
creating advanced
language models.

creating the
hardware
necessary to train
large-scale
language models

facilitate the fine-
tuning of language
models to improve
their accuracy and
performance for
specialized
applications

the deployment
of language
models for real-
world
applications,
offering APIs for
seamless
integration

Infrastructure Layer

Model Creation

AI21labs OpenAI
Adept EleutherAI NVIDIA
ANTHROPIC Google AI

Hardware

habana
SambaNova NVIDIA
cerebras GRAPHCORE

Fine Tuning

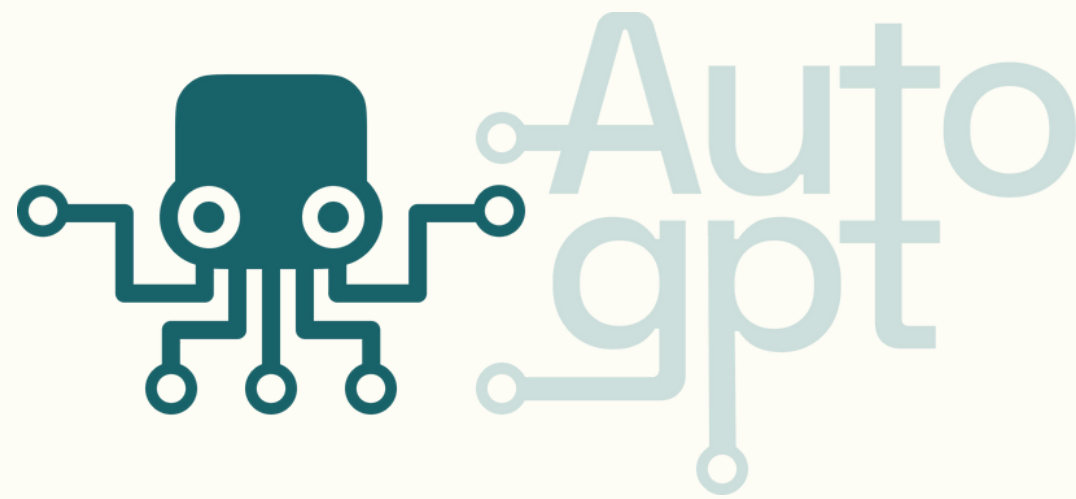
Google AI AWS
OpenAI
Hugging Face

Inference

OpenAI
Hugging Face



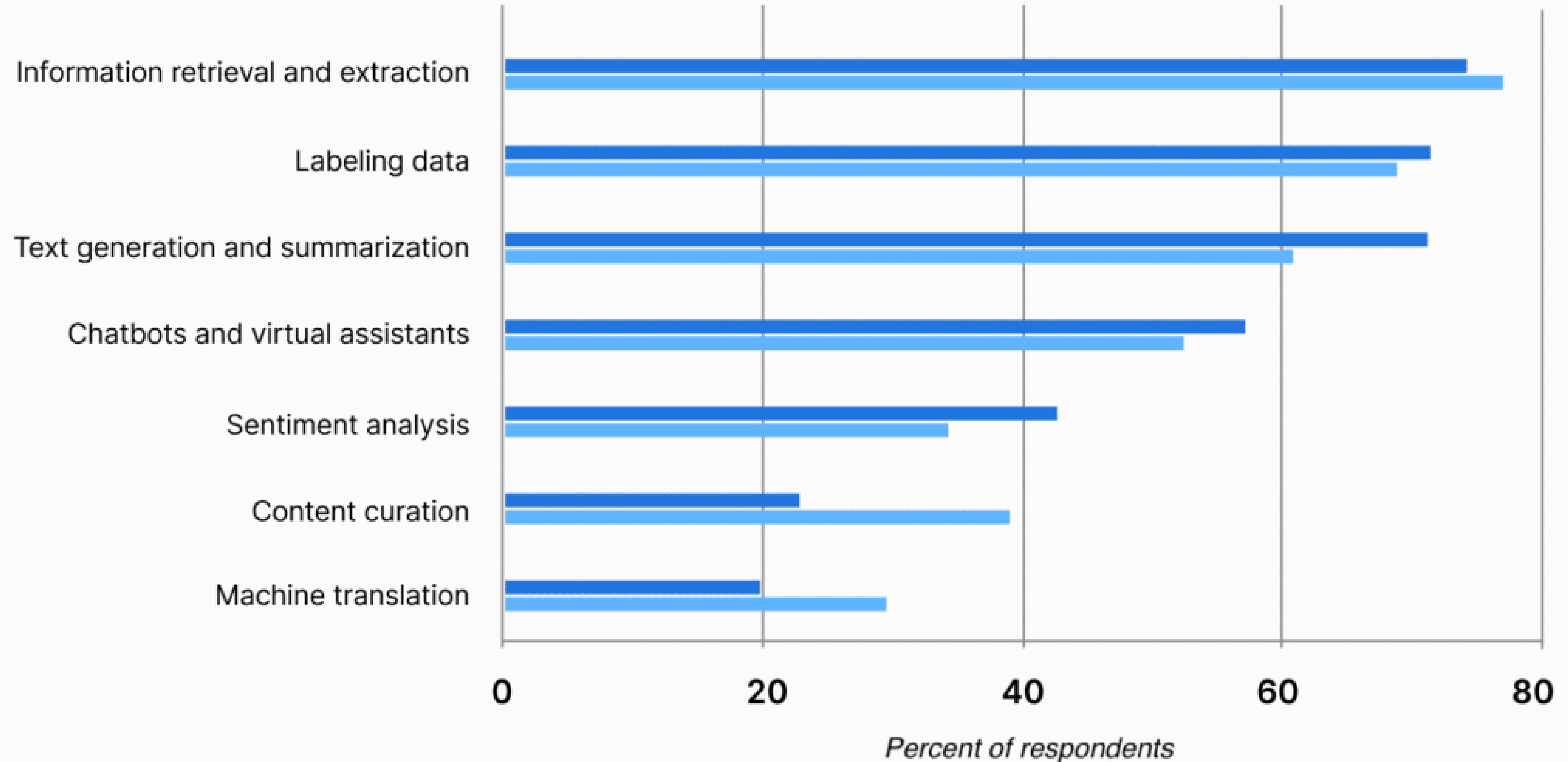
**Your own LLM-
Powered Applications**



**Your own LLM-
Powered Applications**

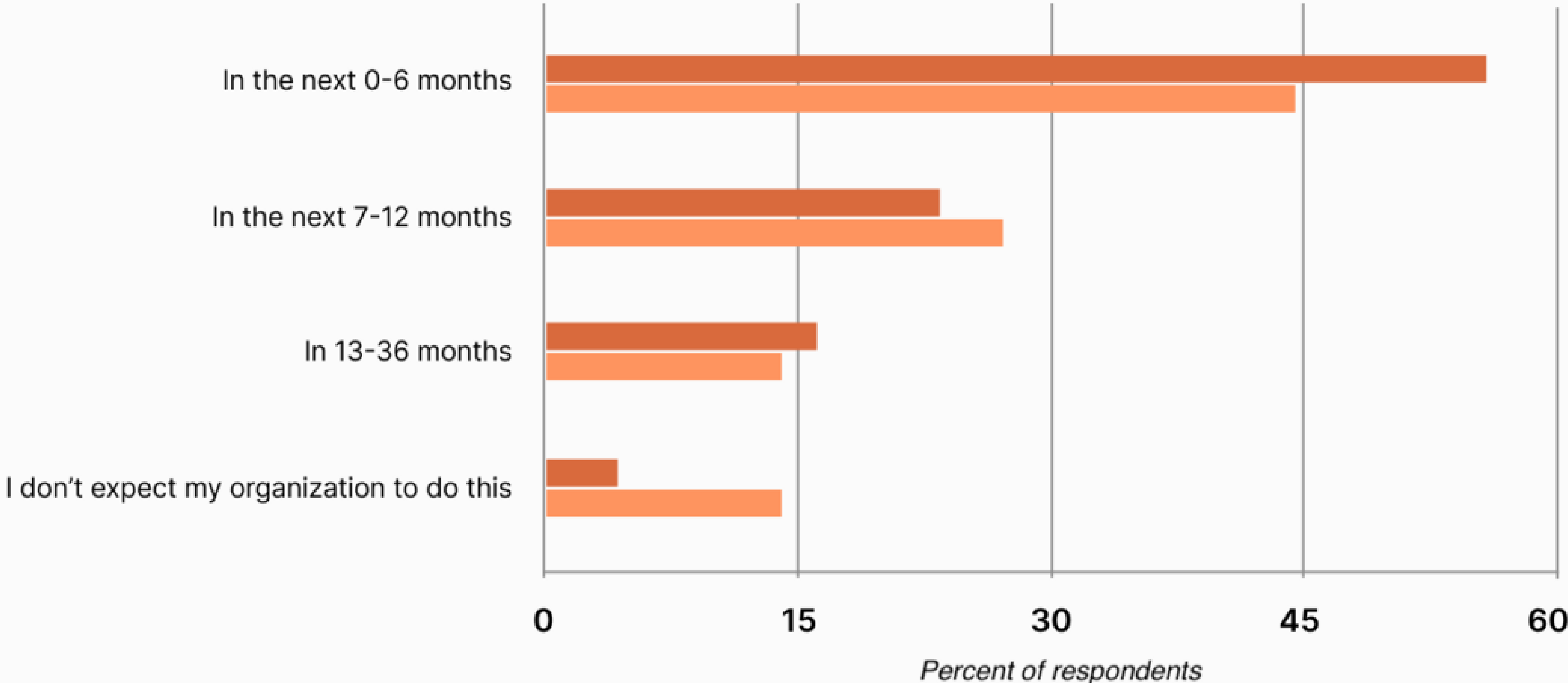
How are you planning to use LLMs?

● The \$billion+ population ● The "other" population

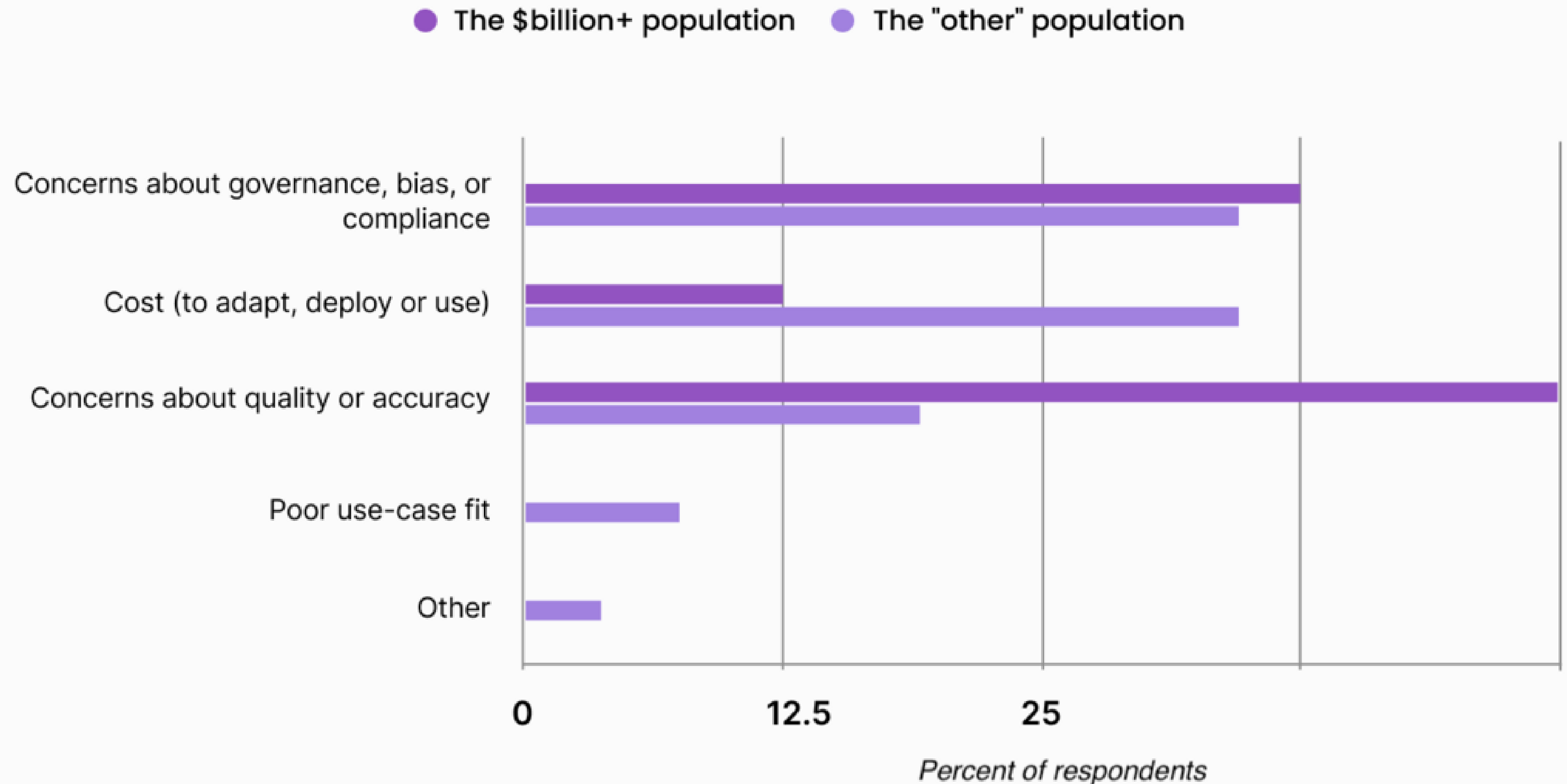


On what timeline do you expect your organization to leverage a large language model in production?

● The \$billion+ population ● The "other" population



What do you expect to be the biggest challenge in leveraging LLMs?



LLMs in 2023 and beyond !



LLMs in 2023 and beyond !

**Some abilities, tasks and solutions
only emerge with scale !**

This is a necessity for a new perspective



This is a necessity for a new perspective

The “yet” perspective:

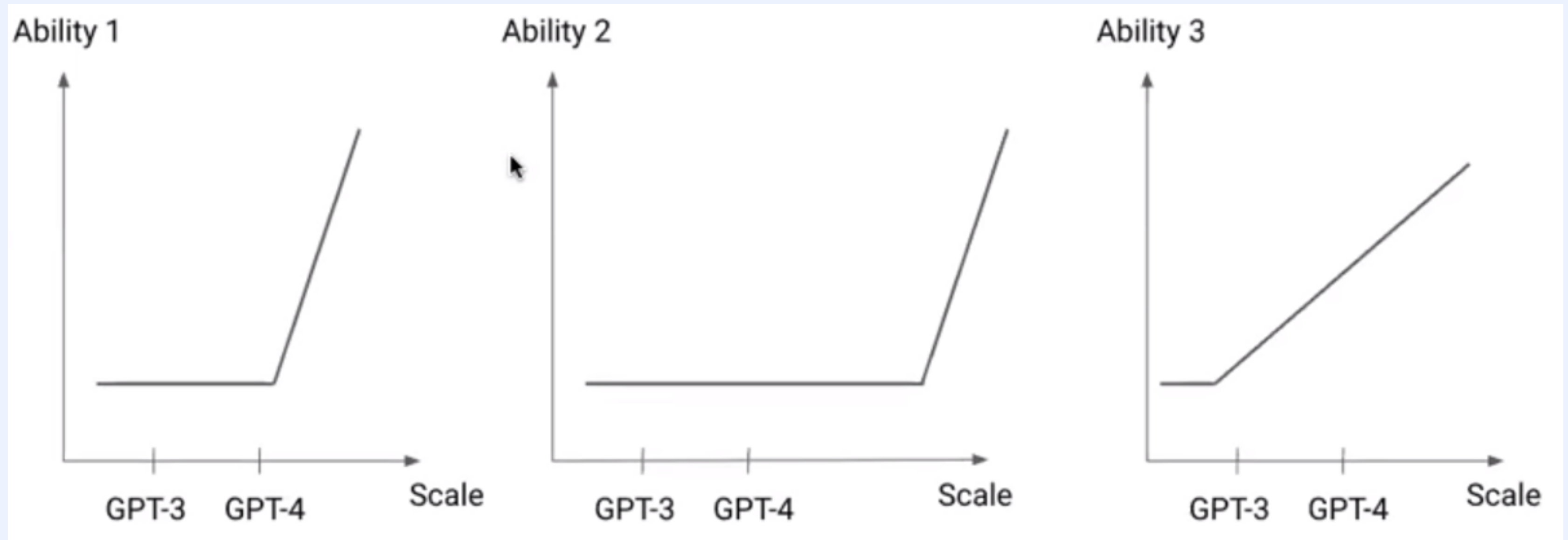
**This idea
doesn't work**



**This idea doesn't
work yet**



The “yet” perspective:



Emergent Abilities with Scale



LLMs also have a lot of limitations !

- inability to provide concise and contextually appropriate responses.
- trained to predict the next token in a sequence
- lacking the crisp
- focused answers desired in specific scenarios



LLMs also have a lot of limitations !

- inability to provide concise and contextually appropriate responses.
- trained to predict the next token in a sequence
- lacking the crisp
- focused answers desired in specific scenarios

Solution (Trick)

structuring questions in a manner that
prompts the model to deliver concise
answers



Solution (Trick)

Q: The square root of x is the cube root of y . What is y to the power of 2, if $x = 4$?

A:



Pretrained model just predicts the next token, which happens to be the answer

Solution (Trick)

The need for efficient and
targeted Post-training



perform an academic search in order to enumerate the key components of the 5G architecture Network (and not the technologies employed) !



ChatGPT

The 5G architecture network consists of various components designed to fulfill the network's high-level objectives such as increased capacity, improved data rate, decreased latency, and better quality of service [7]. The architecture encompasses several elements including:

1. **Device-to-Device (D2D) Communication:** 5G networks facilitate direct communication between devices, enhancing efficiency and enabling new services.
2. **Small Cell Access Points:** Small cells are used to provide increased coverage and capacity in densely populated areas.
3. **Network Cloud:** The network cloud component provides the infrastructure for cloud-based services and applications.
4. **Internet of Things (IoT):** 5G architecture supports the connectivity of a vast array of IoT devices, fostering the development and deployment of IoT applications [7].

Additional components indicated in the architectural framework include:

1. **Control and User Plane Separation (CUPS):** This simplifies the core network by separating control and user data, enhancing scalability and flexibility.
2. **Flexible Network Components:** These are designed to satisfy various service requirements ensuring that the network can cater to a wide range of applications and services.
3. **Unified Database Management:** Centralized management of databases enhances efficiency and simplifies network operations.
4. **End-to-End (E2E) Network Slicing:** This is a foundational component to support diversified 5G services, allowing for the creation of multiple virtual networks on a single physical infrastructure. Network slicing is key to the evolution of 5G network architecture, and is based on Network Function Virtualization (NFV) and Software-Defined Networking (SDN) technologies [7].

These components are part of a well-structured network architecture designed to meet the

Message ChatGPT...



Certified King of LLMs



Certified King of LLMs



Certified King of LLMs



Benchmark	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (includes benchmark-specific training)
MMLU Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM	75.2% 5-shot Flan-PaLM
HellaSwag Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLAMA (validation set)	85.6% ALUM
AI2 Reasoning Challenge (ARC) Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot	85.2% 25-shot	84.2% 8-shot PaLM	85.6% ST-MOE
WinoGrande Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	84.2% 5-shot PALM	85.6% 5-shot PALM
HumanEval Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM	65.8% CodeT + GPT-3.5
DROP (f1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM	88.4 QDGAT



**Certified King of
LLMs**

**Build GPT from
scratch
(a smaller version)**



**What I cannot create, I do not
understand**

-Richard Feynman



THANKS A LOT AND
GOOD LUCK 🍀🧙

