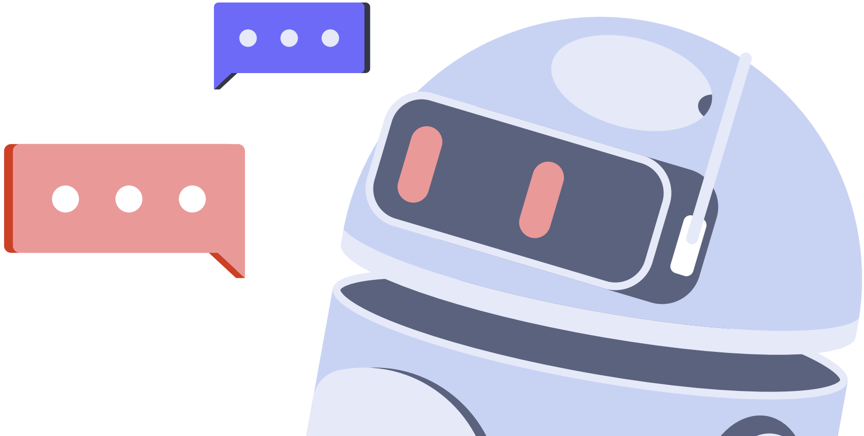
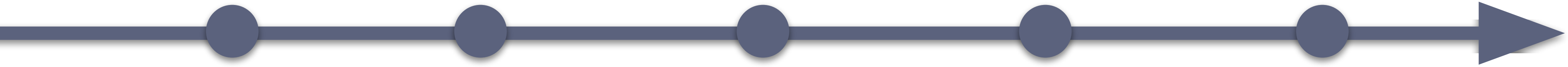
Boosting Business or a Risky Gamble?

Enterprise Perspectives on Large Language Models



Szilvia Hodvogner
2023 ML Forum

Large Language Models (LLMs)



2018
GPT

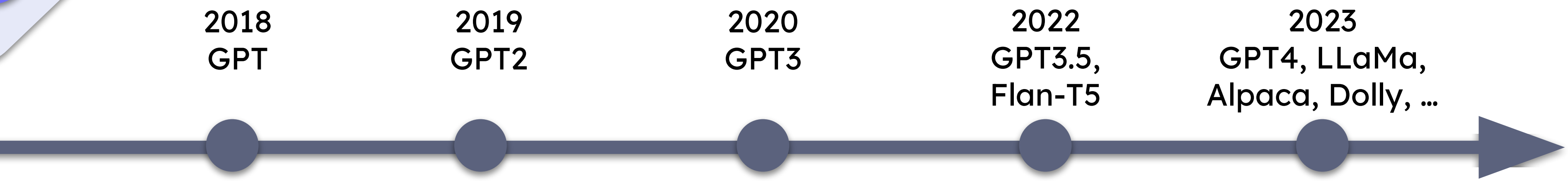
2019
GPT2

2020
GPT3

2022
GPT3.5,
Flan-T5

2023
GPT4, LLaMa,
Alpaca, Dolly, …

Large Language Models (LLMs)



2018
GPT

2019
GPT2

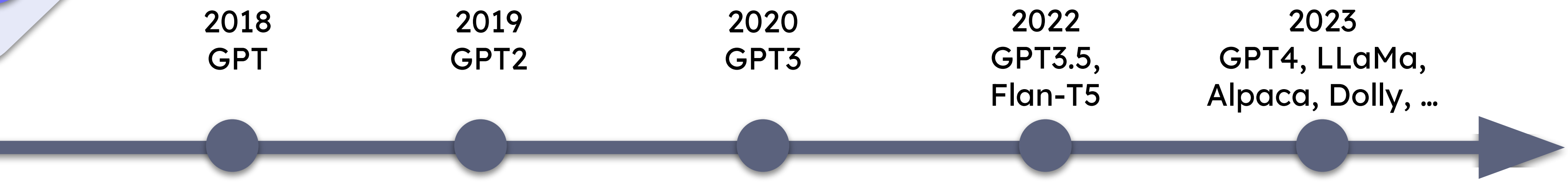
2020
GPT3

2022
GPT3.5,
Flan-T5

2023
GPT4, LLaMa,
Alpaca, Dolly, …

- Transformers with attention
- Billions of parameters
- Petabytes of data from the internet
- Advancements:
  - N-shot learning
  - Reinforcement Learning from Human Feedback (RLHF)
  - Instruction-tuning
  - …

Large Language Models (LLMs)



- Transformers with attention
- Billions of parameters
- Petabytes of data from the internet
- Advancements:
  - N-shot learning
  - Reinforcement Learning from Human Feedback (RLHF)
  - Instruction-tuning
  - …

Goal: Predict the next word → Limitless potential!

Limitless potential… BUT

Ethical and social challenges

↓

Responsible AI Regularization

Model-level constraints

Technology-level limitations

Limitless potential... BUT

Ethical and social challenges

↓

Responsible AI Regularization

- Disruptive
- Effects? - jobs, teaching, content creating
- Legal policies? - training data, biases, alignment

Limitless potential… BUT

Ethical and social challenges

↓

Responsible AI Regularization

Model-level constraints

↓

Parameters, configurations, legal questions

Technology-level limitations

Closed vs open-source models

		Closed development		Open-source model	
		GPT3.5	**GPT4**	**LLaMa**	**Dolly**
Params	# params	175B	1000B	65B	12B
	Context token size	4K	32K	8K	2K
Resources	Inference RAM needs	~750GB	Unknown	~130GB	~32GB
	Pricing ($/1K tokens)	0.002	0.03/0.06	Free	Free
Results	Commonsense reasoning (HellaSwag dataset)	86%	95%	84%	71%
	Q&A, pronoun resolution (WinoGrande dataset)	82%	88%	77%	62%

Closed vs open-source models

		Closed development		Open-source model	
		GPT3.5	**GPT4**	**LLaMa**	**Dolly**
Params	# params	175B	1000B	65B	12B
	Context token size	4K	32K	8K	2K
Resources	Inference RAM needs	~750GB	Unknown	~130GB	~32GB
	Pricing ($/1K tokens)	0.002	0.03/0.06	Free	Free
Results	Commonsense reasoning (HellaSwag dataset)	86%	95%	84%	71%
	Q&A, pronoun resolution (WinoGrande dataset)	82%	88%	77%	62%

Closed vs open-source models

		Closed development		Open-source model	
		GPT3.5	**GPT4**	**LLaMa**	**Dolly**
Params	# params	175B	1000B	65B	12B
	Context token size	4K	32K	8K	2K
Resources	Inference RAM needs	~750GB	Unknown	~130GB	~32GB
	Pricing ($/1K tokens)	0.002	0.03/0.06	Free	Free
Results	Commonsense reasoning (HellaSwag dataset)	86%	95%	84%	71%
	Q&A, pronoun resolution (WinoGrande dataset)	82%	88%	77%	62%

Percentage of correct answers

Who owns what?

Closed development		Open-source model	
GPT3.5	**GPT4**	**LLaMa**	**Dolly**
Commercial use	**Commercial use**	**Non-commercial use**	**Commercial use**

Who owns what?

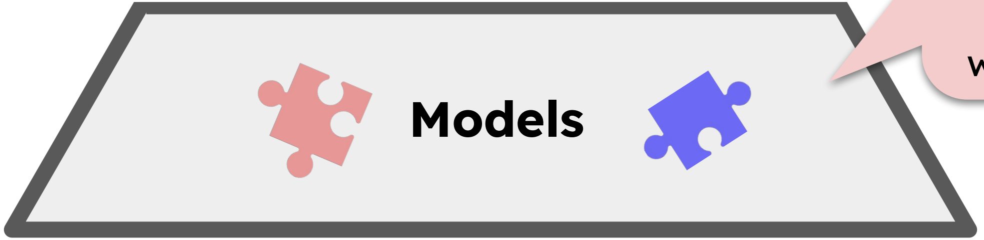
Closed development		Open-source model	
GPT3.5	**GPT4**	**LLaMa**	**Dolly**
Commercial use	**Commercial use**	**Non-commercial use**	**Commercial use**

Not simple!

- Chinchilla: closed development, non-commercial use
- Vicuna: open-source, commercial use licence, but still non-commercial use because LLaMa

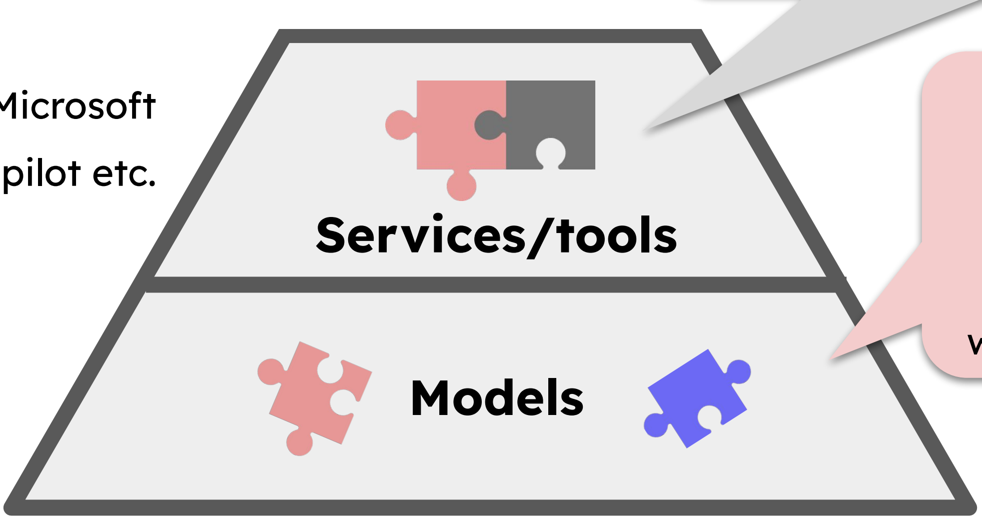
Building blocks

GPT3.5, GPT4, LLaMa,
Dolly, Alpaca, GPT-J,
GPT4All, etc.



Train on custom data
Use in custom workflows

Building blocks



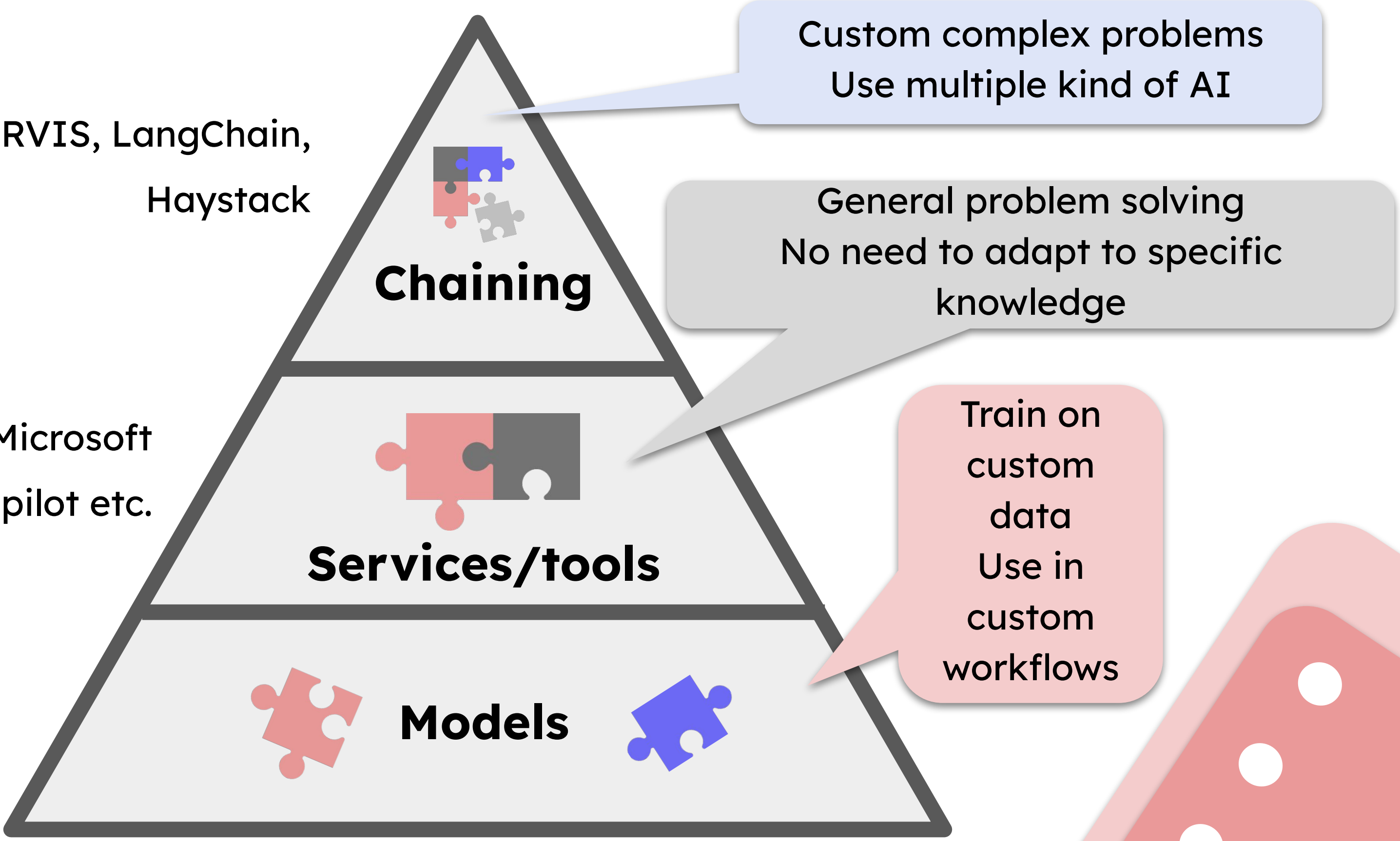
ChatGPT, Bard, Microsoft Bing, GitHub Copilot etc.

GPT3.5, GPT4, LLaMa, Dolly, Alpaca, GPT-J, GPT4All, etc.

General problem solving
No need to adapt to specific knowledge

Train on custom data
Use in custom workflows

Building blocks



JARVIS, LangChain, Haystack

ChatGPT, Bard, Microsoft Bing, GitHub Copilot etc.

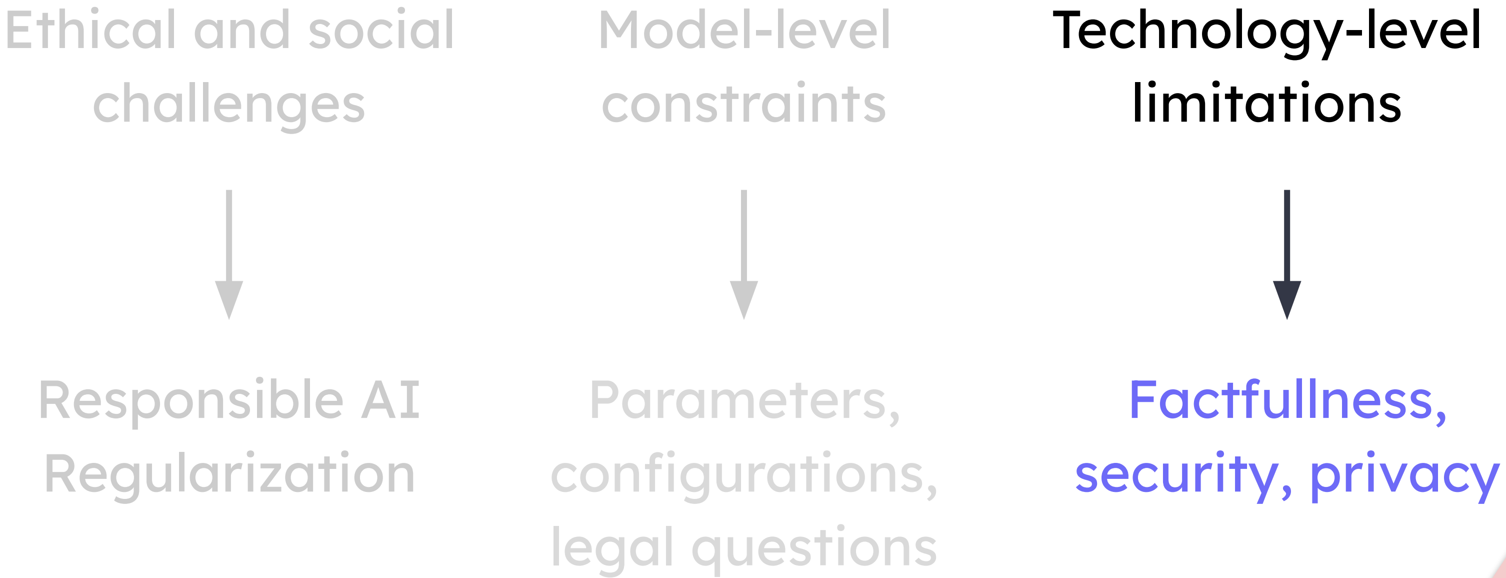
GPT3.5, GPT4, LLaMa, Dolly, Alpaca, GPT-J, GPT4All, etc.

Custom complex problems
Use multiple kind of AI

General problem solving
No need to adapt to specific knowledge

Train on custom data
Use in custom workflows

Limitless potential… BUT



Ethical and social challenges → Responsible AI Regularization

Model-level constraints → Parameters, configurations, legal questions

Technology-level limitations → Factfullness, security, privacy

Factual incorrectness and hallucinations

Hallucinations:

A confident response, but it is nonsensical or unfaithful to the provided source content

Factual incorrectness and hallucinations

Hallucinations:

A confident response, but it is nonsensical or unfaithful to the provided source content

Cause:

Predicting the next word, no reasoning capabilities

Factual incorrectness and hallucinations

Hallucinations:

A confident response, but it is nonsensical or unfaithful to the provided source content

Cause:

Predicting the next word, no reasoning capabilities

Mitigation strategies:

- More data and training
- Fine-tune
- Teaching how to say “I don’t know”
- Middle tier: Query databases, search internet, other models, knowledge graphs

Security & Privacy

Closed-developments - What happens to the upload data?

- Privacy - Data goes through
- Security - Rely on the API

OpenAI

“Retain it 30 day, than dispose”

Azure OpenAI

“Opt out, they cannot use your data”

ChatGPT

“Remove chat history”

Security & Privacy

Closed-developments - What happens to the upload data?

- Privacy - Data goes through
- Security - Rely on the API

Do not share:

- Sensitive data about company, person, or customer
- Proprietary codes
- Emails, meetings drafts

Always check and validate AI response!

OpenAI

“Retain it 30 day, than dispose”

Azure OpenAI

“Opt out, they cannot use your data”

ChatGPT

“Remove chat history”

Security & Privacy

Closed-developments - What happens to the upload data?

- Privacy - Data goes through
- Security - Rely on the API

Open-source solutions

- Privacy - You are safe
- Security - Your concern

OpenAI

“Retain it 30 day, than dispose”

Azure OpenAI

“Opt out, they cannot use your data”

ChatGPT

“Remove chat history”

Security & Privacy

Closed-developments - What happens to the upload data?

- Privacy - Data goes through
- Security - Rely on the API

OpenAI

“Retain it 30 day, than dispose”

Azure OpenAI

“Opt out, they cannot use your data”

ChatGPT

“Remove chat history”

Open-source solutions

- Privacy - You are safe
- Security - Your concern

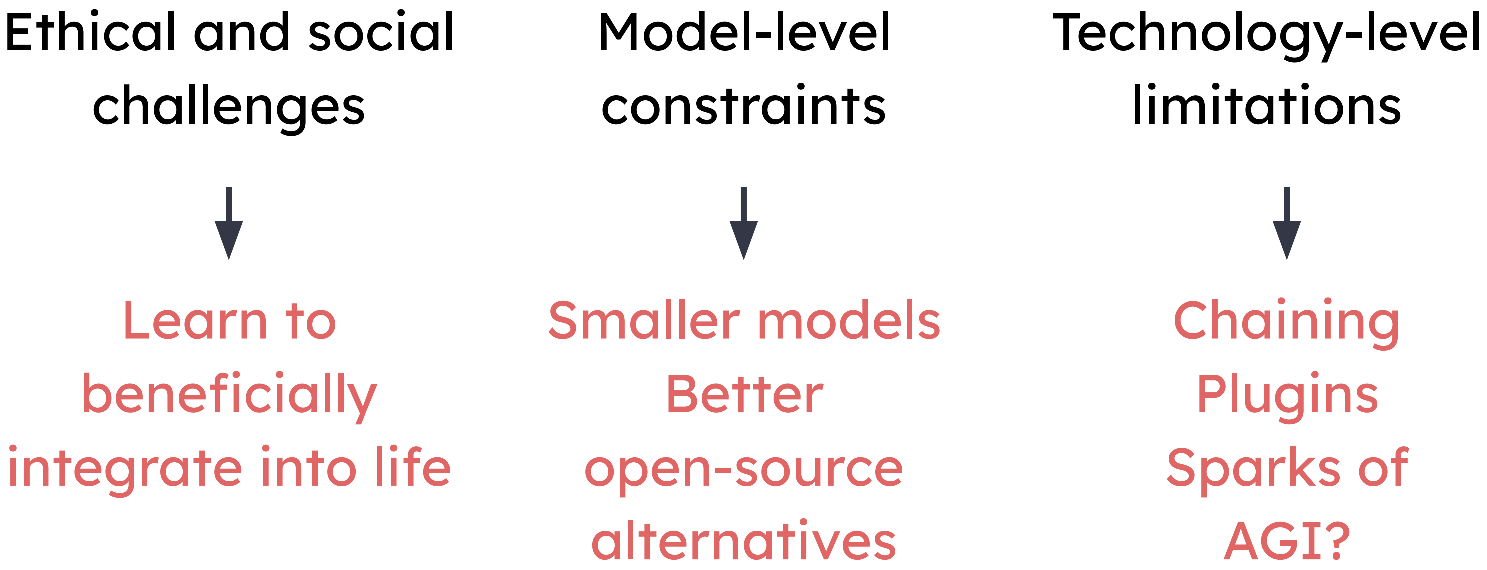
Chaining or combining LLMs with tools

- Caution with autonomous actions!
    - Harmful queries, drop database, send email to wrong contacts, insert obfuscated code, call unintended APIs

Integrating LLMs into a company

	API calls - closed devs (e.g. chatGPT, GPT3.5)	**Local hosting** - open devs (e.g. Dolly, LLaMa, Alpaca)
For ✓	● Easy setup ● Cutting edge models are available ● Low latency	● Higher selection of models ● Lower inference cost ● Independent
Against ✗	● Privacy concerns ● Higher long-term costs ● Depends on a 3rd party API	● Complex and costly setup ● Weaker models

Future



Ethical and social challenges

↓

Learn to beneficially integrate into life

Model-level constraints

↓

Smaller models
Better open-source alternatives

Technology-level limitations

↓

Chaining
Plugins
Sparks of AGI?

Thank you for your attention!

Enterprise Perspectives on Large Language Models