FONCTIONNEMENT DES MODÈLES DE LANGUE EXPLOITATION SUR DONNÉES ALIMENTAIRES

Lundi 29 septembre 2025
Séminaire ALIMining, IRIT, Toulouse

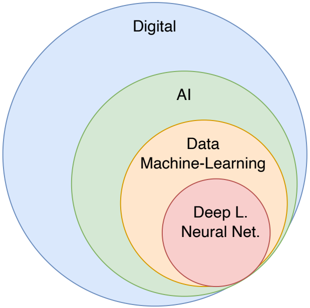
Vincent Guigue
https://vguigue.github.io

MIA PARIS-SACLAY EKINOCS

INRAE AgroParisTech université PARIS-SACLAY

Introduction

Artificial Intelligence & Machine Learning



Input (**X**)	Output (**Y**)	Application
email	spam? (0/1)	spam filtering
audio	text transcript	speech recognition
English	Chinese	machine translation
ad, user info	click? (0/1)	online advertising
image, radar info	position of other cars	self-driving car
image of phone	defect? (0/1)	visual inspection

AI: computer programs that engage in tasks which are, for now, performed more satisfactorily by human beings because they require high-level mental processes.

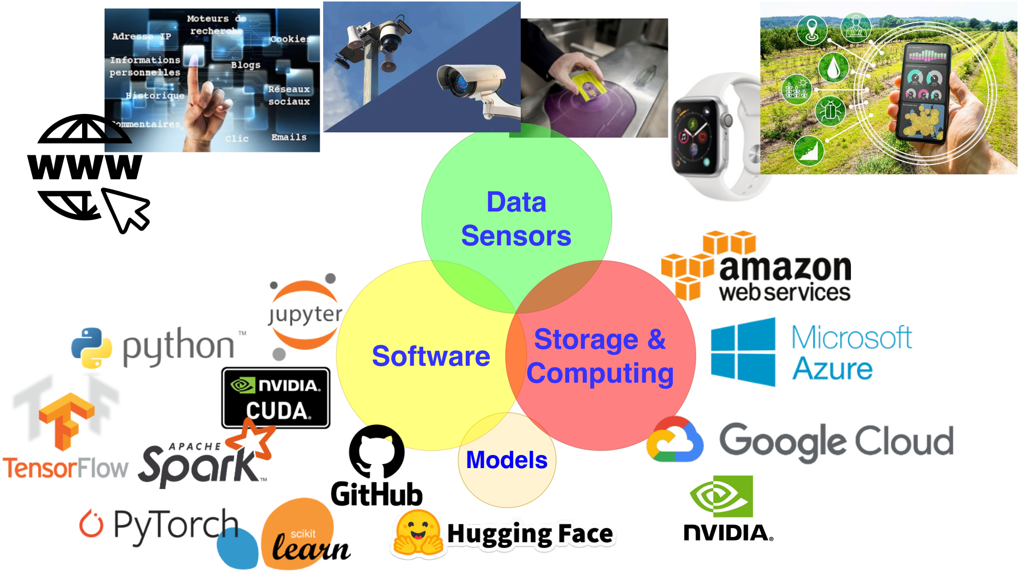
Marvin Lee Minsky, 1956

N-AI (Narrow Artificial Intelligence), dedicated to a single task

$\neq$ **G-AI (General AI)**, which replaces humans in complex systems.

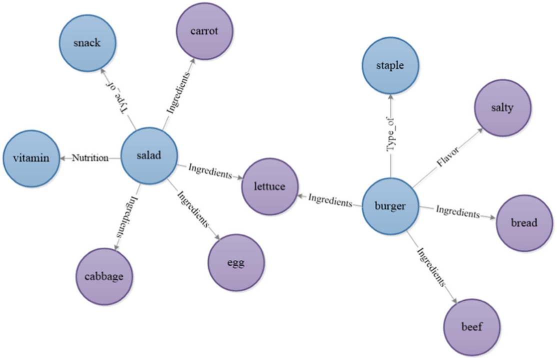
Andrew Ng, 2015

The Ingredients of Machine-Learning



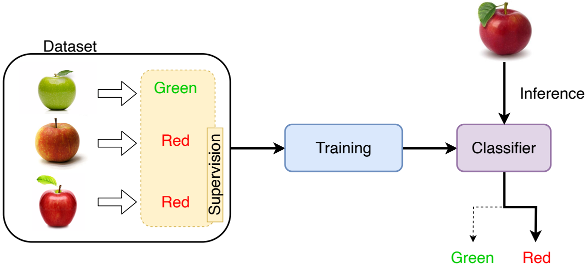
Machine-Learning vs Expert Knowledge

Modeling Expert Knowledge



A relationship extraction method for domain knowledge graph construction, Yu et al. 2020

Machine Learning

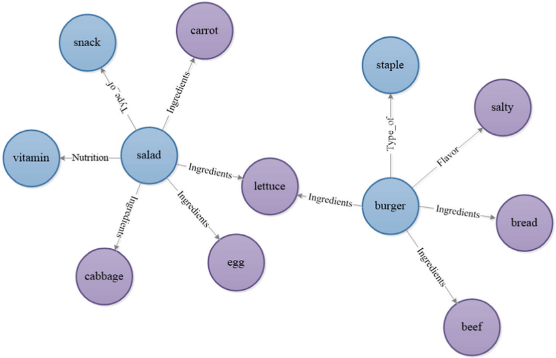


Different behaviors:

different strengths and weaknesses, different costs & requirements

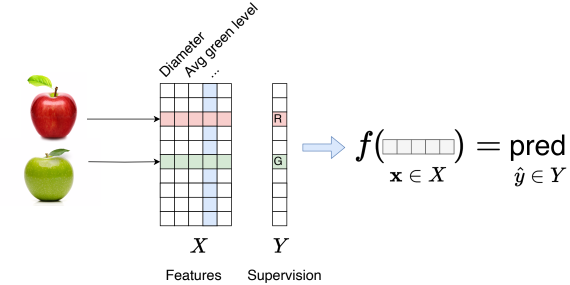
Machine-Learning vs Expert Knowledge

Modeling Expert Knowledge



A relationship extraction method for domain knowledge graph construction, Yu et al. 2020

Machine Learning



Different behaviors:

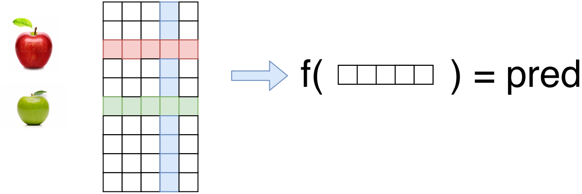
different strengths and weaknesses, different costs & requirements

Deep learning & Representation Learning

[Application to textual data]

From tabular data to text

→ Tabular data
- → Fixed dimension
- → Continuous values



→ Textual data
- → Variable length
- → Discrete values

this new iPhone, what a marvel

An iPhone? What a scam!

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Linguistics [1960-2010]

Rule-based Systems:

* ⟶ {like, love, appreciate} ⟶ * ⟶ #product

* ⟶ {didn't, not, doesn't, don't} ⟶ {like, love, appreciate} ⟶ * ⟶ #product

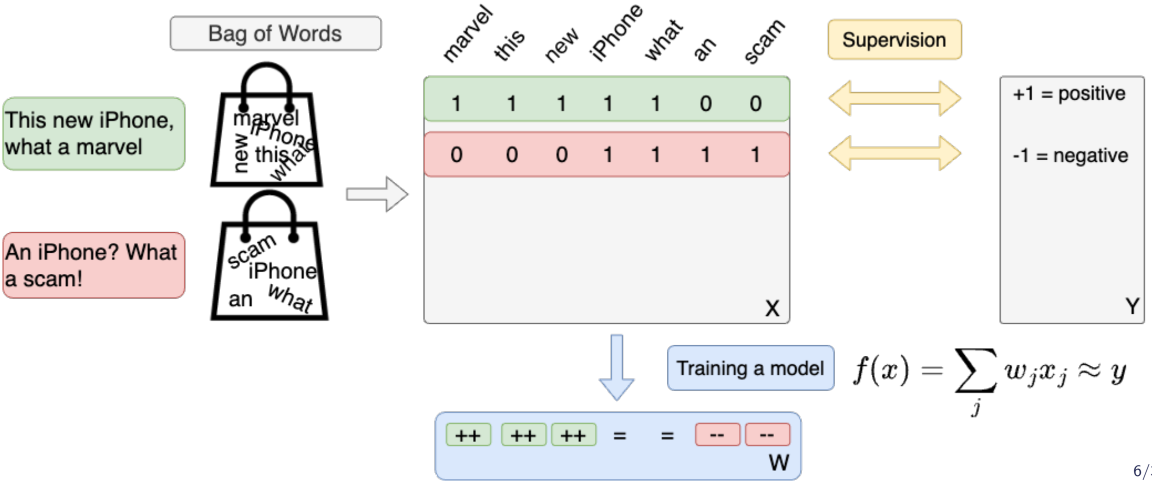
* ⟶ {hate, loathe, detest} ⟶ * ⟶ #product

- Requires expert knowledge
- Rule extraction ⇔ very clean data
- Very high precision
- Low recall
- Interpretable system

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Machine Learning [1990-2015]



$$f(x) = \sum_j w_j x_j \approx y$$

AI + Textual Data: Natural Language Processing (NLP)

NLP = largest scientific community in AI

Linguistics [1960-2010]

- Requires expert knowledge
- Rule extraction ⇔ very clean data

+ Interpretable system
+ Very high precision
− Low recall

Machine Learning [1990-2015]

- Little expert knowledge needed
- Statistical extraction ⇔ robust to noisy data

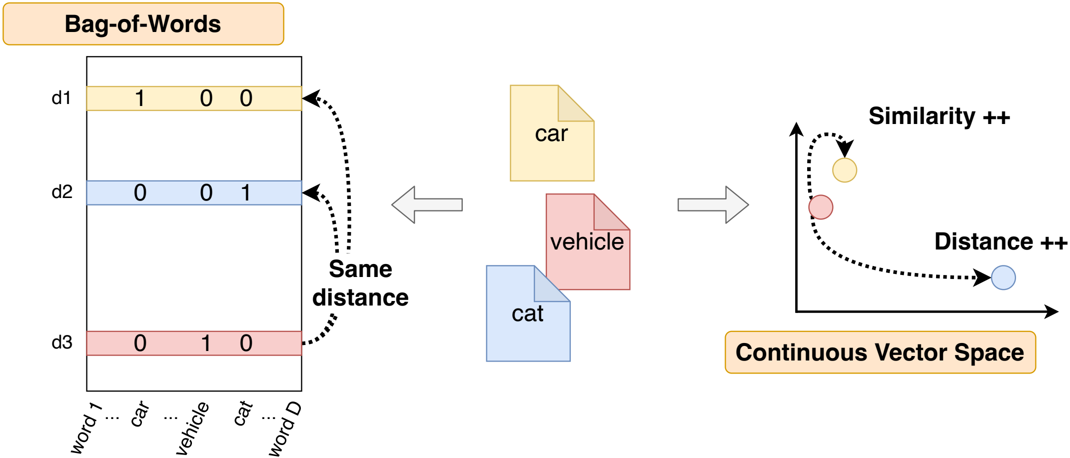
≈ Less interpretable system
− Lower precision
+ Better recall

Precision = criterion for acceptance by industry

→ Link to metrics

Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]

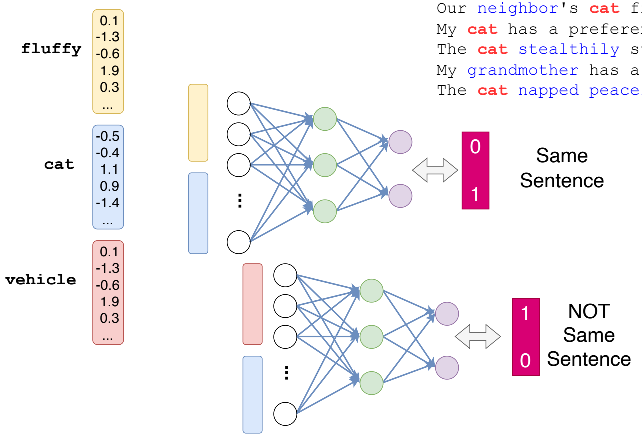


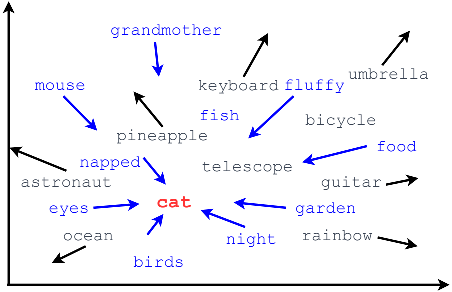
LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]

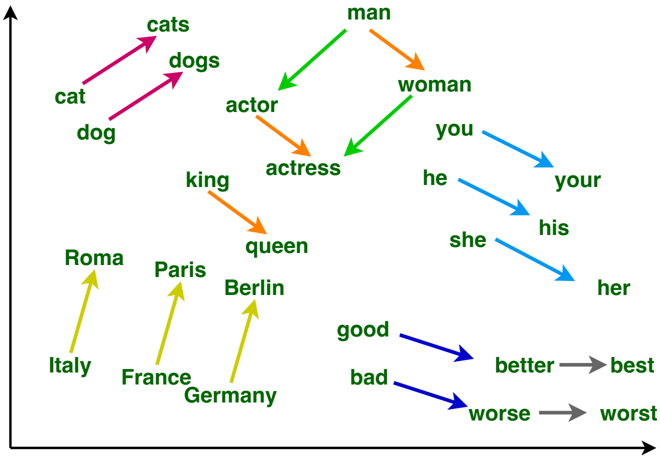
The fluffy cat napped lazily in the sunbeam.
I adopted a stray cat from the shelter last week.
My cat loves to chase after toy mice.
The black cat stealthily crept through the dark alley.
I often find my cat perched on the windowsill, watching birds.
She gently stroked her cat's fur as it purred contentedly.
Our neighbor's cat frequently visits our backyard.
My cat has a preference for fish flavored cat food.
The cat stealthily stalked a mouse in the garden.
My grandmother has a collection of porcelain cat figurines.
The cat napped peacefully in the warm sunlight.





Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]



- Semantic Space:

  similar meanings

  $\Leftrightarrow$

  close positions

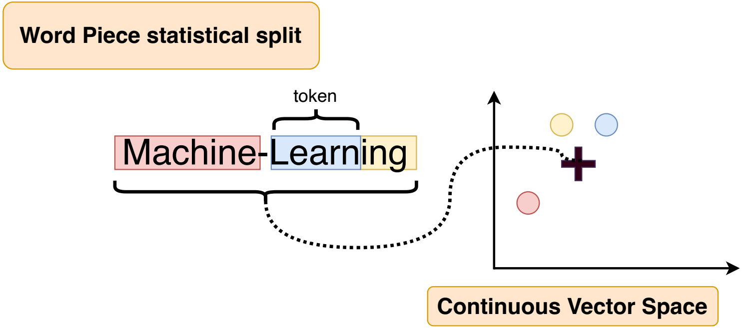
- Structured Space: grammatical regularities, basic knowledge, ...

Distributed representations of words and phrases and their compositionality, Mikolov et al. NeurIPS 2013

Deep/Representation Learning for Text Data

From Bag of Words to Vector Representations [2008, 2013, 2016]

From Words to Tokens

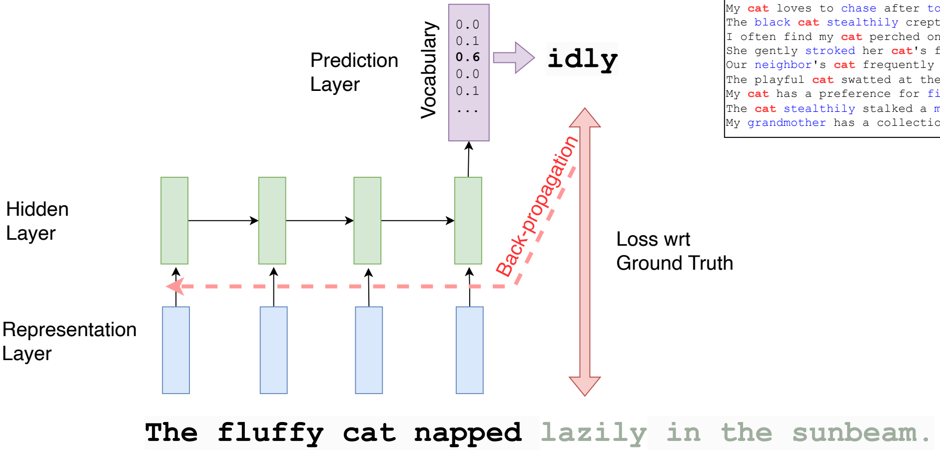


- Representation of unknown words
- Adaptation to technical domains
- Resistance to spelling errors

Enriching word vectors with subword information. Bojanowski et al. TACL 2017.

Aggregating word representations: towards generative AI

- Generation & Representation
- New way of learning word positions




```
The fluffy cat napped lazily in the sunbeam.
I adopted a stray cat from the shelter last week.
My cat loves to chase after toy mice.
The black cat stealthily crept through the dark alley.
I often find my cat perched on the windowsill, watching birds.
She gently stroked her cat's fur as it purred contentedly.
Our neighbor's cat frequently visits our backyard.
The playful cat swatted at the dangling string with its paw.
My cat has a preference for fish flavored cat food.
The cat stealthily stalked a mouse in the garden.
My grandmother has a collection of porcelain cat figurines.
```

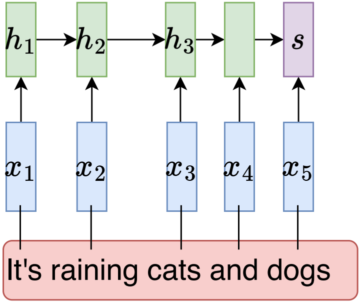
Corpus

Sequence to Sequence Learning with Neural Networks, Sutskever et al. NeurIPS 2014

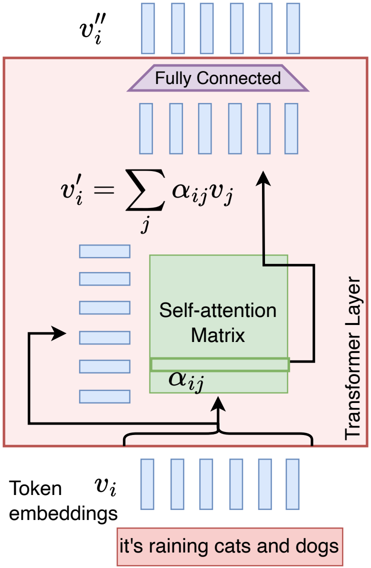
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:



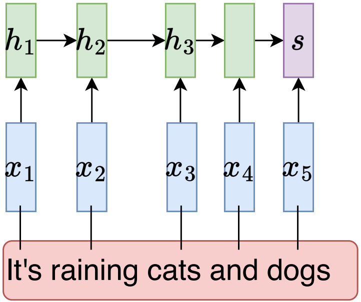
Attention is all you need, Vaswani et al. NeurIPS 2017

Sequence to Sequence Learning with Neural Networks, Sutskever et al. NeurIPS 2014

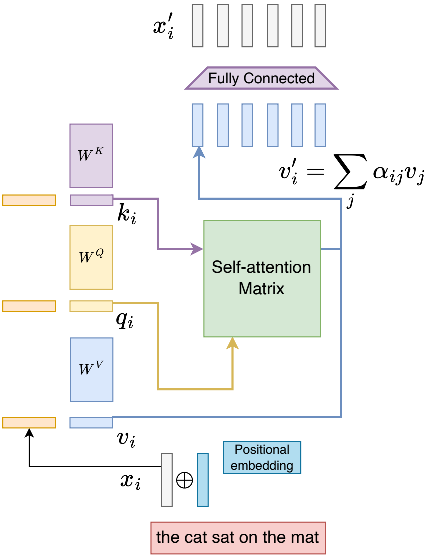
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



Transformer:



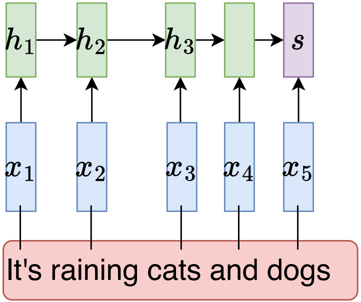
Attention is all you need, Vaswani et al. NeurIPS 2017

Sequence to Sequence Learning with Neural Networks, Sutskever et al. NeurIPS 2014

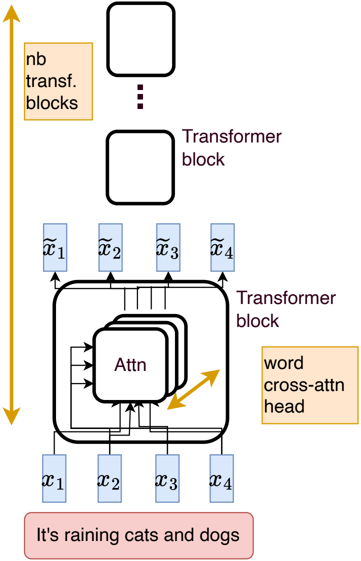
Transformer architecture: state-of-the-art aggregation

Recurrent Neural Network:

$$h_{t+1} = h_t W_1 + x_{t+1} W_2$$



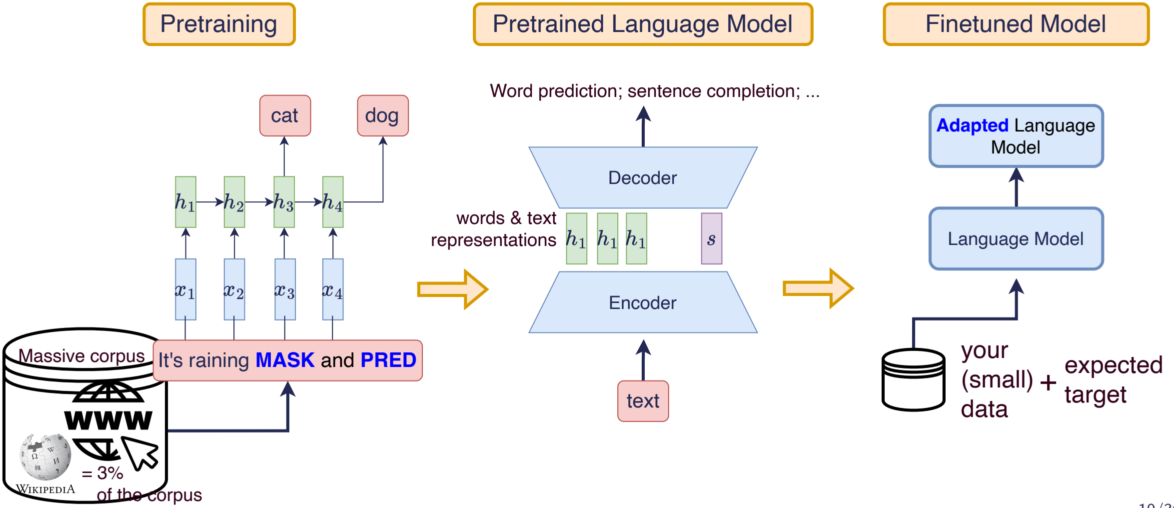
Transformer:



Attention is all you need, Vaswani et al. NeurIPS 2017

A new developpement paradigm since 2015

- Huge dataset + huge archi. ⇒ unreasonable training cost
- Pre-trained architecture + 0-shot / finetuning



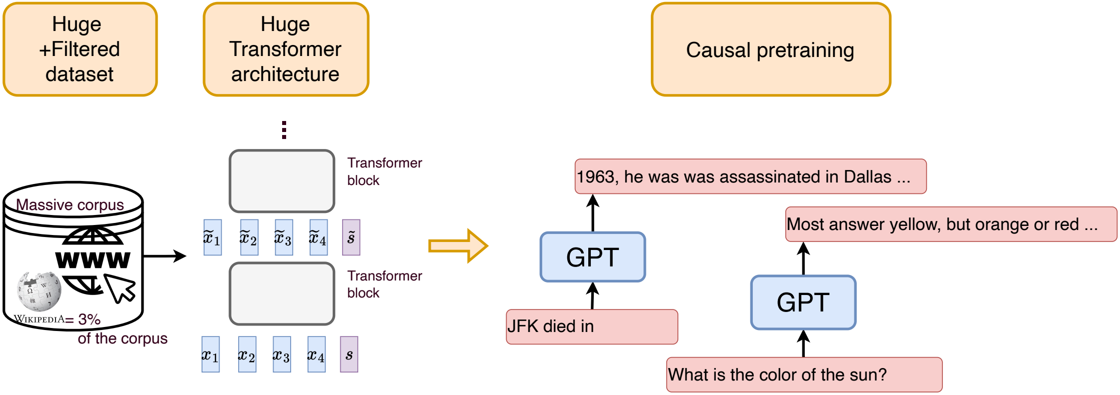
chatGPT

November 30, 2022

1 million users in 5 days
100 million by the end of January 2023
1.16 billion by March 2023

The Ingredients of chatGPT

0. Transformer + massive data (GPT)



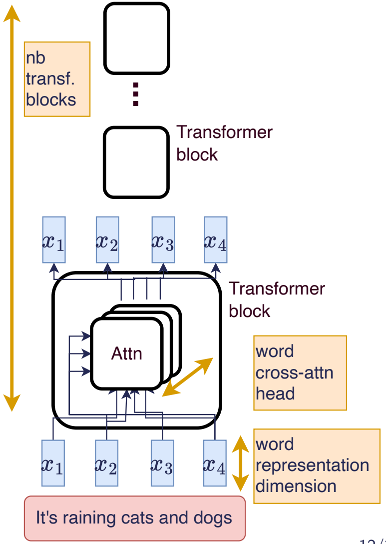
- Grammatical skills: singular/plural agreement, tense concordance
- (Parametric) Knowledge: entities, names, dates, places

Language Models are Few-Shot Learners, Brown et al. 2020

The Ingredients of chatGPT

1. More is better! (GPT)

- + more input words [500 ⇒ 2k, 32k, 100k]
- + more dimensions in the word space [500-2k ⇒ 12k]
- + more attention heads [12 ⇒ 96]
- + more blocks/layers [5-12 ⇒ 96]

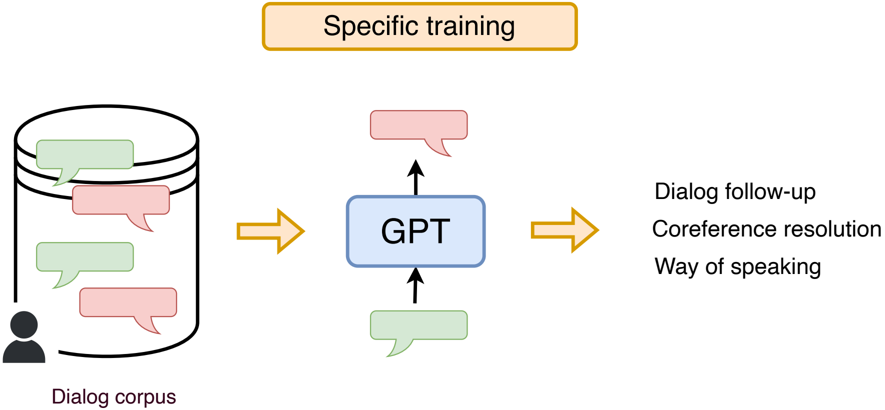


175 Billion parameters... What does it mean?

- $1.75 \cdot 10^{11} \Rightarrow 300$ GB + 100 GB (data storage for inference) $\approx$ 400GB
- NVidia A100 GPU = 80GB of memory (=20k€)
- Cost for (1) training: 4.6 Million €

The Ingredients of chatGPT

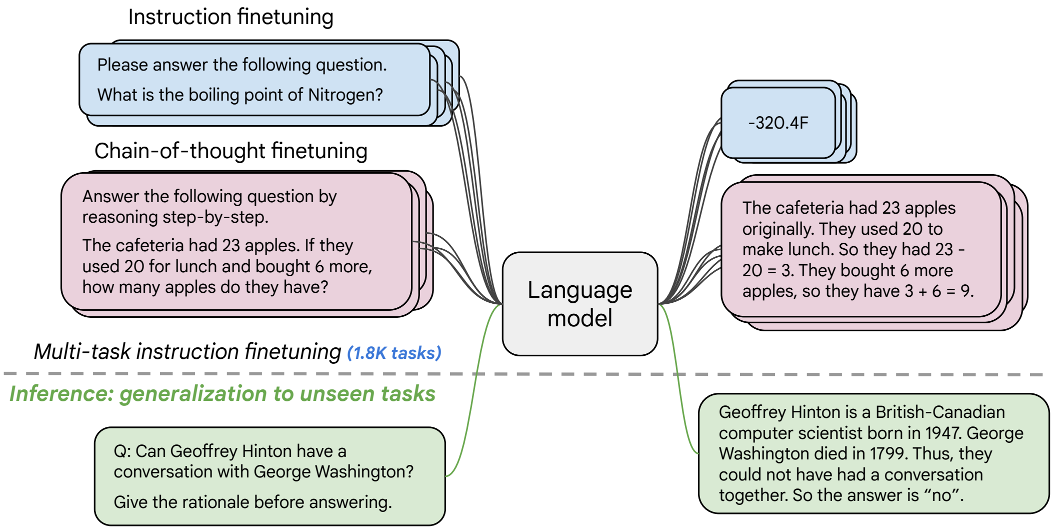
2. Dialogue Tracking



- **Very clean** data — Data generated/validated/ranked by humans

The Ingredients of chatGPT

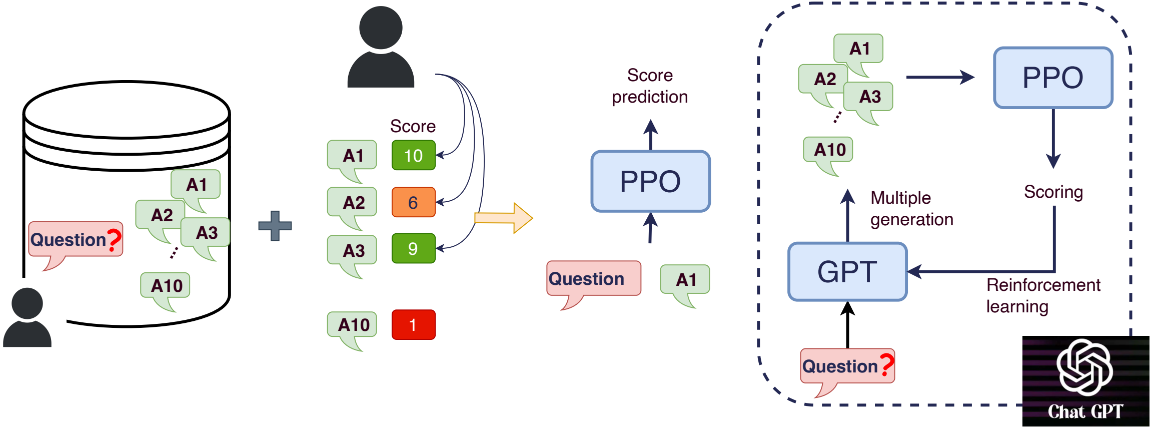
3. Fine-tuning on different (±) complex reasoning tasks



Scaling Instruction-Finetuned Language Models, Chung et al., JMLR 2024

The Ingredients of chatGPT

4. Instructions + answer ranking



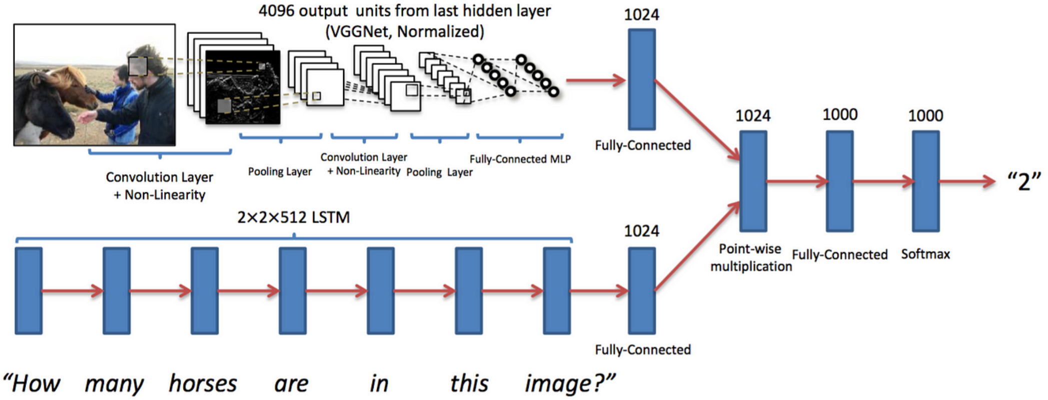
- Database created by humans
- Response improvement
- ... Also a way to avoid critical topics = censorship

Training language models to follow instructions with human feedback, Ouyang et al., 2022

GPT4 & Multimodality

Merging information from text & image. **Learning** to exploit information jointly

The example of VQA: visual question answering

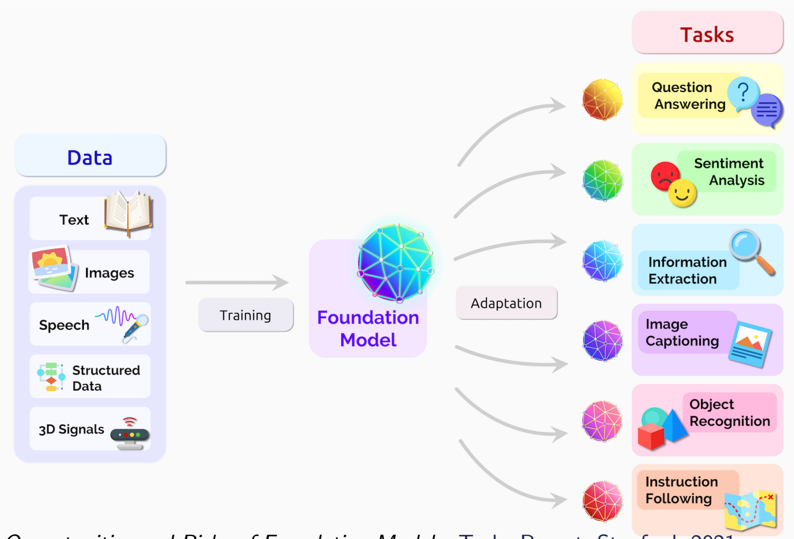


⇒ Backpropagate the error ⇒ modify word representations + image analysis

VQA: Visual Question Answering, arXiv, 2016 , A. Agrawal et al.

Towards Larger Foundation Models?

- Let the modalities enrich each other

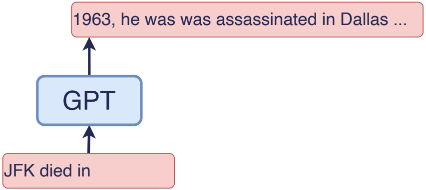


On the Opportunities and Risks of Foundation Models, Tech. Report, Stanford, 2021
Bommasani et al.

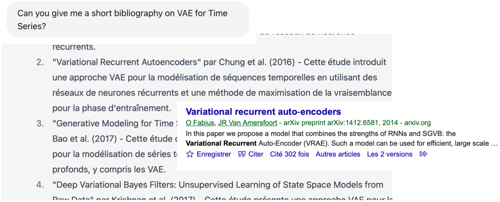
Machine Learning Limits

chatGPT and the relationship with truth

1. **Likelyhood** = grammar, agreement, tense concordance, logical sequences...
   $\Rightarrow$ Repeated knowledge
2. Predict the most **plausible** word...
   $\Rightarrow$ produces **hallucinations**
3. **Offline** functioning
4. chatGPT $\neq$ **knowledge graphs**
5. Brilliant answers...
   And silly mistakes!
   + we cannot predict the errors



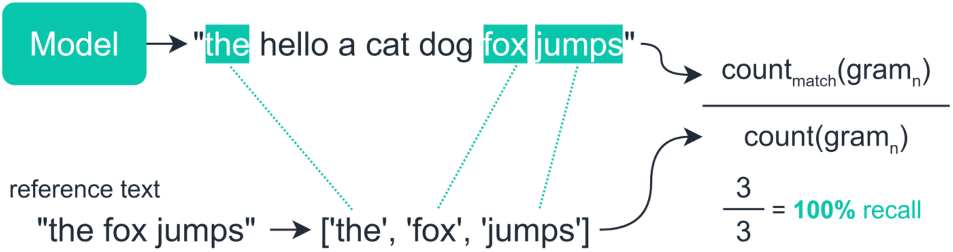
Example: producing a bibliography



Generative AI: how to evaluate performance?

The critical point today

- How to evaluate against ground truth?
- How to evaluate system confidence / plausibility of generation?

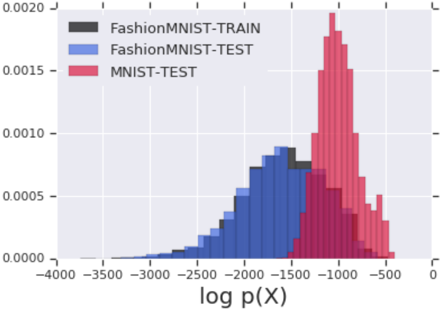


The Ultimate Performance Metric in NLP, J. Briggs, Medium 2021

Generative AI: how to evaluate performance?

The critical point today

- How to evaluate against ground truth?
- How to evaluate system confidence / plausibility of generation?



Plausibility

Train

Test

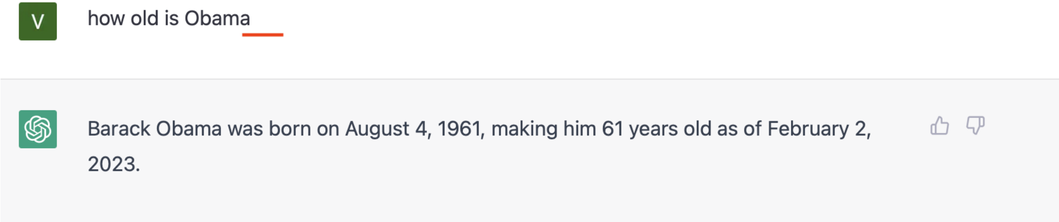
Do Large Language Models Know What They Don't Know?, Yin et al. , ACL, 2023

Do Deep Generative Models Know What They Don't Know?, Nalisnick et al. , ICLR, 2019

Stability/predictability

- Difficult to bound a behavior
- Impossible to predict good/bad answers

⇒ Little/no use in video games

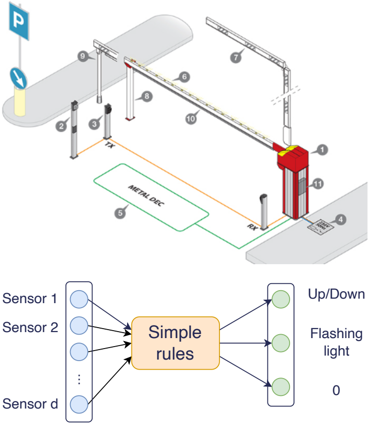


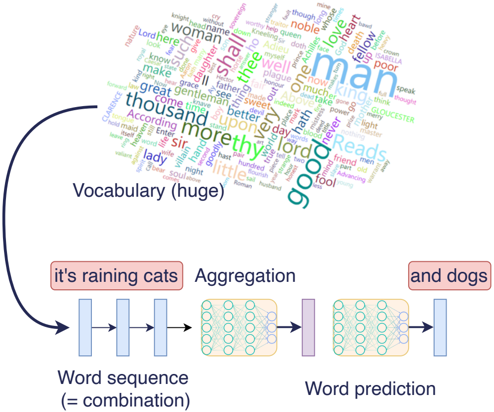
Stability/predictability

- Difficult to bound a behavior
- Impossible to predict good/bad answers

⇒ Little/no use in video games

Explainability... And complexity





- Simple system
- Exhaustive testing of inputs/outputs
- **Predictable & explainable**

- Large dimension
- Complex non-linear combinations
- **Non-predictable & non-explainable**

Explainability... And complexity

Interpretability *vs* Post-hoc Explanation

Neural networks = **non-interpretable** (almost always)

too many combinations to anticipate

Neural networks = **explainable a posteriori** (almost always)

[Uber Accident, 2018]

- Simple system
- Exhaustive testing of inputs/outputs
- **Predictable** & **explainable**

- Large dimension
- Complex non-linear combinations
- **Non-predictable** & **non-explainable**

Transparency : open source / open weight

- Can I modify it? — **Adaptation**
- What training data was used? — **Data contamination / skills**
- What editorial stance / censorship is involved? — **Access to information**
- Why this answer? — **Explainability / interpretability**

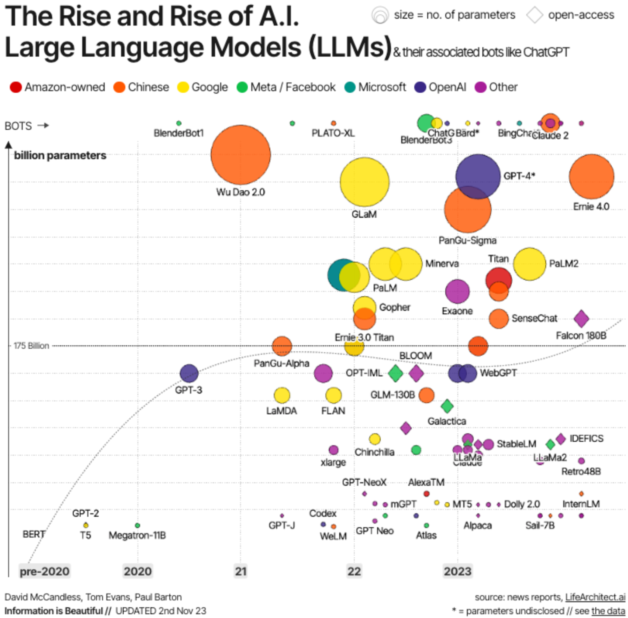
Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta Llama 2	BigScience BLOOMZ	OpenAI GPT-4	stability.ai Stable Diffusion 2	Google PaLM 2	ANTHROPIC Claude 2	cohere Command	AI21labs Jurassic-2	Inflection Inflection-1	amazon Titan Text	Average
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	57%	52%	47%	47%	41%	39%	31%	20%	20%	13%	

https://crfm.stanford.edu/fmti/May-2024/index.html

Costs / Frugality



Parameters

1998	LeNet-5	= 0.06M
2011	Senna	= 7.3M
2012	AlexNet	= 60M
2017	Transformer	= 65M / 210M
2018	ELMo	= 94M
2018	BERT	= 110M / 340M
2019	GPT2	= 1,500M
2020	GPT3	= 175,000M
2025	Llama-4	= 2,000,000M

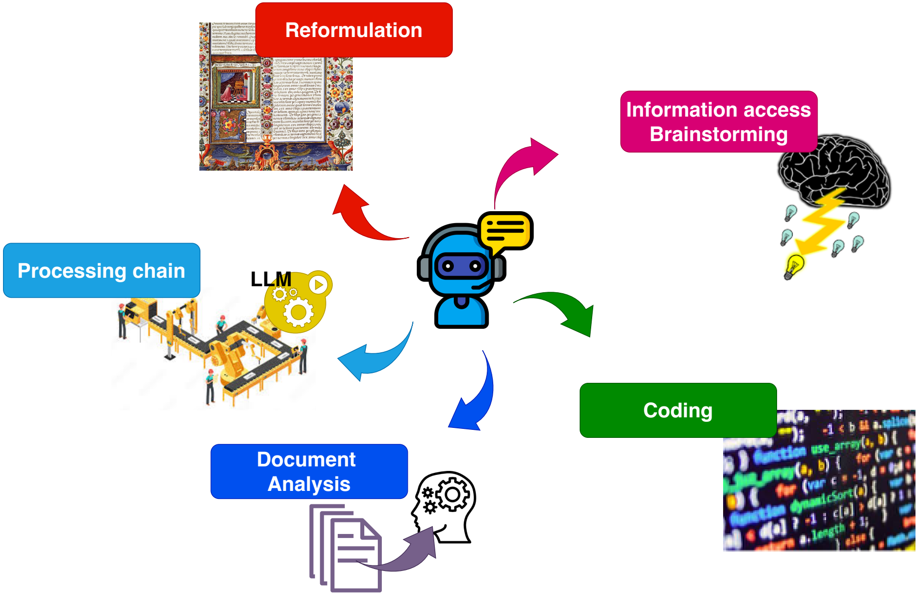
Everything beyond the LLM's capabilities/training

- Simple calculations (multiplication, division)
- Generating *n*-syllable animal names (in progress)
- Playing chess
- Follow (complex) causal reasoning
- ...

Large Language Models Uses

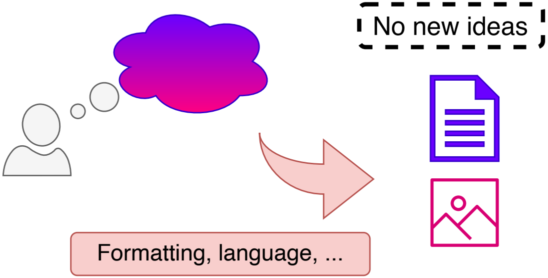
[in nutrition research]

Key uses in 5 pictures



(1) Formatting information

A fantastic tool for **formatting**



- Personal assistant
  - Standard letters, recommendation letters, cover letters, termination letters
  - Translations
- Meeting reports
  - Formatting notes
- Writing scientific articles
  - Writing ideas, in French, in English

$\Rightarrow$ No new information, just writting, cleaning up, ...

(1) Nutrition use : Input standardization (?)

⇒ opportunity to fuse heterogeneous information

Food.

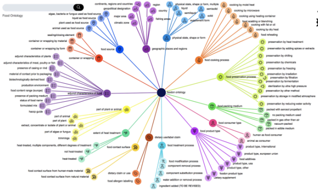
INCA2 Individual national study of food consumption in France Database

anses
agence nationale de sécurité sanitaire
alimentation, environnement, travail

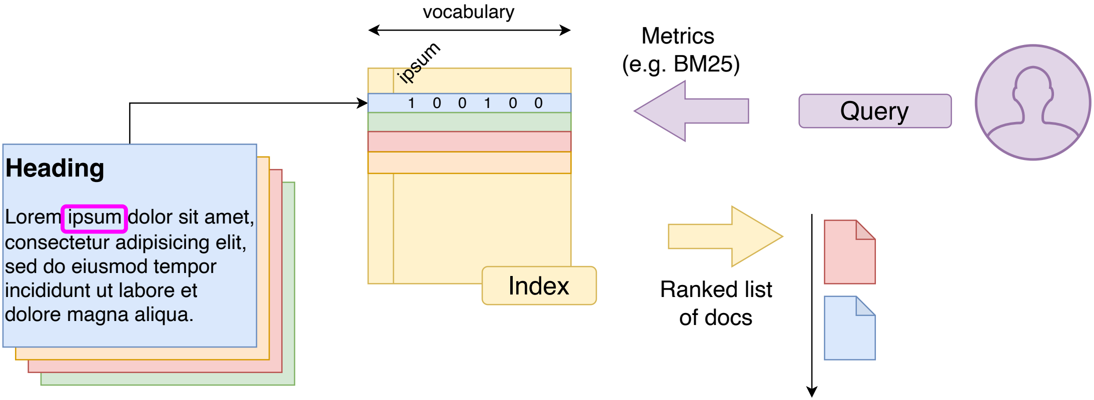
ARTICLE OPEN

FoodOn: a harmonized food ontology
traceability, quality control and data in

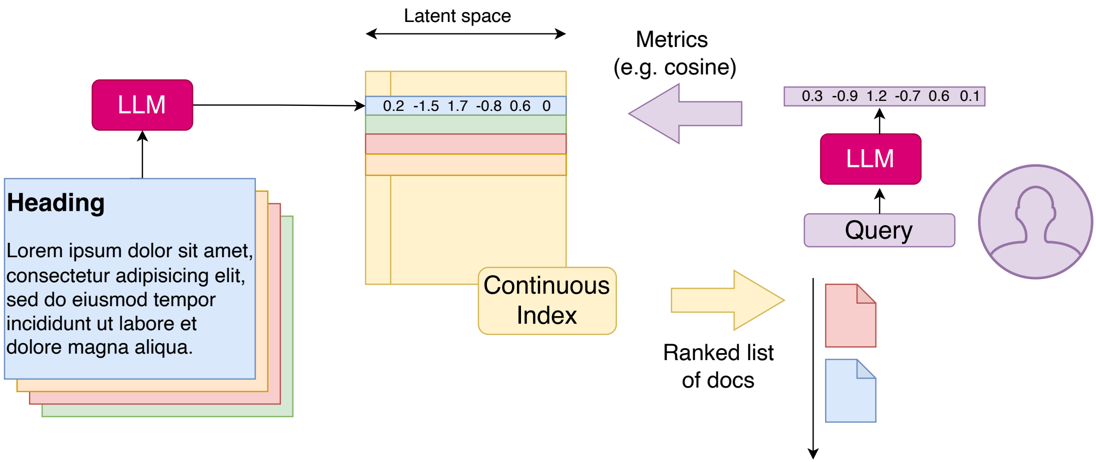
J. Griffiths[2,8], Gurinder S. Gosal[1], Pier L. Buttigieg[3], Ro
Brinkman[2] and William W. L. Hsiao[1,2,7]



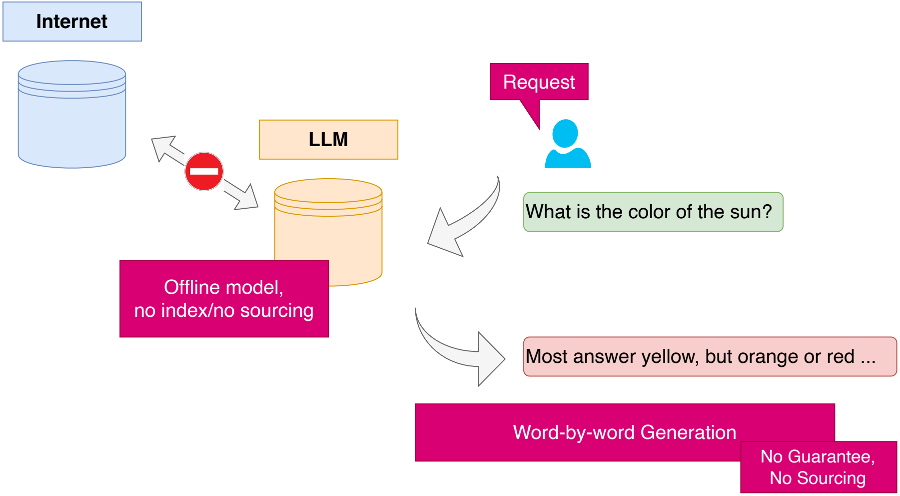
(1) Chat & RAG : a new way to access information



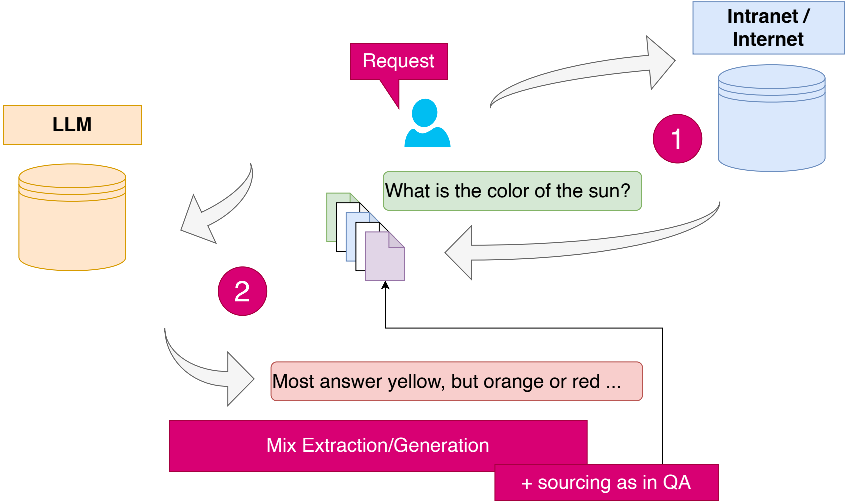
(1) Chat & RAG : a new way to access information



(1) Chat & RAG : a new way to access information



(1) Chat & RAG : a new way to access information

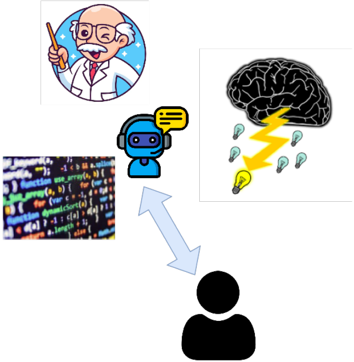


$\Rightarrow$ A way to build a *reliable* chatbot to advise users?

- Parametric memory *vs* Information Retrieval

(2) Brainstorming / Course Planning / Statistics Review

- **Find** inspiration [writer's block syndrome]
- **Organize** ideas quickly
- **Avoid omissions** / increase confidency
- **Search** in a targeted way, adapted to one's needs



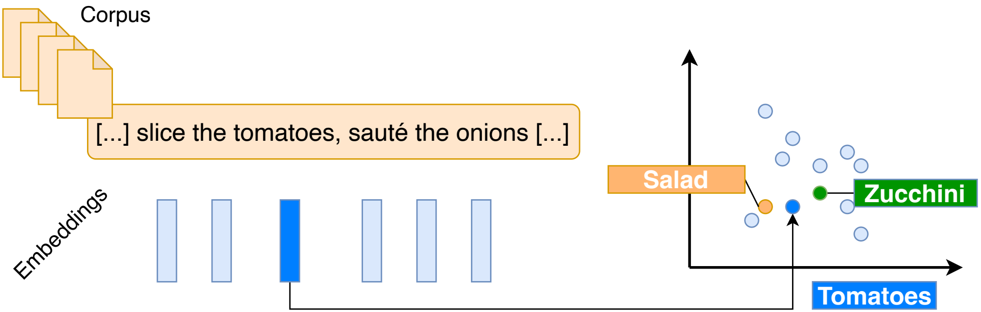
$\Rightarrow$ Impressive answers, sometimes incomplete or partially incorrect... But often useful

3 reference articles on the use of transformers in recommendation systems
What is the purpose of the log-normal Poisson law?
Propose 10 sections for a course on Transformers in AI

- In which areas are LLMs reliable?
- What are the risks for primary information sources?
- What societal risks for information?

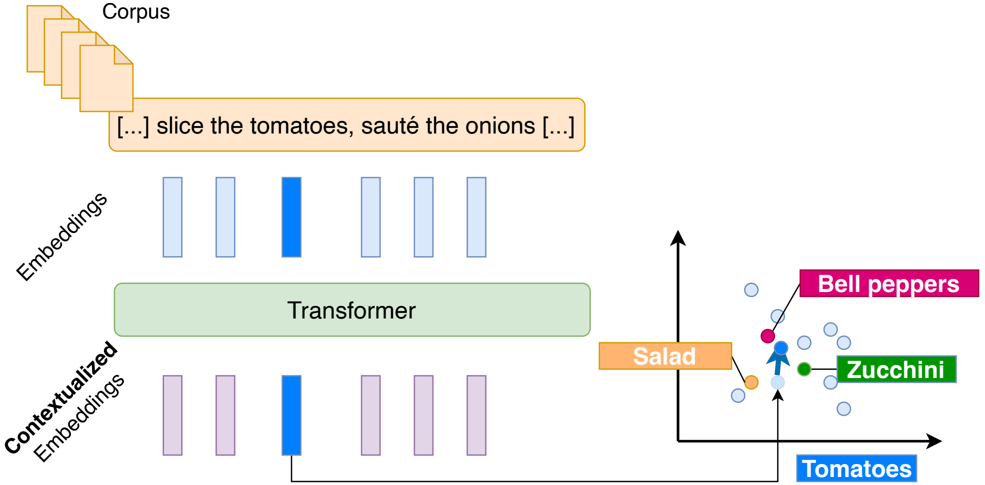
(2) Internal knowledge exploitation for nutrition

- Brainstorming in the kitchen: which application for cooking?
- Ingredient substitution... At every scale: Ingredient, Food, Dish



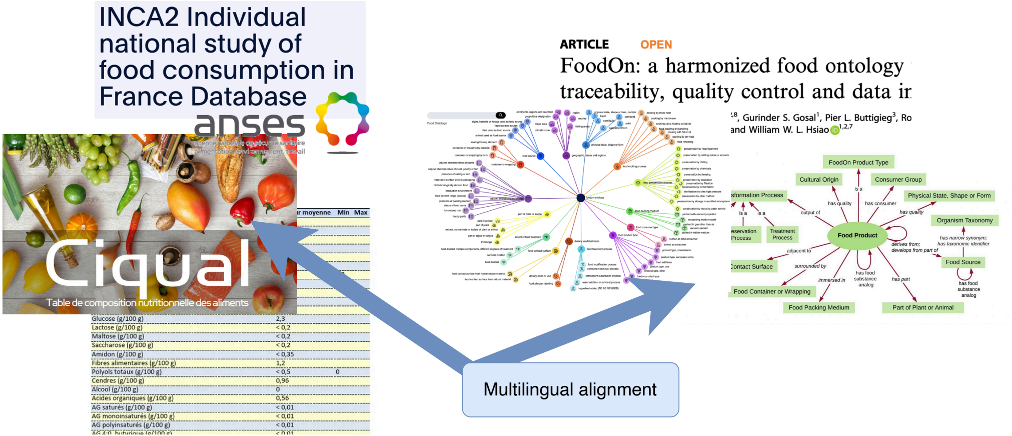
(2) Internal knowledge exploitation for nutrition

- Brainstorming in the kitchen: which application for cooking?
- Ingredient substitution... At every scale: Ingredient, Food, Dish
- ++ Upgrade by contextualization



(2) Internal knowledge exploitation for nutrition

- Brainstorming in the kitchen: which application for cooking?
- Ingredient substitution... At every scale: Ingredient, Food, Dish
- ++ Upgrade by contextualization
- Interoperability and ontologies



(2) Internal knowledge exploitation for nutrition

- Brainstorming in the kitchen: which application for cooking?
- Ingredient substitution... At every scale: Ingredient, Food, Dish
- ++ Upgrade by contextualization
- Interoperability and ontologies

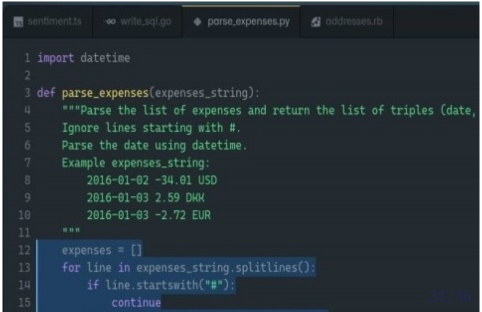
(3) Coding: Different Tools, Different Levels

- Providing solutions to exercises
- Learning to code or getting back into it
  - New languages, new approaches (ML?)
  - Benefit from explanations...

But how to handle mistakes?
- Help with a library [*getting started*]
- Faster coding

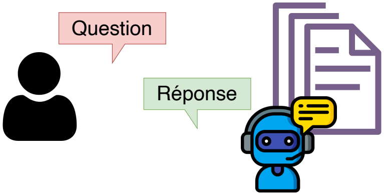
- What about copyrights?
  - What impact on future code processing?
- How to adapt teaching methods?
- How many calls are needed for code completion?

What about the carbon footprint?
- What is the risk of error propagation?



(4) Document Analysis

- Summarizing documents / articles
- Dialoguing with a document database
- Assistance in writing reviews
- FAQs, internal support services within companies
- Technology watch
- Generating quizzes from lecture notes



- Will articles still be read in the future?
  - Should we make our articles NotebookLM-proof?
- How to save time while remaining honest and ethical?

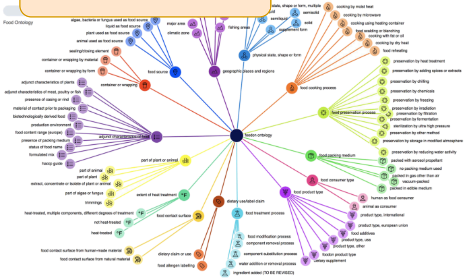
(4) Information Extraction in Nutrition

- Ontology building (mostly textual data)

Raw & noisy data → Validated ontology

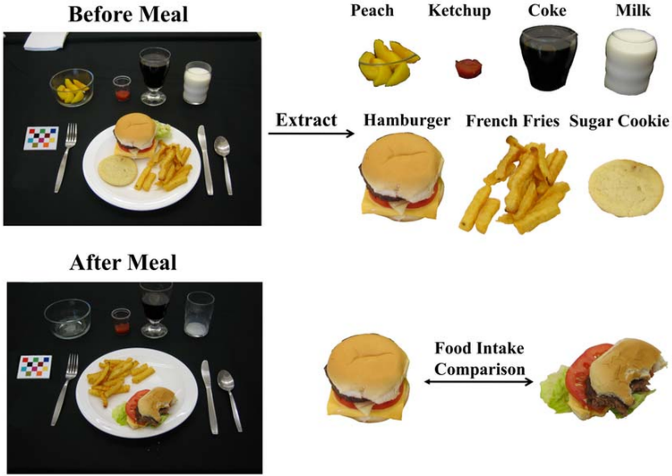
marmiton

Food.



(4) Information Extraction in Nutrition

- Ontology building (mostly textual data)
- Image analysis

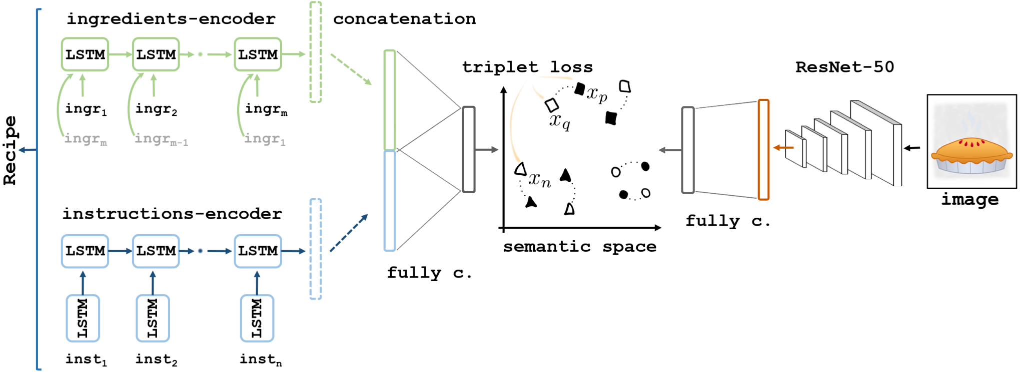


- Food recognition
- Segmentation
- Estimation of quantities

An Overview of The Technology Assisted Dietary Assessment Project at Purdue University., Khanna et al. , 2010

(4) Information Extraction in Nutrition

- Ontology building (mostly textual data)
- Image analysis
- Multimodal analysis + algorithmic process



Images & Recipes: Retrieval in the cooking context, SIGIR 2018
Carvalho et al.

(4) Information Extraction in Nutrition

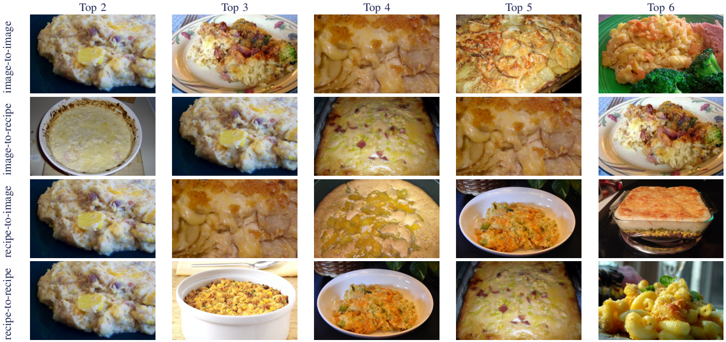
- Ontology building (mostly textual data)
- Image analysis
- Multimodal analysis + algorithmic process

	ingr (ingredients)	**instr (cooking instructions)**	**image**
Pizza	1) *pizza dough* 2) *hummus* 3) *arugula* 4) *cherry / grape tomatoes* 5) *pitted greek olives* 6) *crumbled feta cheese* … …	1) *Cut the dough into two 8-ounce sized pieces.* 2) *Roll the ends under to create round balls.* 3) *Then using a well-floured rolling pin, roll the dough out into 12-inch circles.* 4) *Place the dough circles on sheets of parchment paper.* … …	
Pecan Pie	1) *unsalted butter* 2) *eggs* 3) *condensed milk* 4) *sugar* 5) *vanilla extract* 6) *chopped pecans* 7) *chocolate chips* … …	1) *Preheat the oven to 375 degrees F.* 2) *In a large bowl, whisk together the melted butter and eggs until combined.* 3) *Whisk in the sweetened condensed milk, sugar, vanilla, pecans, chocolate chips, butterscotch chips, and coconut.* … …	

Images & Recipes: Retrieval in the cooking context, SIGIR 2018
Carvalho et al.

(4) Information Extraction in Nutrition

- Ontology building (mostly textual data)
- Image analysis
- Multimodal analysis + algorithmic process

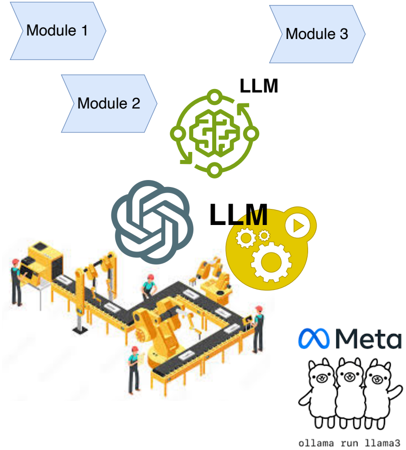


Images & Recipes: Retrieval in the cooking context, SIGIR 2018
Carvalho et al.

(5) LLM in a Production Pipeline / Agentic AI

- Run LLM locally
- Extract knowledge
- Sort documents / generate summaries
- Generate examples to train a model
  [Teacher/student - distillation]
- Generate variants of examples $\nearrow\nearrow$ increase dataset size
  [Data augmentation]

$\Rightarrow$ Integrate the LLM into a processing pipeline
= little/less supervision = **Agentic AI**



- Can I train models on generated data?
- How much does it cost? ($ + CO_2) Need for GPUs?
- How good are open-weight models?

(5) LLM in a Production Pipeline / Agentic AI

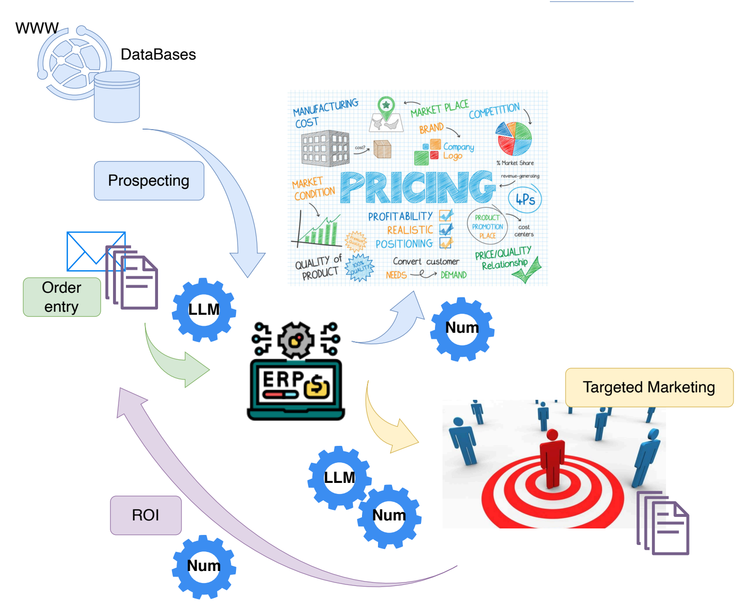
Module 3

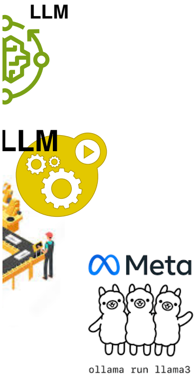
- Run LLM
- Extract kr
- Sort docu
- Generate e
- Generate v
dataset siz

⇒ Integrate t

=

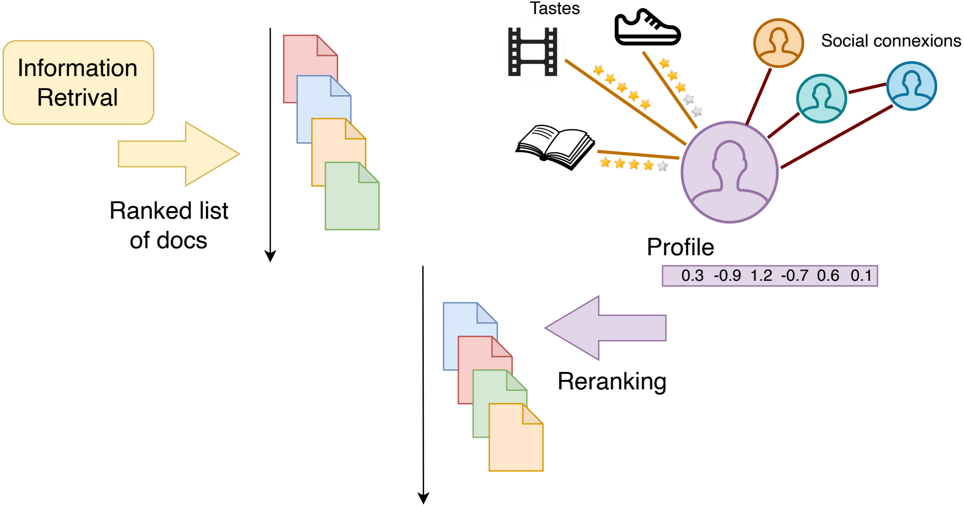
- Can I t
- How m
- How go





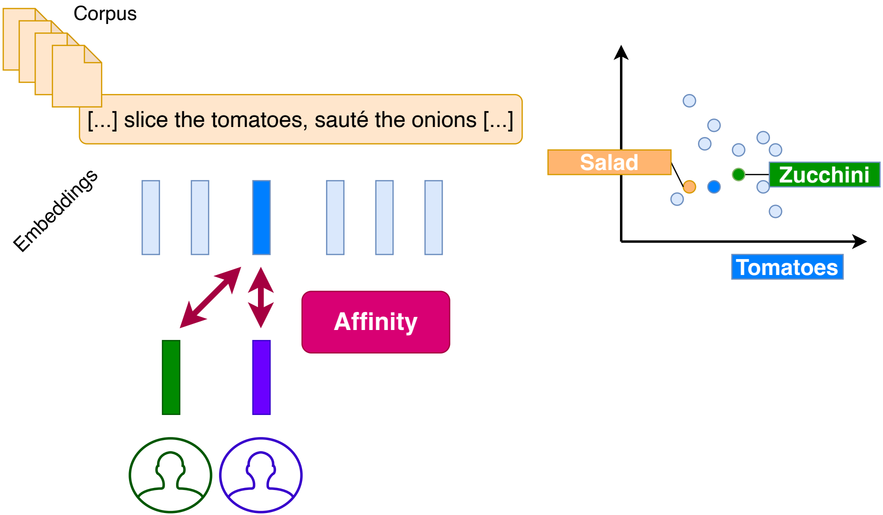
(5) What about RecSys in Nutrition?

Profiling is roughly everywhere in Information Retrieval



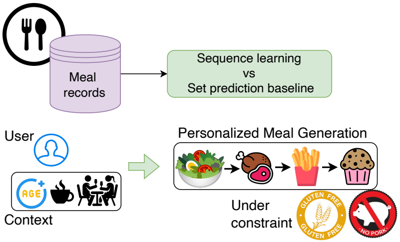
(5) What about RecSys in Nutrition?

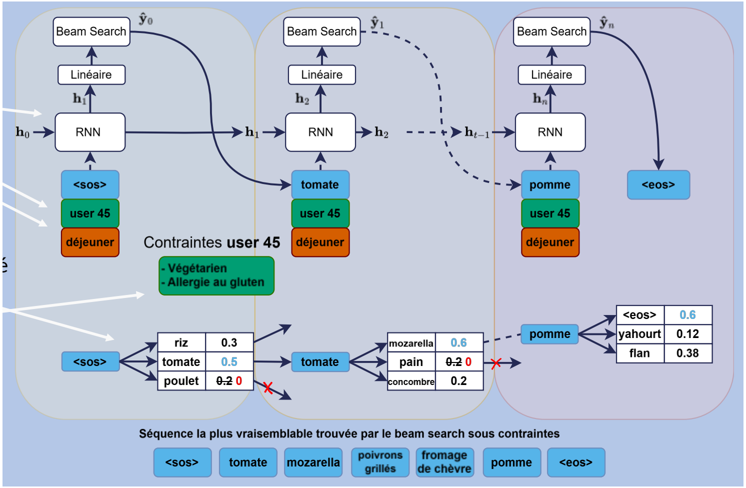
Opportunities in nutrition : modeling user preferences



(5) What about RecSys in Nutrition?

Building consistent proposals... With expert constraints





Génération séquentielle prenant en compte des informations contextuelles en nutrition , CAp 2025
Combeau et al.

Conclusion

New tools for new opportunities

LLMs offer new perspectives in nutrition:

- A natural and convenient interface for users
  - enabling dialogue, plate analysis, and personalized advice
- Accessible on multiple devices, from computers to smartphones and smart kitchens (Alexa, Google Assistant, ...)
- A means to unify and connect existing nutritional resources
- A powerful tool to extract and structure knowledge $\Rightarrow$ enrich databases
- A modular component for next-generation recommender systems