



Centre Suisse d'Intelligence Artificielle
pour les PME – CSIA-PME

AI-DAYS@Hes-so

27–28.1.25

GENÈVE

29.1.25

LAUSANNE



Workshop 1:
RAG: Unveiling the Power of Retrieval-Augmented
Generation

- Célien Donzé, HEIA-FR / HES-SO
- Henrique Marques Reis, HE-ARC, HES-SO
- Jonathan Guerne, HE-ARC, HES-SO

la Mobilière

opi >>>
industries
technologies

alp+ict
western
switzerland
digital
cluster

Hes·so

eksperiens
LA RENCONTRE DU PROJET

25

Schedule - Morning

8:30:00 AM	0:05	Opening remarks	Sébastien Rumley	HEIA-FR, HES-SO / Swiss Ai center
8:35:00 AM	0:30	The Alps research infrastructure at CSCS: enabling world-class ML research in Switzerland	Fawzi Mohamed	The Swiss National Supercomputing Centre (CSCS), ETH Zurich
9:05:00 AM	0:18	SCITAS: On-premise and Cloud Infrastructure driving HPC & AI Scientific Computing at EPFL	Gilles Fourestey	Warm-up
9:23:00 AM	0:18	Picterra's Infrastructure: Scaling ML for Geospatial Imagery Analysis	Julien Rebetez	
9:41:00 AM	0:19	Securing AI Infrastructure: Strategies & Common Pitfalls	Terry Vogelsang	
10:00:00 AM	0:30	Is your infrastructure ready for AI workloads ?	Jean-Baptiste Thomas	Principal Field Solutions Architect – Pure Storage EMEA
10:30:00 AM	0:20	Break - Offered by the Swiss Ai Center		
10:50:00 AM	0:18	Enabling ressources optimisation thru Monitoring of Mixed GPU Setups	Martin Roch-Neirey	HEIA-FR, HES-SO
11:08:00 AM	0:18	Enabling headache-free heterogeneous GPU resource sharing for a small cluster	Abele Mălan, Romain De Laage	UNINE
11:26:00 AM	0:18	Efficient GPU Resource Sharing with Kubernetes and Coder	Dorian Gambin	Use-cases
11:44:00 AM	0:18	Unlocking Performance: vGPU Acceleration on Cisco UCS X-Series with VMware vSphere	Jérémy Gamba	
12:02:00 PM	0:18	GPU infrastructure at UNIL, an attempt at continuous adaptation	Emmanuel Jeanvoine	
12:20:00 PM	0:55	Lunch - Offered by PureStorage and Exoscale		

Schedule - Afternoon



12:20:00 PM	0:55	Lunch - Offered by PureStorage and Exoscale		
1:15:00 PM	0:30	GPU hyperspecialisation	Antoine Coetsier	Exoscale
1:45:00 PM	0:18	Building a SLURM Cluster with Existing Heterogeneous Hardware	Marc Stadelman	Centre for Artificial Intelligence, ZHAW School of Engineering
2:03:00 PM	0:19	Slurm & heterogenous users: avoiding main pitfalls	Ljiljana Dolamic	ar Campus
2:22:00 PM	0:19	OpenOnDemand	Adrien Albert, Yann Sagon	UNIGE
2:41:00 PM	0:19	Where to ask for help when building or operating a GPU cluster?	Marco Merkel	HPC-DoltNow
3:00:00 PM	0:30	Break - Offered by E4		
3:30:00 PM	0:20	OAR: a Versatile Resource and Job Management System to Tame Complexity	Olivier Richard	Pd SA
3:50:00 PM	0:30	Private LLM for industrial users	Daniele Cremonini	E4
4:20:00 PM	0:05	Mini-break		
4:25:00 PM	0:50	Panel/group discussion		
5:15:00 PM		End		

Scaling

Streching

Who are we ?



- Célien Donzé



- Henrique Marques Reis
- Jonathan Guerne

Outline



Technical background



break



Hands on



break



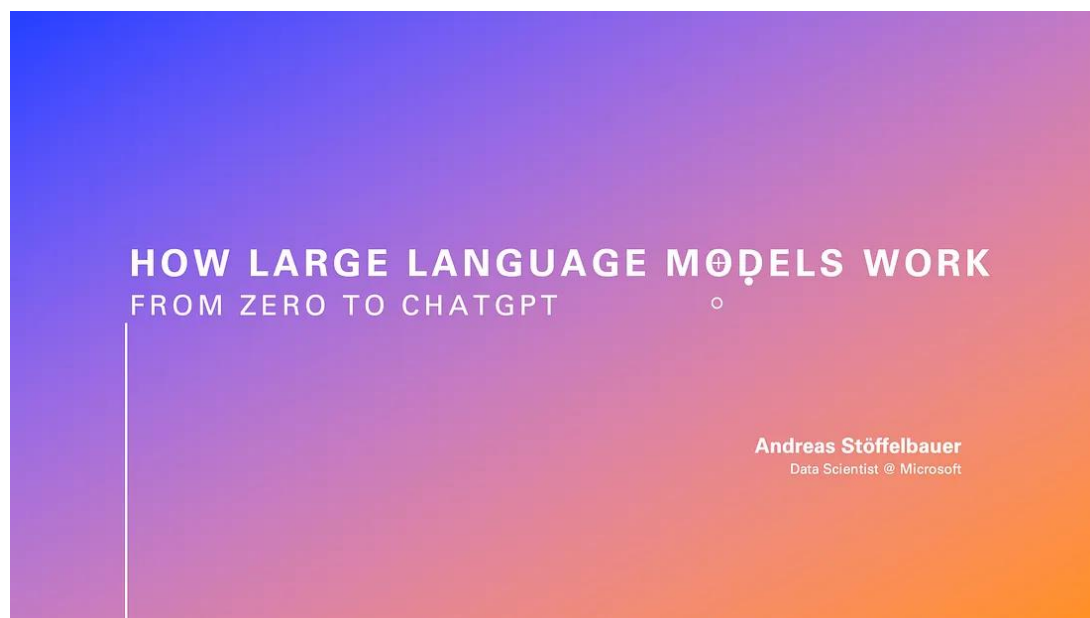
Final discussion



Total duration: 2h30

A bit of background

About LLM: how do they work



Excellent resource from Andreas Stöffelbauer on [Medium](#)

- Large pretrained models for word prediction
- Fine-tuned for better alignment
- Improved via reinforcement learning from human feedback (RLHF)

Natural language generation

After training: We can **generate text** by predicting **one word at a time**

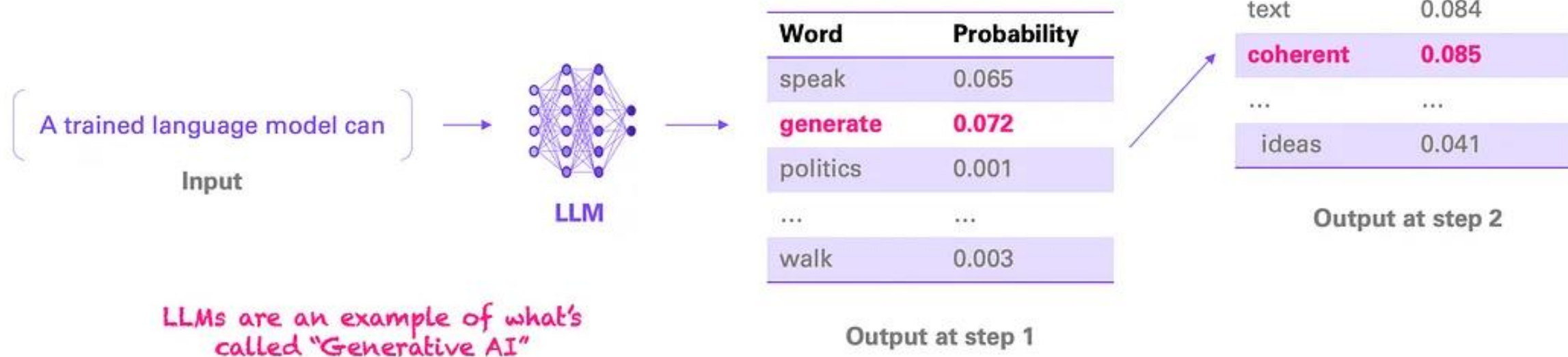


Illustration from Andreas Stöffelbauer

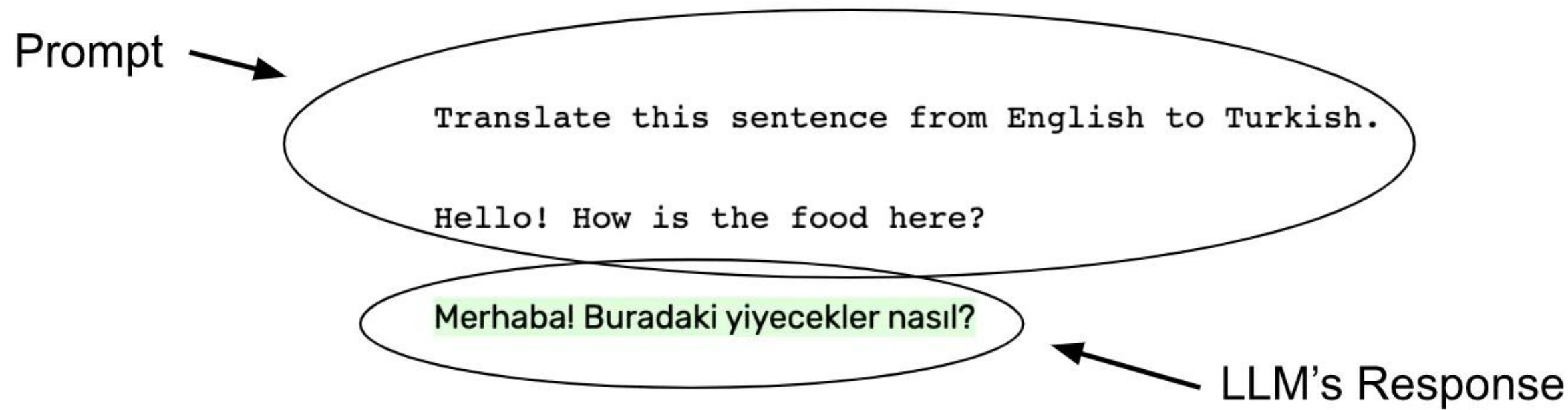
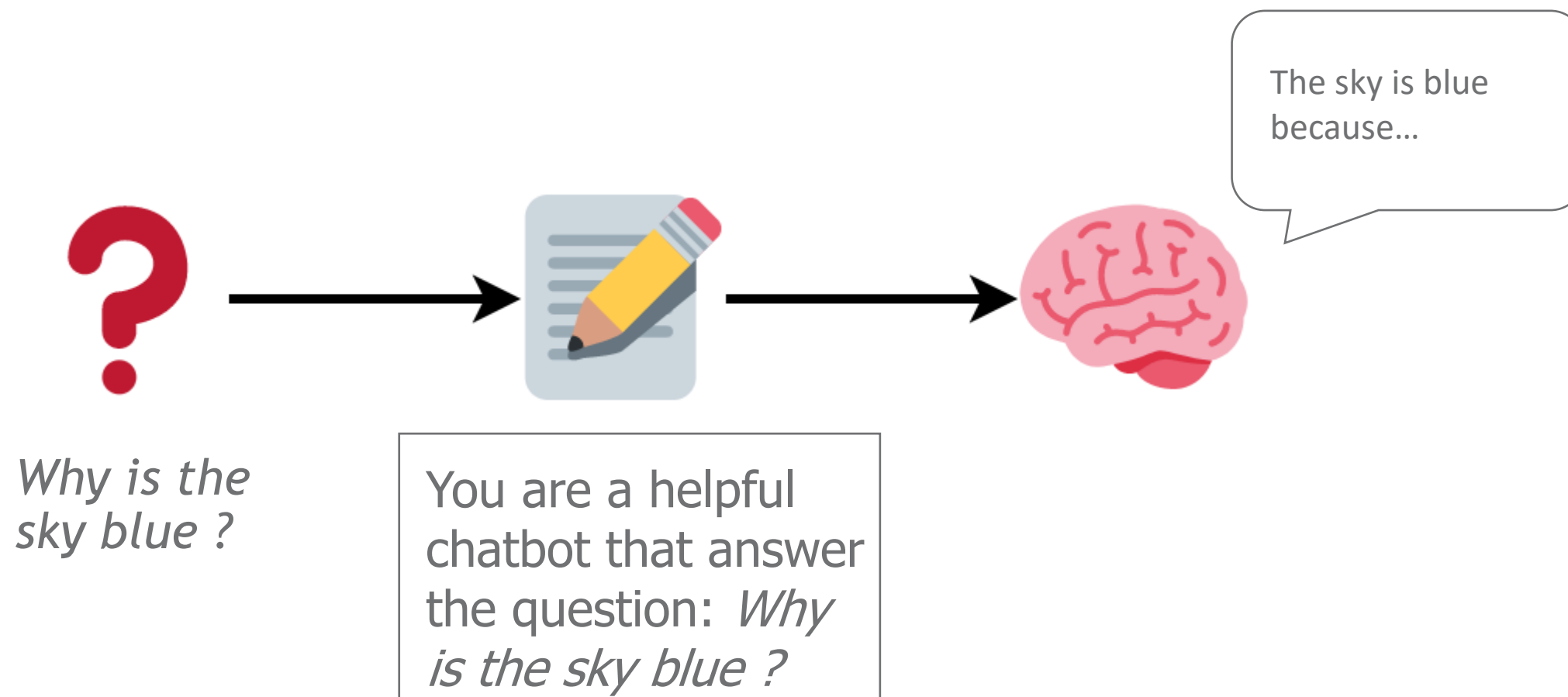


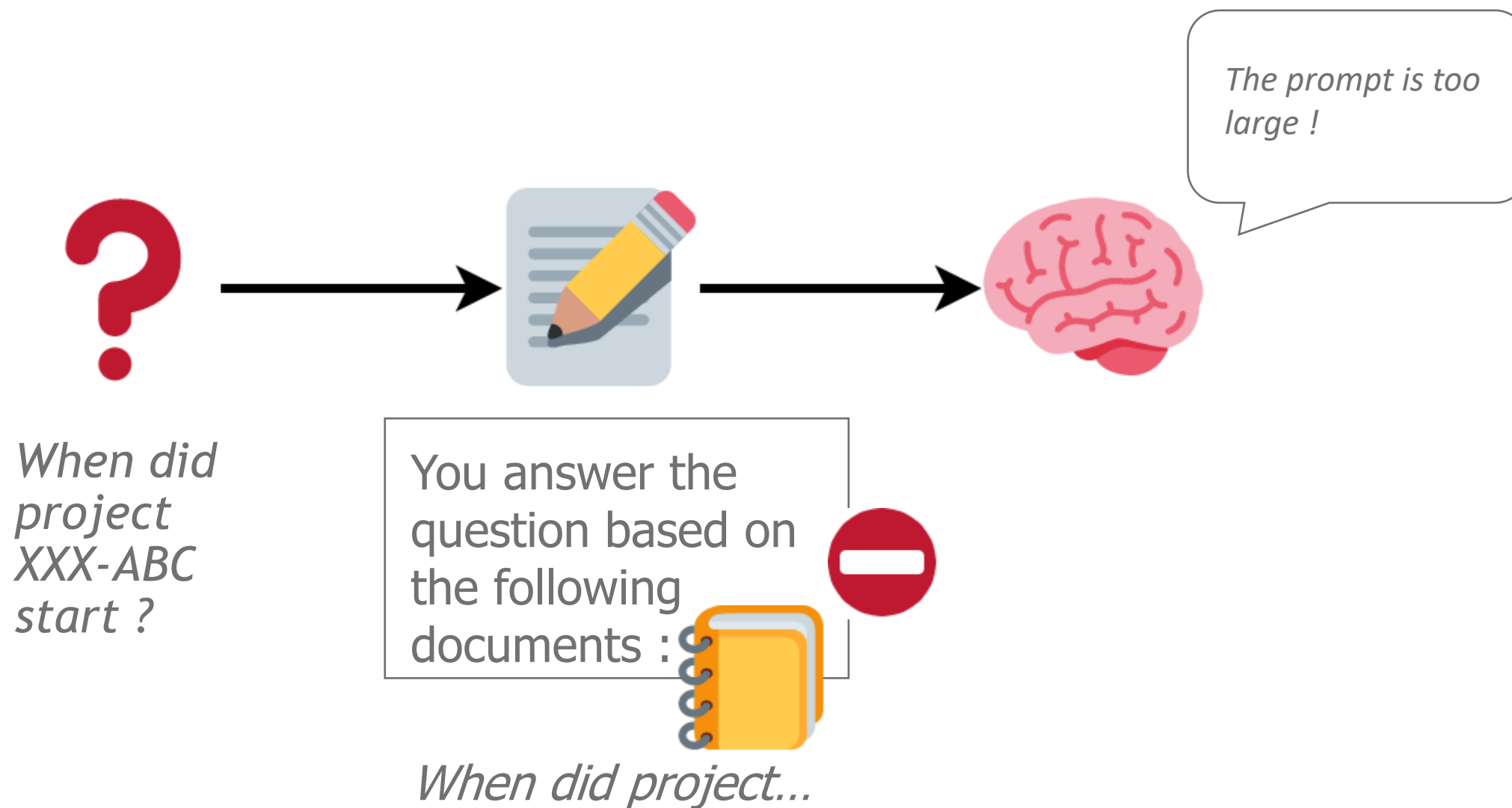
Illustration from Raza Habib and Sinan Ozdemir on [humanloop](https://humanloop.ai)

Interaction with LLMs : a first simplification



How can I ask questions about my own documents ?

Document query: a first intuition

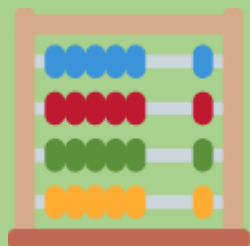


The prompt is too large !

- There are too many documents submitted in the prompt
- Are they all relevant to answer the question ? Probably not...
- How can we find the most relevant documents / sections of document to submit ?
- With the help of **Similarity Search**

Similarity Search: how does it work ?

- Measure the similarity between two texts, i.e. find a distance between texts
- It's easier to compute distances on numbers than words
- How can we turn words into numbers ? With **Embeddings**

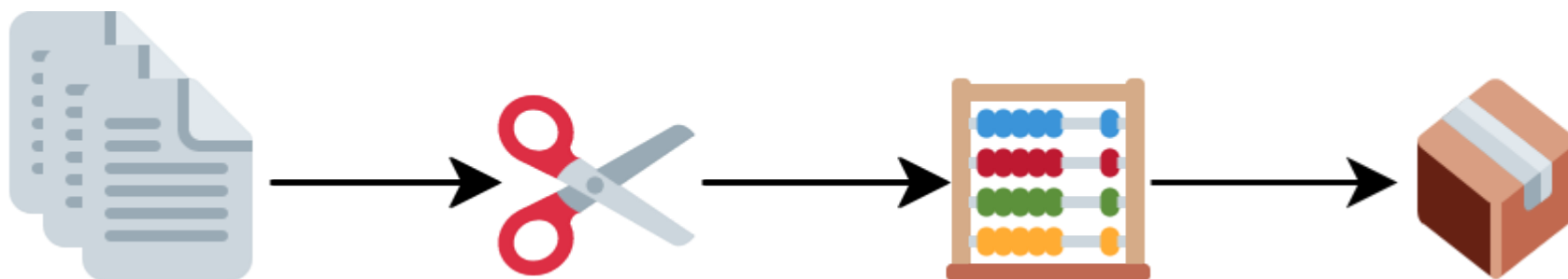


Embedding can be seen as a collection of numbers; the simplest form of embedding would be to count the presence of each word in a sentence. E.g. "Hello world, hello Switzerland" →

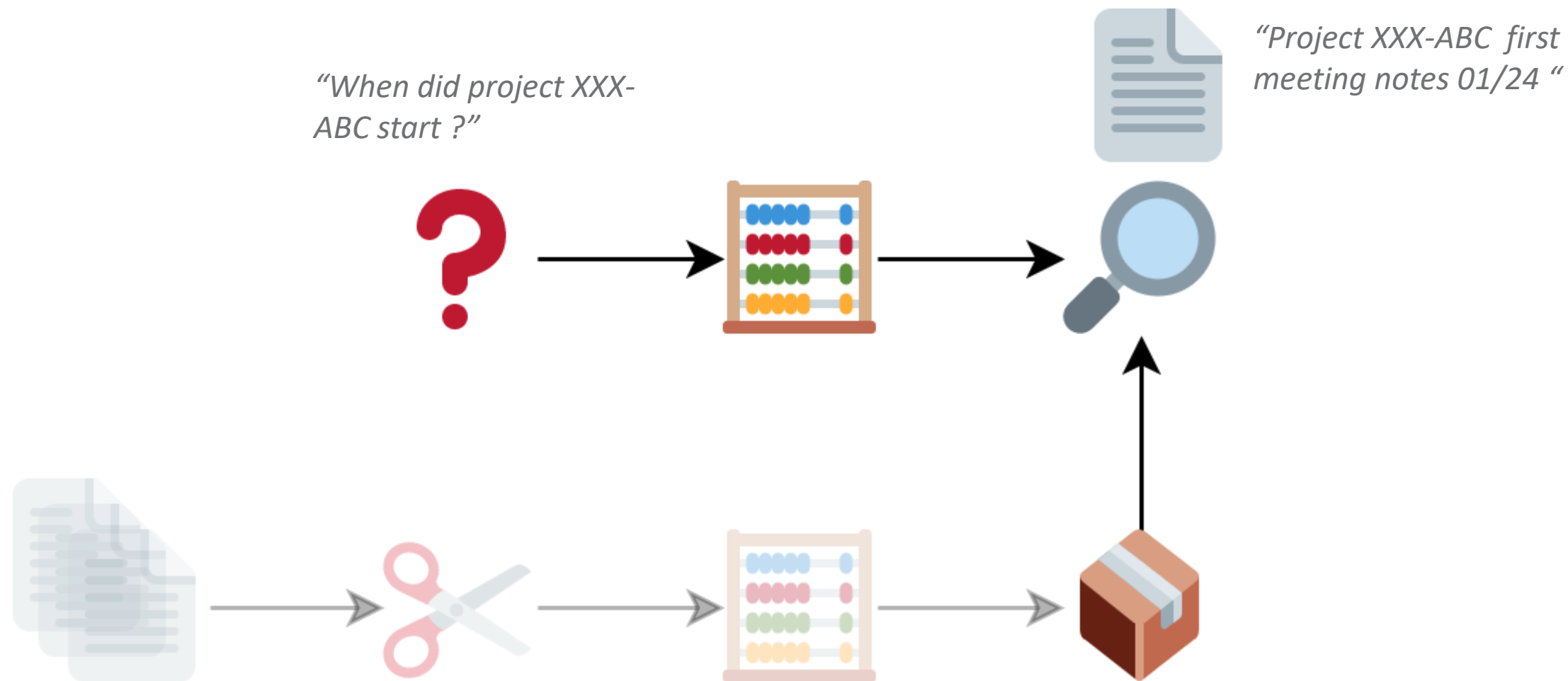
...	<i>cat</i>	<i>dog</i>	<i>Hello</i>	<i>switzerland</i>	<i>world</i>
	0	0	2	1	1

Turning documents into embeddings

- Synthetising a whole document into a single embedding is too broad
- Each document must first be sliced into **chunks** of text
- Those chunks can then be converted to embeddings
- Those embeddings can then be stored in dedicated database called **vectorstore**



Exploring the vectorstore



You answer the question “*when did project XXX-ABC start*” based on “*Project XXX-ABC first meeting notes 01/24*”



Questions ?

... now we are ready to code

<https://github.com/swiss-ai-center/workshop-rag>



Ressources

- <https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f>
- <https://humanloop.com/blog/prompt-engineering-101>