AI and Computing Challenges: Harnessing Large Language Models at Thomson Reuters

Adrian Alan Pol

Applied Machine Learning Scientist at **Thomson Reuters Labs**

17 June 2024

Thomson Reuters

What do we do?

Thomson Reuters is the most trusted provider of essential news, information, and tools for professionals in the legal, tax, accounting, compliance, government, and media markets.

Our customers rely on us to deliver the intelligence, technology, and human expertise they need to find trusted answers that inform their most important decisions.

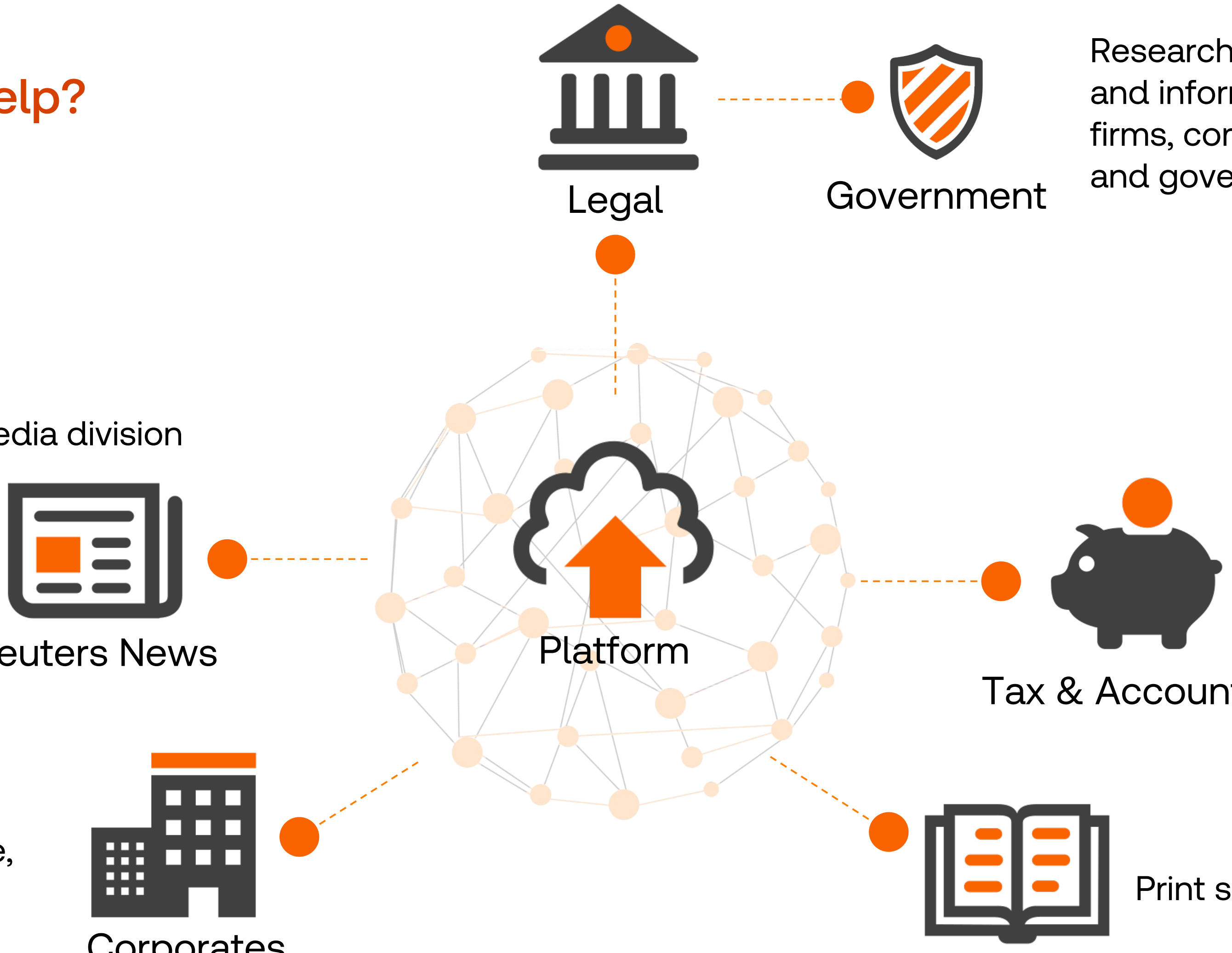
Operations in **100+ countries**

25,000+ employees

100+ years of business

Thomson Reuters

Who do we help?



Research platforms, legal software, and information services, to law firms, corporate legal departments, and government agencies

News and media division

Tax and accounting, risk management, regulatory compliance, and business intelligence

Print solutions

Thomson Reuters Labs™

AI at Thomson Reuters

30+ Years of AI at Thomson Reuters.

Thomson Reuters Labs™ is the innovation and applied research arm of Thomson Reuters, with almost 200 talented colleagues operating globally.

Through rapid prototyping of solutions and continuous knowledge sharing, we support our organization and customers with the understanding and application of new technologies to their businesses.

We work collaboratively across our core customer segments to identify, de-risk, and activate future-ready opportunities in AI, machine learning, data science and emerging technologies.

AI Applied Research Scientists
AI algorithm development

AI Applied Research Engineers
End-to-end AI/ML Lifecycle management

Human-Centered AI Design
AI Experience Strategy, Research & Design

Program Leads

FUNCTIONAL SKILLS

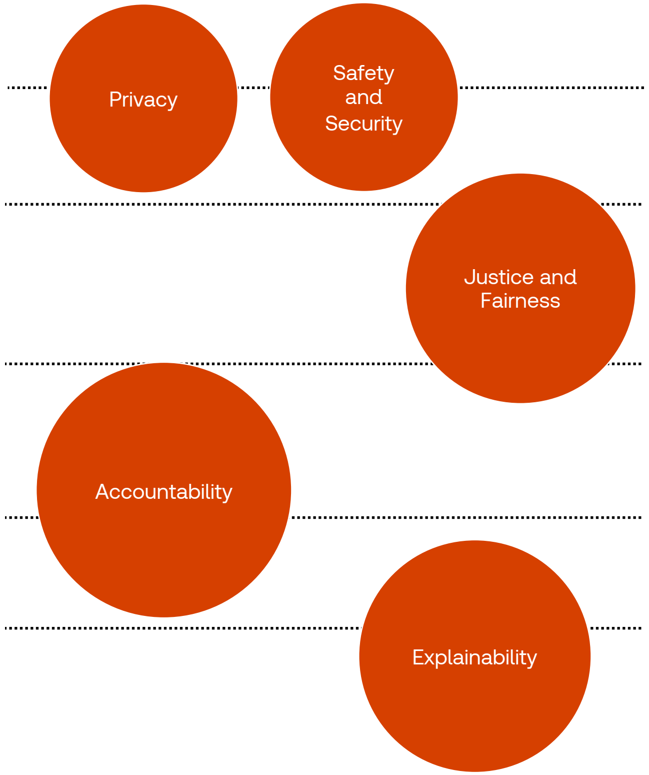
BUSINESS / PROBLEM-SOLVING SKILLS

Exploratory AI Problem-Solving

Business & Customer Engagement

Thomson Reuters Labs™

Our AI principles

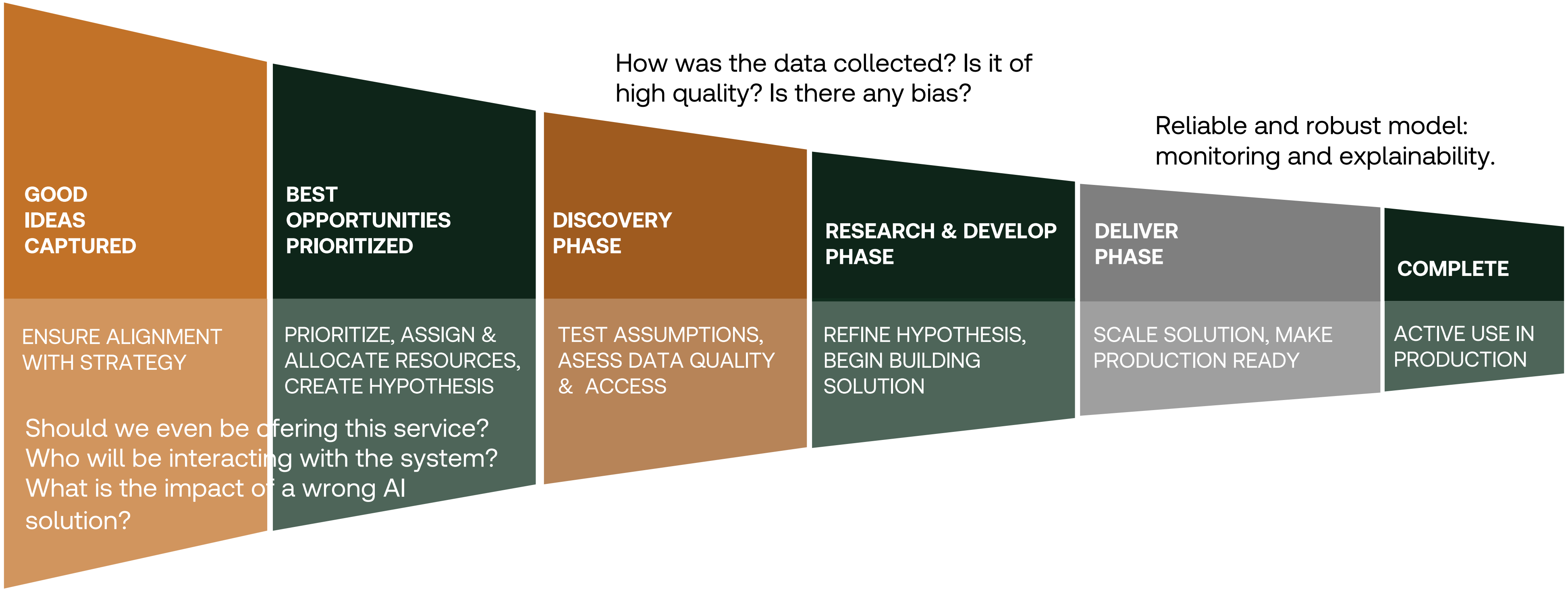


Thomson Reuters has adopted the following principles to promote trustworthiness in our continuous design, development, and deployment of artificial intelligence:

1. That Thomson Reuters will prioritize **safety, security, and privacy** throughout the design, development and deployment of our AI products and services.
2. That Thomson Reuters will strive to maintain a **human-centric approach**, and will strive to design, develop and deploy AI products and services **that treat people fairly**.
3. That Thomson Reuters aims to design, develop and deploy AI products and services that are **reliable** and that help empower people to make **efficient, informed, and socially beneficial decisions**.
4. That Thomson Reuters will maintain **appropriate accountability measures** for our AI products and services.
5. That Thomson Reuters will **implement practices** intended to **make the use of AI** in our products and services **interpretable**.

Prototyping at TR Labs™

Proof-of-concept average duration of six months



GOOD IDEAS CAPTURED

ENSURE ALIGNMENT WITH STRATEGY

BEST OPPORTUNITIES PRIORITIZED

PRIORITIZE, ASSIGN & ALLOCATE RESOURCES, CREATE HYPOTHESIS

DISCOVERY PHASE

TEST ASSUMPTIONS, ASESS DATA QUALITY & ACCESS

RESEARCH & DEVELOP PHASE

REFINE HYPOTHESIS, BEGIN BUILDING SOLUTION

DELIVER PHASE

SCALE SOLUTION, MAKE PRODUCTION READY

COMPLETE

ACTIVE USE IN PRODUCTION

Should we even be ofering this service?
Who will be interacting with the system?
What is the impact of a wrong AI solution?

How was the data collected? Is it of high quality? Is there any bias?

Reliable and robust model: monitoring and explainability.

Start with the end user

A design thinking approach

Who is going to use this and how can we make them more successful?

Define the problem

What is the hypothesis we want to test?

Do we have what we need?

What does success look like?

Build and deliver

What is the minimum viable solution?

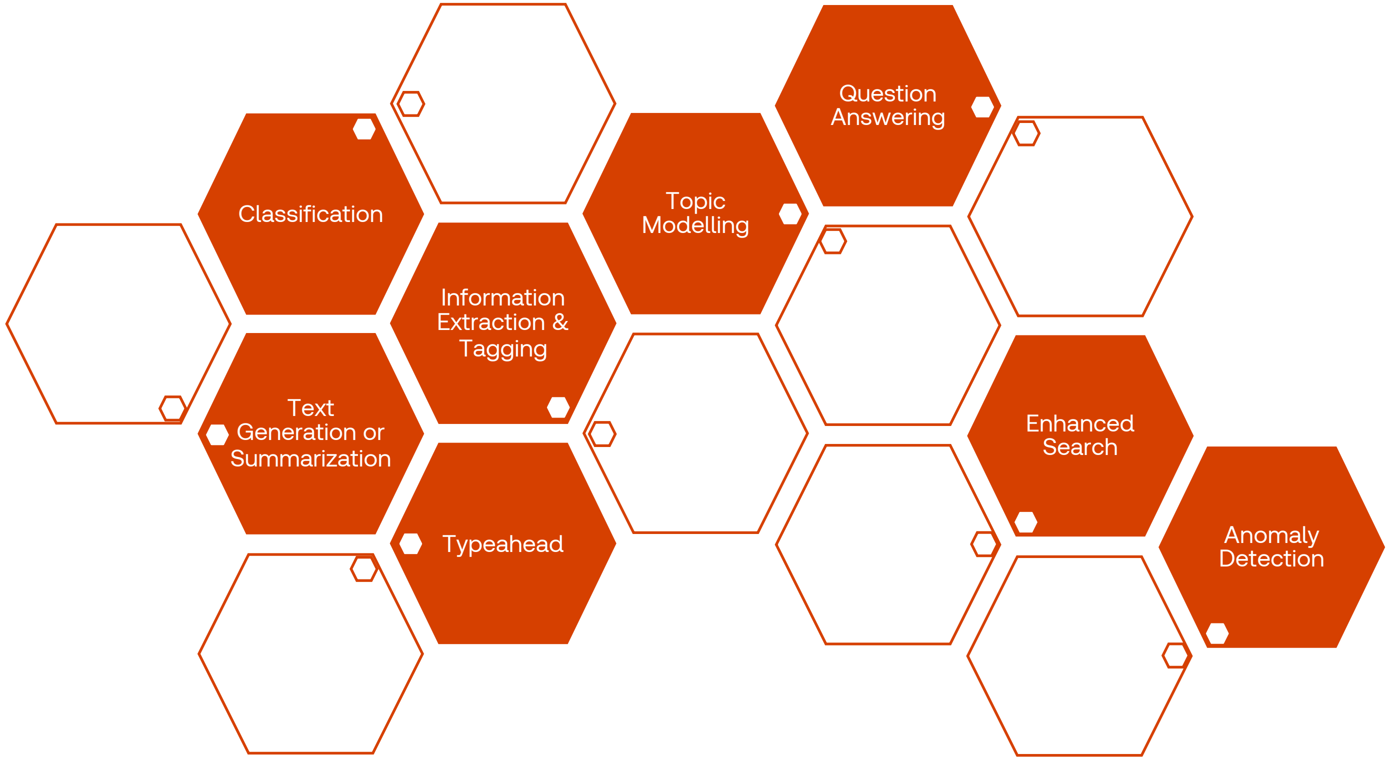
Constant iteration and co-creation involving end-users

Prototyping at TR Labs™

Types of problems

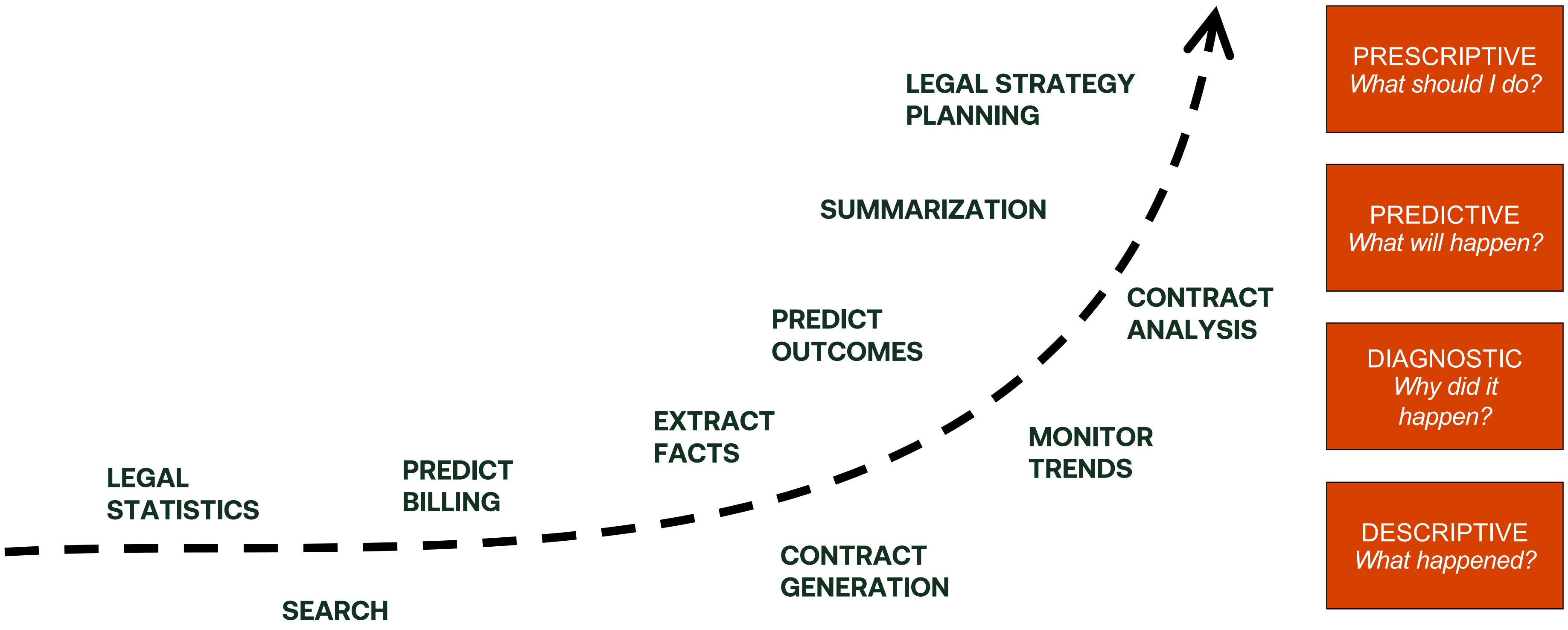
- Automating legal documents, contracts, and tax reports **generation**.
- **Summarizing** complex legal and financial documents.
- Enhancing **information retrieval**.
- **Automating** tax research and tax calculation.
- **Customizing** legal and tax information for each client.

and many more



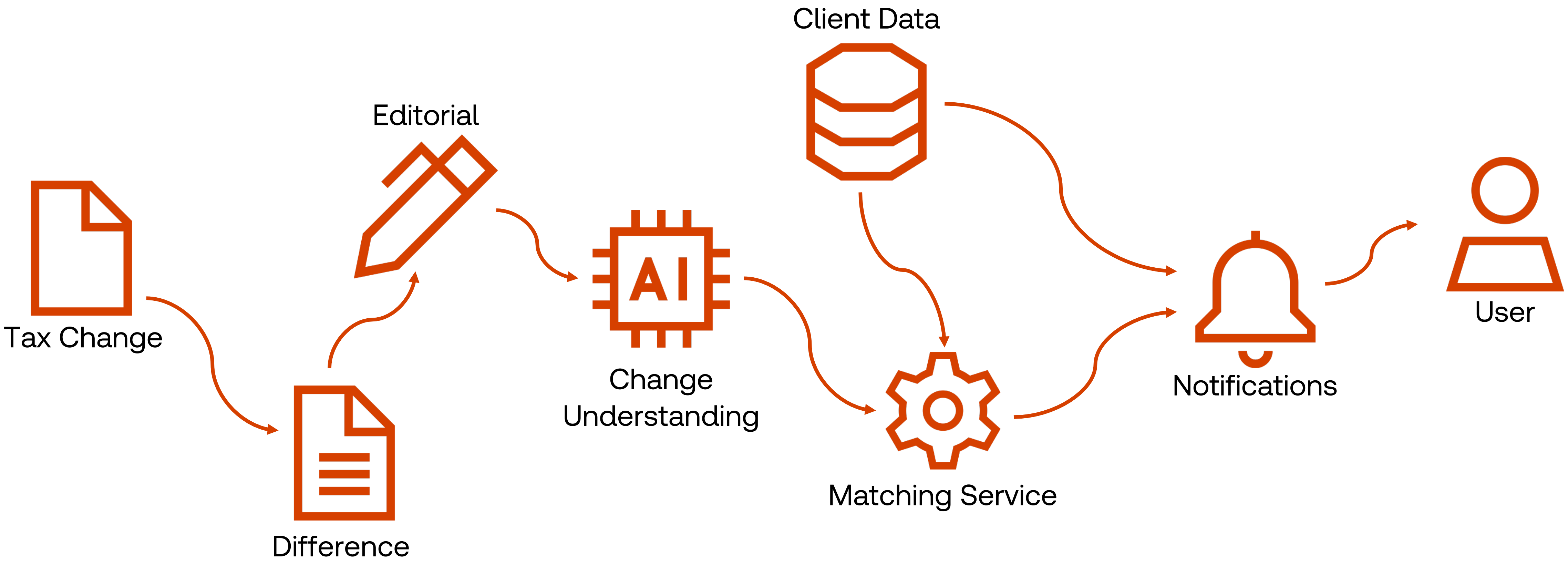
Prototyping at TR Labs™

Impact and feasibility of the challenges



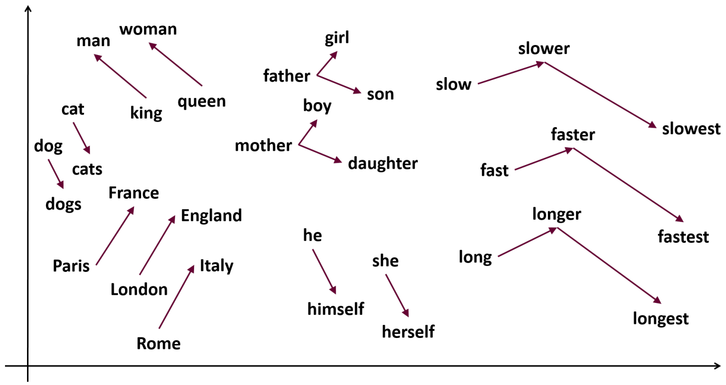
Example: Tax Regulatory Insights

Notify customers on impacting tax changes



Leveraging LLMs without compromising AI principles

Basics: Text Representation



Transform text to vectors to enable machine learning

TF-IDF

Term Frequency-Inverse Document Frequency

- Document represented by **numeric value for each word** (including all words in the corpus).
- **TF:** word occurrence in the document.
- **IDF:** word occurrence in all documents.
- How representative is a word for a document?

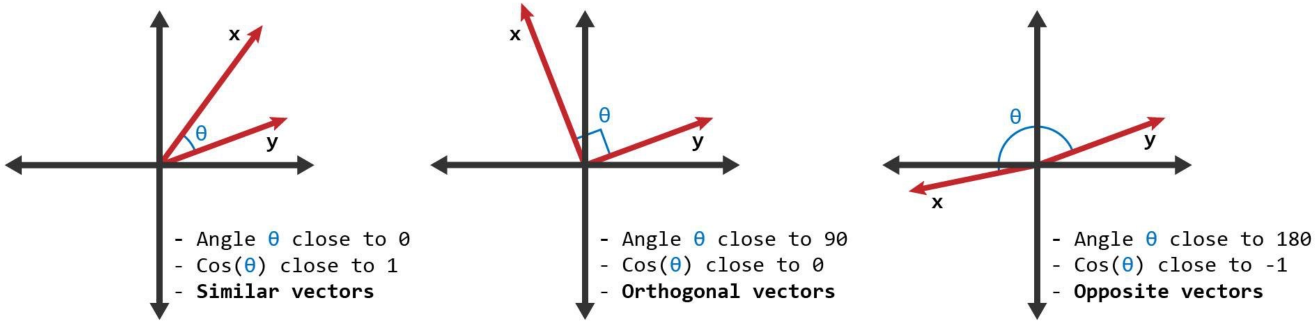
Word2Vec

(Trainable) dictionary of vector for each word

- Document represented by **vector for each word**.
- Each vector is n-dimensional.
- Semantically similar words are close in the vector space.
- Useful if semantically similar words indicate the same thing, e.g. the same class.
- There are also **multi-lingual** word vectors.

Basics: Natural Language Processing Tasks

Examples of classes of problems



- Angle θ close to 0
- Cos(θ) close to 1
- **Similar vectors**

- Angle θ close to 90
- Cos(θ) close to 0
- **Orthogonal vectors**

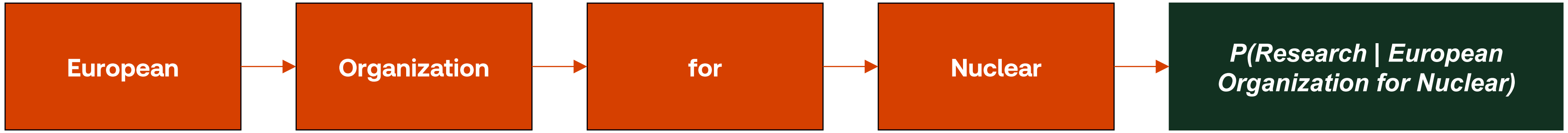
- Angle θ close to 180
- Cos(θ) close to -1
- **Opposite vectors**

- Auto Tagging & Classification
  - Multi-class, multi-label classification.
- Search & Text Similarity
  - BM25 or cosine similarity between document representations.
- Summarization
  - Extractive: find most representative sentence(s) of the text.
  - Abstractive: compose new sentence(s) summarizing the text.
- Clustering
  - Find groups of similar texts (represent text + use clustering algorithm).
- Translation
  - Can’t just translate word by word, generative deep neural models.

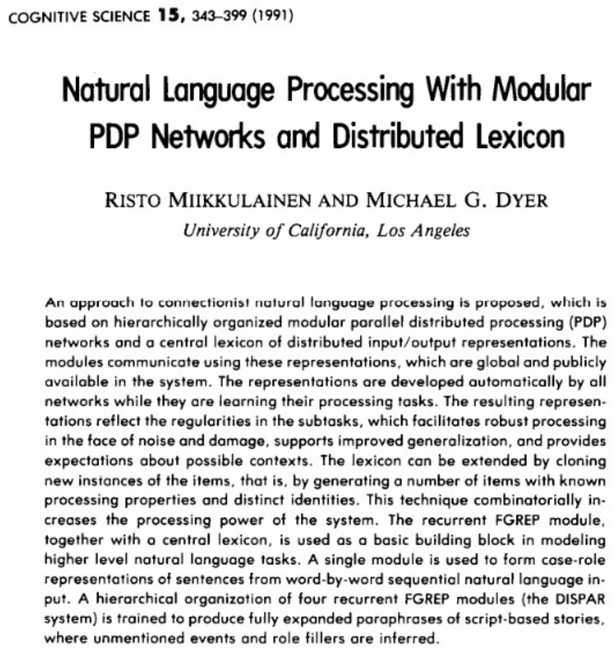
(Large) Language Models

From simple idea to a universal tool

- Take the sequence, vectorize it and generate token probabilities.
  - Basically, next token prediction.



- The bigger the models, the more emerging capabilities.
- With the proper user interface, LLMs become valuable tools and can solve many downstream tasks, e.g. text generation, classification, summarization, paraphrasing, entity extraction and translation.



COGNITIVE SCIENCE **15**, 343–399 (1991)

Natural Language Processing With Modular PDP Networks and Distributed Lexicon

RISTO MIIKKULAINEN AND MICHAEL G. DYER

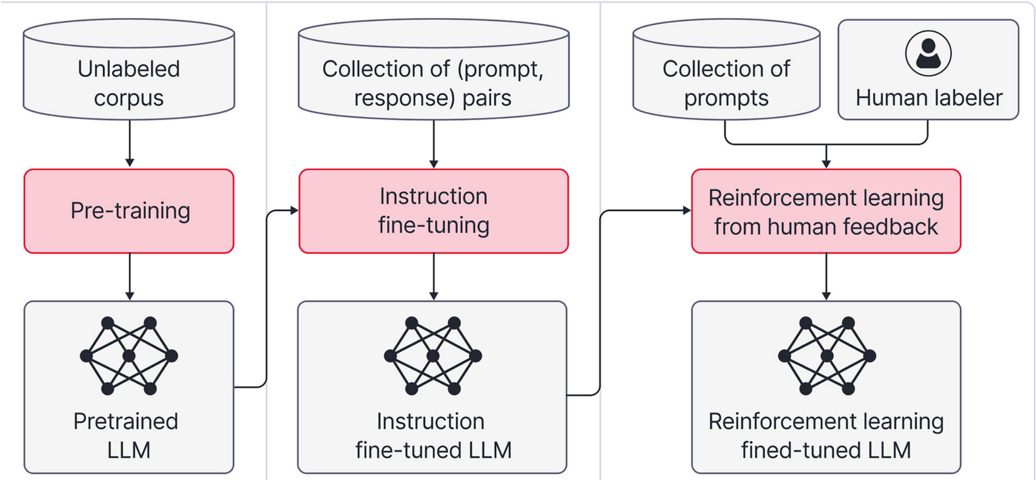
University of California, Los Angeles

An approach to connectionist natural language processing is proposed, which is based on hierarchically organized modular parallel distributed processing (PDP) networks and a central lexicon of distributed input/output representations. The modules communicate using these representations, which are global and publicly available in the system. The representations are developed automatically by all networks while they are learning their processing tasks. The resulting representations reflect the regularities in the subtasks, which facilitates robust processing in the face of noise and damage, supports improved generalization, and provides expectations about possible contexts. The lexicon can be extended by cloning new instances of the items, that is, by generating a number of items with known processing properties and distinct identities. This technique combinatorially increases the processing power of the system. The recurrent FGREP module, together with a central lexicon, is used as a basic building block in modeling higher level natural language tasks. A single module is used to form case-role representations of sentences from word-by-word sequential natural language input. A hierarchical organization of four recurrent FGREP modules (the DISPAR system) is trained to produce fully expanded paraphrases of script-based stories, where unmentioned events and role fillers are inferred.

Large Language Model Training

Why not train from scratch

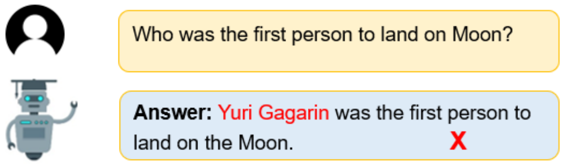
- **Computational resources:** high electricity costs.
- **Data costs:** the cost of acquiring, curating, and storing data.
- **Personnel costs:** researchers, engineers, and specialists.
- **Infrastructure costs:** physical infrastructure needed to support the training process.


Unlabeled corpus
Pre-training
Pretrained LLM
Collection of (prompt, response) pairs
Instruction fine-tuning
Instruction fine-tuned LLM
Collection of prompts
Human labeler
Reinforcement learning from human feedback
Reinforcement learning fined-tuned LLM

Hallucinations

Using the LLMs in production remains a challange

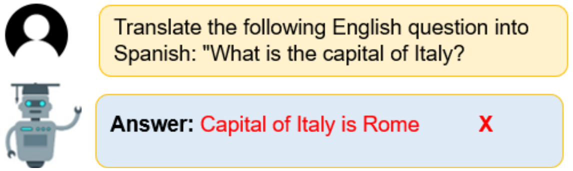
- Output does not align with facts or the user's input (*convincing but factually incorrect information).*
  - *Factuality hallucinations* (inconsistency or fabrication).



Who was the first person to land on Moon?

Answer: Yuri Gagarin was the first person to land on the Moon. X

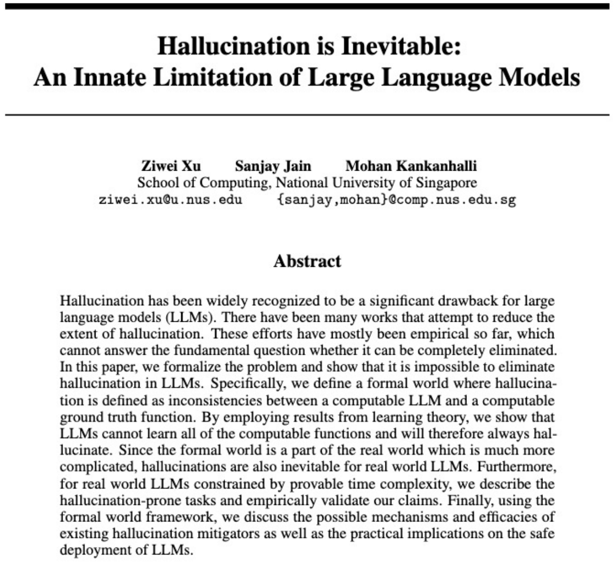
  - *Faithfulness hallucinations* (instruction, context or logical inconsistency).



Translate the following English question into Spanish: "What is the capital of Italy?

Answer: Capital of Italy is Rome X

- Other LLM issues: staleness, customization, attribution, revisions.



Hallucination is Inevitable:
An Innate Limitation of Large Language Models

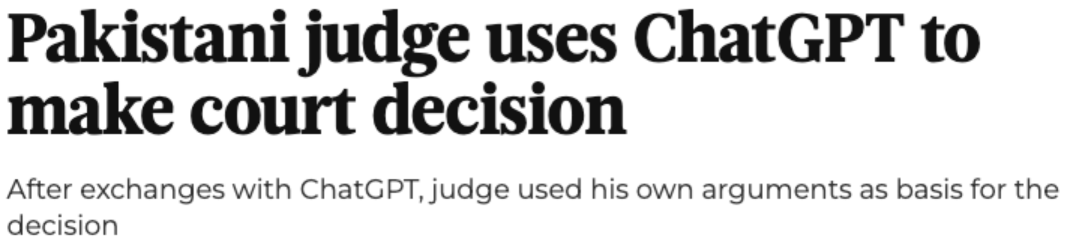
Ziwei Xu **Sanjay Jain** **Mohan Kankanhalli**
School of Computing, National University of Singapore
ziwei.xu@u.nus.edu {sanjay,mohan}@comp.nus.edu.sg

Abstract

Hallucination has been widely recognized to be a significant drawback for large language models (LLMs). There have been many works that attempt to reduce the extent of hallucination. These efforts have mostly been empirical so far, which cannot answer the fundamental question whether it can be completely eliminated. In this paper, we formalize the problem and show that it is impossible to eliminate hallucination in LLMs. Specifically, we define a formal world where hallucination is defined as inconsistencies between a computable LLM and a computable ground truth function. By employing results from learning theory, we show that LLMs cannot learn all of the computable functions and will therefore always hallucinate. Since the formal world is a part of the real world which is much more complicated, hallucinations are also inevitable for real world LLMs. Furthermore, for real world LLMs constrained by provable time complexity, we describe the hallucination-prone tasks and empirically validate our claims. Finally, using the formal world framework, we discuss the possible mechanisms and efficacies of existing hallucination mitigators as well as the practical implications on the safe deployment of LLMs.

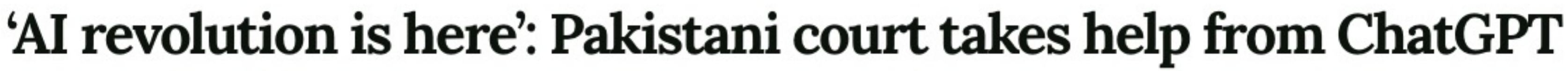
Hallucinations: Risks

High stakes in legal and tax domain

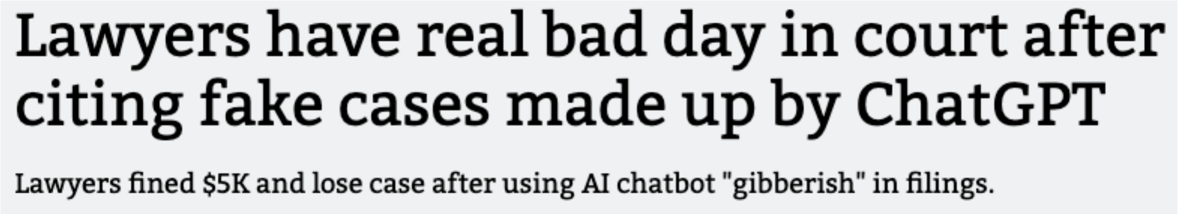


Pakistani judge uses ChatGPT to make court decision

After exchanges with ChatGPT, judge used his own arguments as basis for the decision

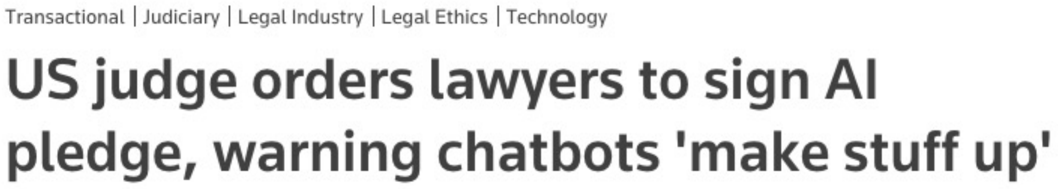


'AI revolution is here': Pakistani court takes help from ChatGPT



Lawyers have real bad day in court after citing fake cases made up by ChatGPT

Lawyers fined $5K and lose case after using AI chatbot "gibberish" in filings.



Transactional | Judiciary | Legal Industry | Legal Ethics | Technology

US judge orders lawyers to sign AI pledge, warning chatbots 'make stuff up'

Hallucinations: Causes

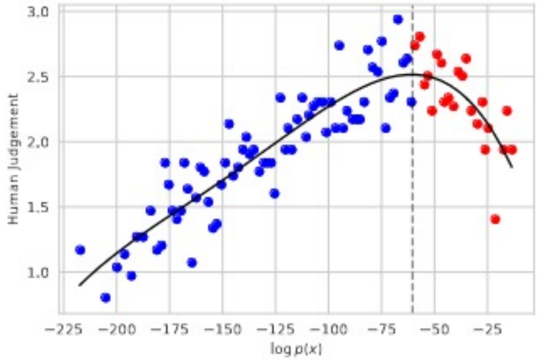
Three roots of hallucinations

- **Data**: flawed or biased training data can lead LLMs to learn incorrect patterns and factual errors.
  - Misinformation and biases: falsehoods, duplication, social bias.
  - Boundaries: domain knowledge deficiency, outdated factual knowledge.
  - Knowledge shortcut.
  - Knowledge recall.
- **Training**: issues during pre-training or alignment processes.
  - Pre-training: architecture flaws, exposure bias.
  - Alignment: capability, belief misalignment.
- **Inference**: stochastic nature of the decoding strategies can introduce randomness.
  - Sampling randomness.
  - Imperfect decoding representation: context size, softmax bottleneck.

Hallucinations: Inference Randomness

Not falling for the likelihood trap

- **The likelihood trap**: counter-intuitive observation that high likelihood sequences are often low quality.



https://arxiv.org/abs/2004.10450

- Deterministic decoding strategies are unsuitable for good interface.
- Use stochastic decoding strategy, e.g., *top-k*, *top-p*.

Hallucinations: Detection

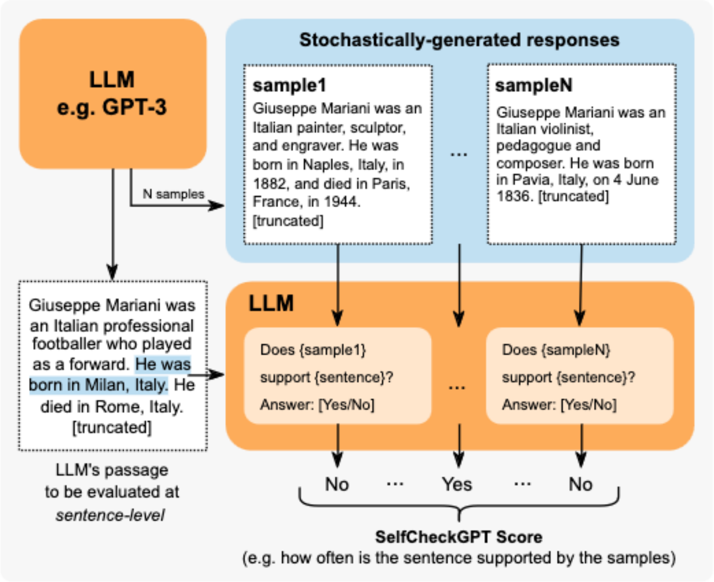
Measuring hallucinations

- **Confidence score**: output reflects LLM’s confidence in the generated text's correctness.
    - Uncertainty estimation: use internal model states (e.g. token probability) to identify hallucination.
- **Consistency check**: compare the output with knowledge sources.
    - Measure overlap using, e.g. n-grams.
    - Prompting based.
- **Benchmarks**: standardized datasets designed to test for hallucinations.
- Error detection and analysis protocols or model and data drift monitoring systems should be included standard deployment stategies.

Hallucinations: Detection Example

SelfCheckGPT

- Compare multiple responses from a model using the same query.
- Measure consistency between those responses, using, e.g., simple prompt.
- Drawback: more calls, more tokens lead to higher costs.



LLM
e.g. GPT-3

N samples

Stochastically-generated responses

sample1
Giuseppe Mariani was an Italian painter, sculptor, and engraver. He was born in Naples, Italy, in 1882, and died in Paris, France, in 1944. [truncated]

...

sampleN
Giuseppe Mariani was an Italian violinist, pedagogue and composer. He was born in Pavia, Italy, on 4 June 1836. [truncated]

Giuseppe Mariani was an Italian professional footballer who played as a forward. He was born in Milan, Italy. He died in Rome, Italy. [truncated]

LLM's passage
to be evaluated at
sentence-level

LLM

Does {sample1}
support {sentence}?
Answer: [Yes/No]

...

Does {sampleN}
support {sentence}?
Answer: [Yes/No]

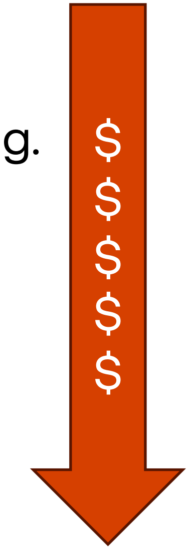
No ... Yes ... No

SelfCheckGPT Score
(e.g. how often is the sentence supported by the samples)

Hallucinations: Mitigation

Mitigation strategies have different cost

- **Model solutions**: parameter setting (e.g. temperature), model choices.
- **Prompt engineering**: guiding model behaviour with defined rules and limitations via prompting.
- **External memory**: e.g. **Retrieval-Augmented Generation (RAG)**.
- **Ensemble methods**: combining multiple models, e.g. LLM Blender.
- **Model fine-tuning**: refining the weights for specific downstream tasks and datasets.
- **Oversight and editorial control**: fact-checking and verification by human experts.



Retrieval Augmented Generation (RAGs)

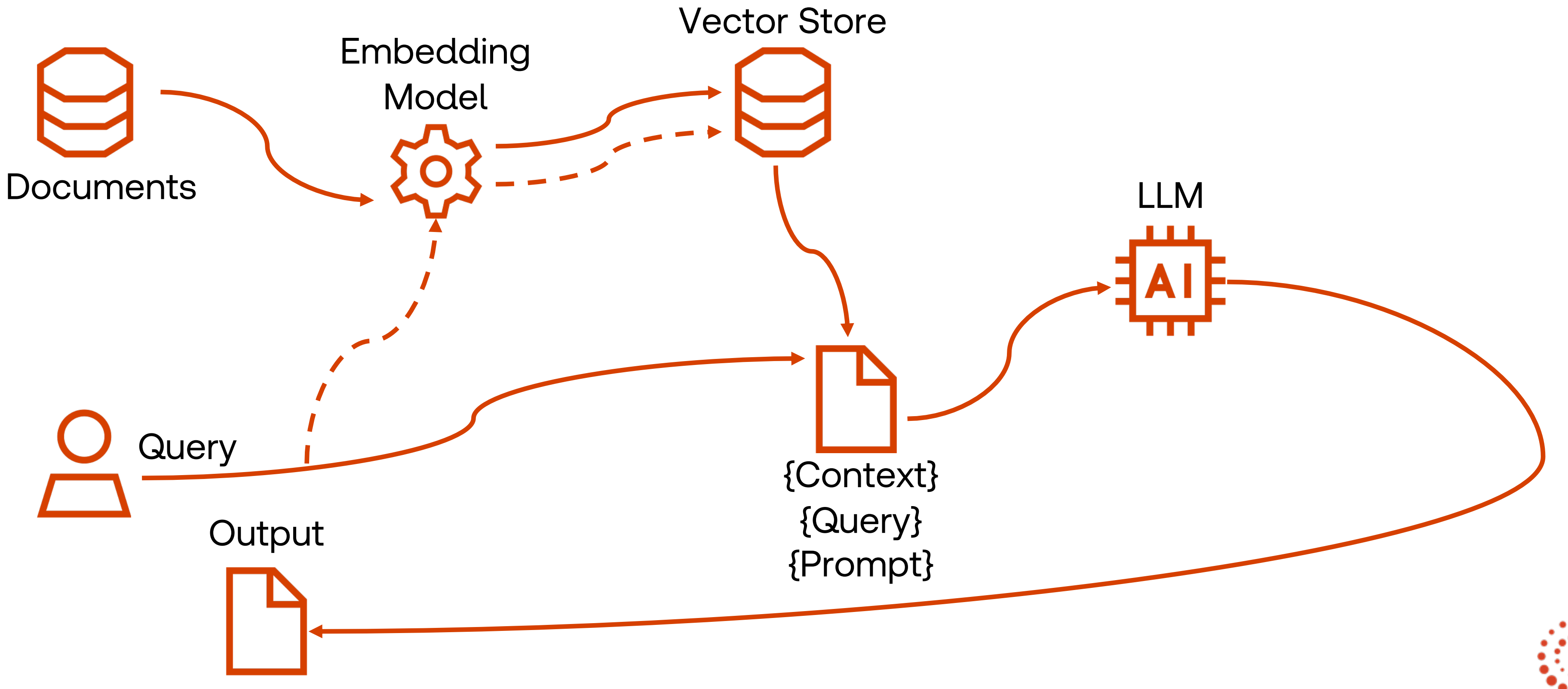
External memory: enhance performace

- **Integration of external knowledge within context**: retrieve relevant information during inference, enabling the generation of factually grounded outputs.
- **Improved factual accuracy**: mitigate the issue of hallucinations.
- Efficient **customization** and **revisions**.
- **Explainability**: opens avenues to mitigate the issue of attribution.
- **Adaptability to evolving knowledge**: mitigates the issue of staleness.

Retrieval Augmented Generation (RAGs)

Non-parametric memory

- During inference retrieve relevant information and agument the prompt.



Fine-Tuning

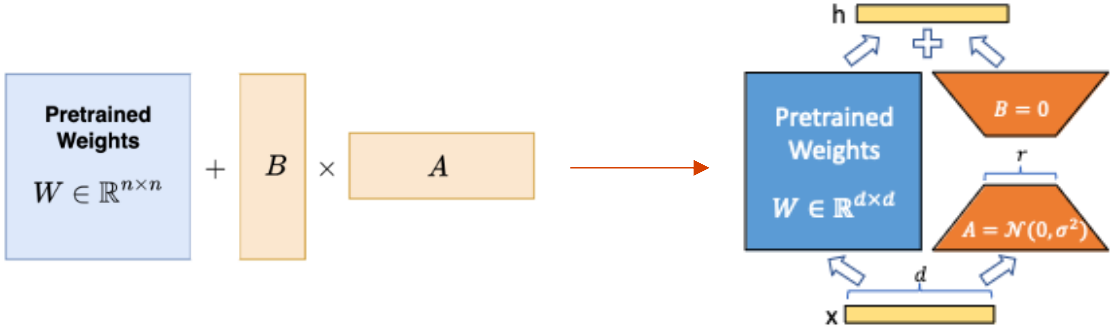
Tradeoffs of fine-tuning

- **Customization:** moving away from general-purpose to domain-specific, using a focused dataset.
- **Performance boost**: build upon the already acquired knowledge and retain old skills.
- **Resource efficiency**: requires less data and computational resources than training from scratch.
- **Data optimization**: fine-tuning relies on the quality of the data used.
- But it is **still costly** and introduces a risk of **forgetting**.

Fine-Tuning: LoRA

Lower fine-tuning duration and costs

- Low-Rank Adaptation (LoRA), a method for adapting LLMs for specific tasks.
- Use rank-decomposition matrices to achieve adaptation and reduce the number of trainable parameters.
- Parameters fused with original weights - avoids introducing additional inference latency.


Pretrained Weights
$W \in \mathbb{R}^{n \times n}$
$+$
B
$\times$
A
h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

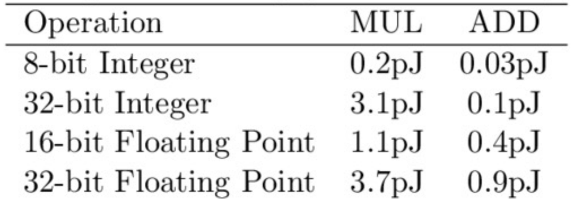
- Stanford Alpaca: a protocol to fine-tune open-source LLM with self-instruct and LoRA for a **few hundred dollars.**

Stanford Alpaca

Quantization

Reducing cost with fewer bits

- Model training and inference uses floating point representation.
- **Quantization** results in fixed precision weights and activations.
  - Reduction in computational cost for matrix multiplication and addition.
  - Reduction in data transfer and the memory needed for storing tensors.
- Quantization reduces energy, latency (training and inference) and area.
- Quantization aware training, quantizes and fine-tunes to recover lost performance.



Operation	MUL	ADD
8-bit Integer	0.2pJ	0.03pJ
32-bit Integer	3.1pJ	0.1pJ
16-bit Floating Point	1.1pJ	0.4pJ
32-bit Floating Point	3.7pJ	0.9pJ

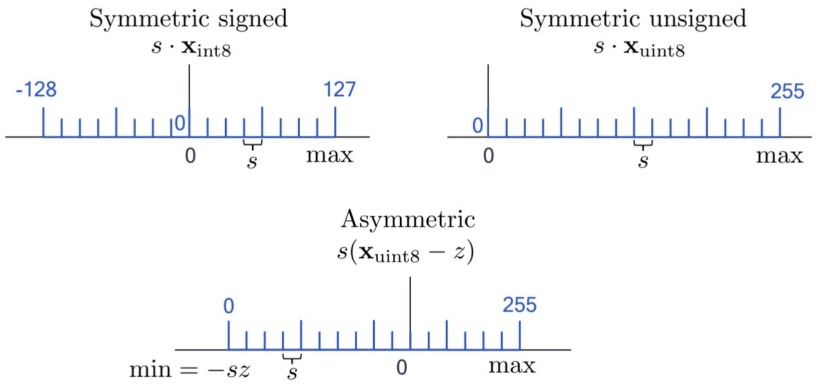
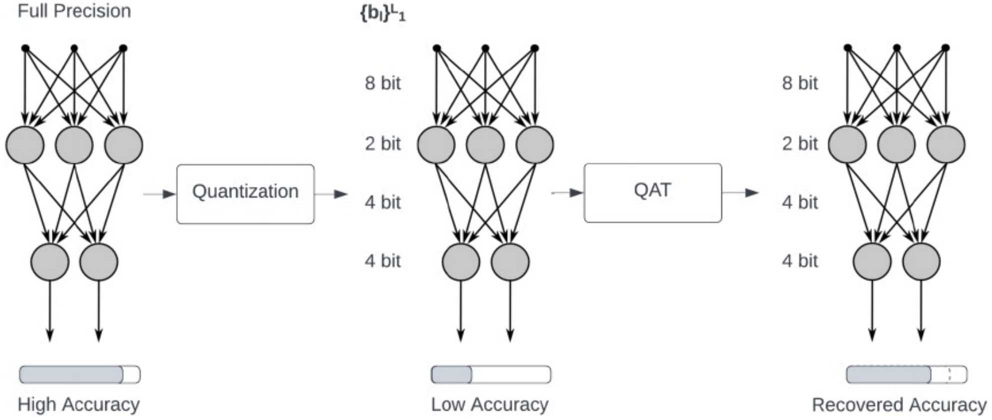

Symmetric signed
$s \cdot \mathbf{x}_{\text{int8}}$
-128
127
0
0
s
max
Symmetric unsigned
$s \cdot \mathbf{x}_{\text{uint8}}$
0
255
0
s
max
Asymmetric
$s(\mathbf{x}_{\text{uint8}} - z)$
0
255
$\min = -sz$
s
0
max

Figure 3: A visual explanation of the different uniform quantization grids for a bit-width of 8. s is the scaling factor, z the zero-point. The floating-point grid is in black, the integer quantized grid in blue.

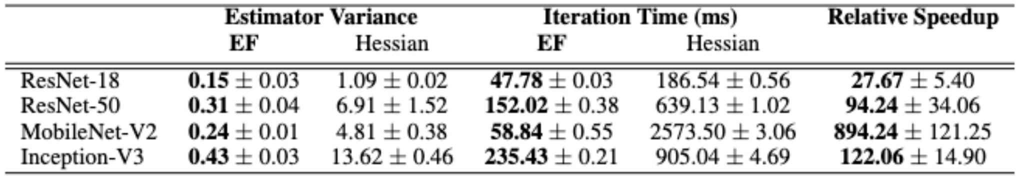
Heterogeneous Quantization

KT-CEVA collaboration: optimal quantization configuration

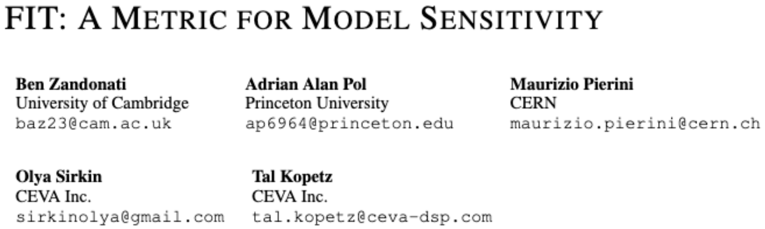
- Quantization granularity: homogeneous or heterogeneous (per-layer, -channel, -group).



- Accuracy drop is minimised by considering the **sensitivity** of layers to quantization.



	Estimator Variance		Iteration Time (ms)		Relative Speedup
	EF	Hessian	EF	Hessian	
ResNet-18	**0.15** ± 0.03	1.09 ± 0.02	**47.78** ± 0.03	186.54 ± 0.56	**27.67** ± 5.40
ResNet-50	**0.31** ± 0.04	6.91 ± 1.52	**152.02** ± 0.38	639.13 ± 1.02	**94.24** ± 34.06
MobileNet-V2	**0.24** ± 0.01	4.81 ± 0.38	**58.84** ± 0.55	2573.50 ± 3.06	**894.24** ± 121.25
Inception-V3	**0.43** ± 0.03	13.62 ± 0.46	**235.43** ± 0.21	905.04 ± 4.69	**122.06** ± 14.90



FIT: A METRIC FOR MODEL SENSITIVITY

Ben Zandonati
University of Cambridge
baz23@cam.ac.uk

Adrian Alan Pol
Princeton University
ap6964@princeton.edu

Maurizio Pierini
CERN
maurizio.pierini@cern.ch

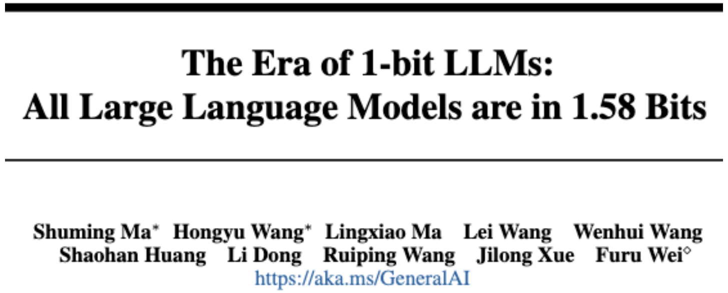
Olya Sirkin
CEVA Inc.
sirkinolya@gmail.com

Tal Kopetz
CEVA Inc.
tal.kopetz@ceva-dsp.com

Quantization of Language Models

KT-CEVA collaboration: optimal quantization configuration

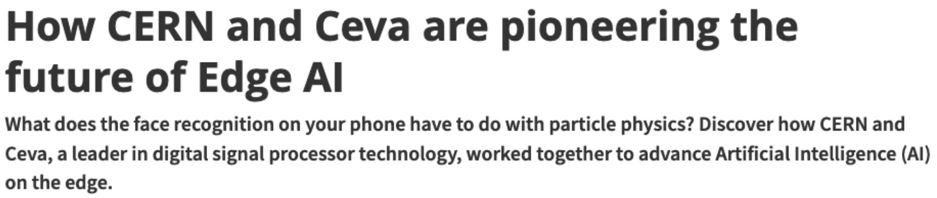
- Large Language Models can be quantized, too.



The Era of 1-bit LLMs:
All Large Language Models are in 1.58 Bits

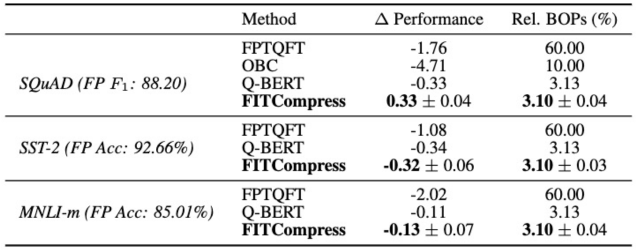
Shuming Ma* Hongyu Wang* Lingxiao Ma Lei Wang Wenhui Wang
Shaohan Huang Li Dong Ruiping Wang Jilong Xue Furu Wei◇
https://aka.ms/GeneralAI

- *FITCompress* (KT-CEVA collaboration), compressed BERT to 3.10% original size with minimal performance loss.



How CERN and Ceva are pioneering the future of Edge AI

What does the face recognition on your phone have to do with particle physics? Discover how CERN and Ceva, a leader in digital signal processor technology, worked together to advance Artificial Intelligence (AI) on the edge.



	Method	Δ Performance	Rel. BOPs (%)
SQuAD (FP F_1: 88.20)	FPTQFT	-1.76	60.00
	OBC	-4.71	10.00
	Q-BERT	-0.33	3.13
	FITCompress	**0.33** ± 0.04	**3.10** ± 0.04
SST-2 (FP Acc: 92.66%)	FPTQFT	-1.08	60.00
	Q-BERT	-0.34	3.13
	FITCompress	**-0.32** ± 0.06	**3.10** ± 0.03
MNLI-m (FP Acc: 85.01%)	FPTQFT	-2.02	60.00
	Q-BERT	-0.11	3.13
	FITCompress	**-0.13** ± 0.07	**3.10** ± 0.04

- *FITCompress* was integrated into CEVA’s products.

Fine-Tuning: QLoRA

Accelarate fine-tuning

- Fine-tuning 65B parameter models on a single GPU is difficult because of memory requirements.
- Quantized LoRA (QLoRA) reduces memory usage during fine tuning by using 4-bit quantization.
- QLoRA can be combined with other fine-tuning techniques, e.g. Alpaca.

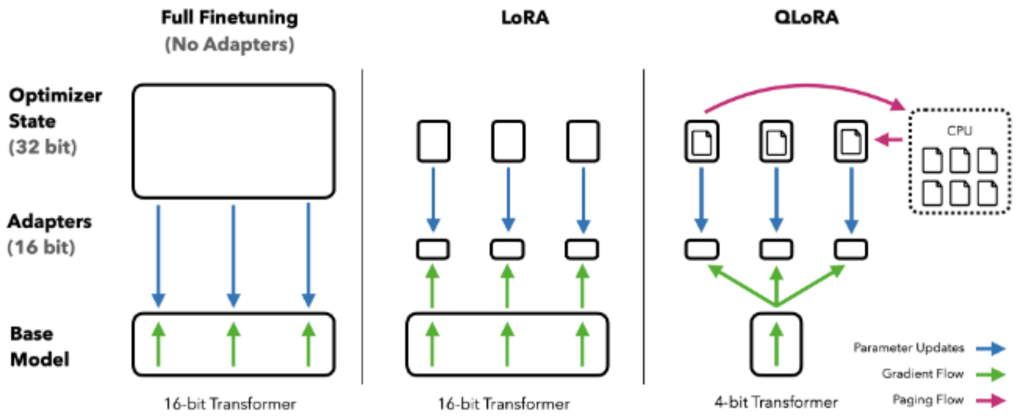


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

Summary

and Miscellaneous

- LLMs became a standard tool for solving NLP problems.
- Based on current understanding, some level of hallucinations is expected.
    - To mitigate this issue, common strategies are prompt engineering, RAGs and fine-tuning.
    - RAGs remain a popular and robust solution for grounding LLMs.
- Collaboration with model providers is crucial to optimize prompts and costs.
- During the design phase, we focus on minimizing calls and tokens.
- Leveraging cloud-based infrastructure for its scalability and potential cost benefits.
- Efficient fine-tuning for faster delivery.

LLMs at CERN?

Moving from CERN to Big Tech as a Postdoc

What skills do you have?

Hard skills:

- Strong foundation in math and statistics.
- Programming experience
- Handling large datasets.
- Data analysis and problem-solving.
- Experience with computing tools.
- Experimental design.

Soft skills:

- Analytical skills.
- Communication.
- Collaboration.
- Leadership.

You are missing NLP experience… let’s brainstorm.

Not on the LLM train yet? Don’t worry, I got this. ;)

Shifter Assistant

During data taking, part of the shifter's job is to note any relevant information for future operator crews, detector experts and data certification personnel.

Based on the past logs (and optionally some other inputs from the shifter), generate a set of high-quality comments on the apparatus's state via simple **typeahead**.

Additionally, extending this approach by including apparatus status from data monitoring tools via a language interface is a great way to increase the expert’s visibility of the detector issues.

Research Assistant

Copious data piling up in CERN Twiki is an issue when onboarding a new tool, especially for students who have not been around CERN long enough. The search functionality often fails to deliver, and legacy documentation may lead to wasted hours.

The interaction with Twiki can extended to a chat interface using the **RAG** approach, and the relevant content may be leveraged to provide question-specific answers.

When documentation is updated, so is the chatbot.

Educational Tools

Science communication and outreach have long been CERN’s goals.

CERN may offer a new interface for interacting with the public (**text-to-speech**, for example).

Maria Sklodowska-Curie’s avatar can clearly and engagingly answer questions about scientific phenomena in multiple languages. Grounding the system using CERN’s guide material can mitigate the hallucinations.

Thank you!

adrianalan.me

Adrian Alan Pol

Applied Machine Learning Scientist

Thomson Reuters Labs