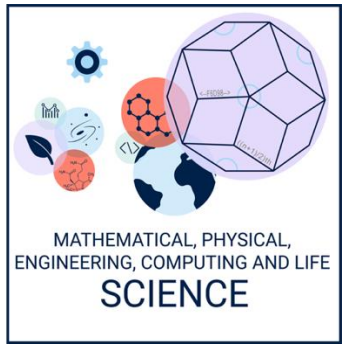
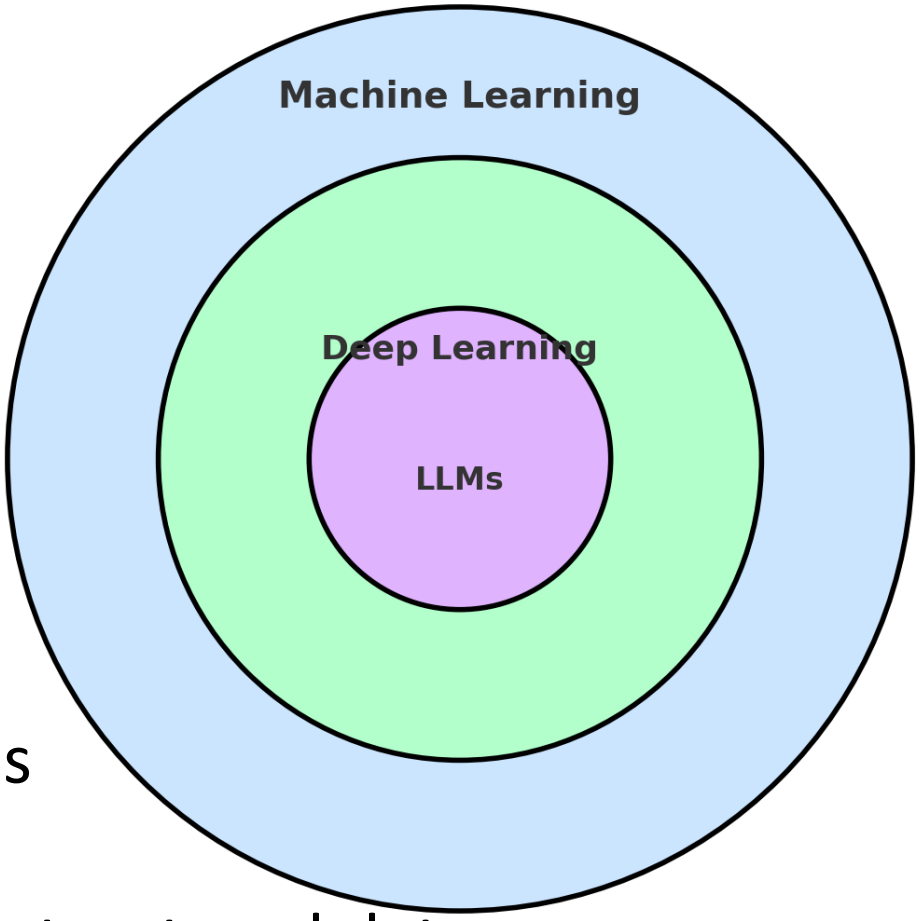


Where the AI future happens

Jim Naismith
Sept 2025

Hierarchy of AI



Broadly intelligent machines, able to make decisions

Pattern recognition by machine learning using high structured data

Deep learning using unstructured data (ingesting audio, text)

Large Language model, “understanding” and “reasoning”

AI @ Oxford

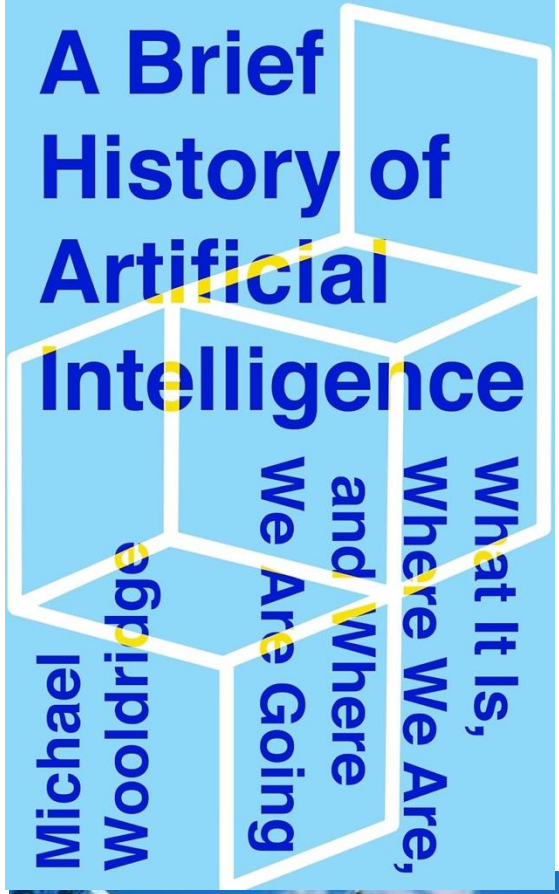
AI will change everything, there will be a pre-AI and post-AI Oxford

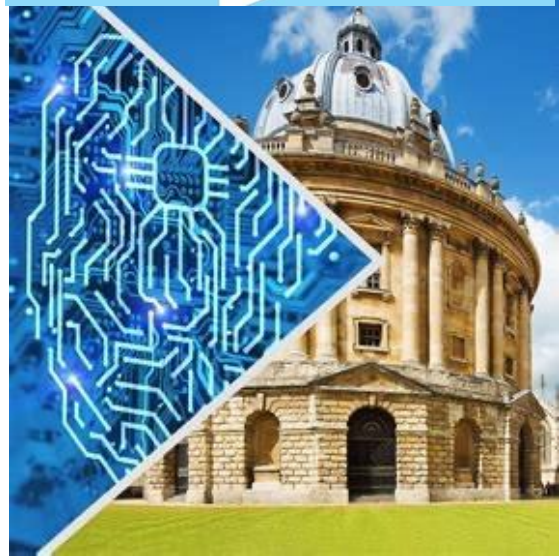
*AI is a tool, not a religion **nor** the end of human brain*

Everyone at Oxford must be able to use AI tools

Students, researchers & senior professionals must be understand their limitations, risks, & opportunities

AI first in our teaching, research and administration





AI @ Oxford

Professor Sir Nigel Shadbolt FRS

Pioneer in AI research

Professor of Computer Science & Ethics in AI

Coordinating AI Research across Oxford

AI @ Oxford

Professor Anne Trefethen FREng

Pioneer in scientific computing

Professor of Engineering

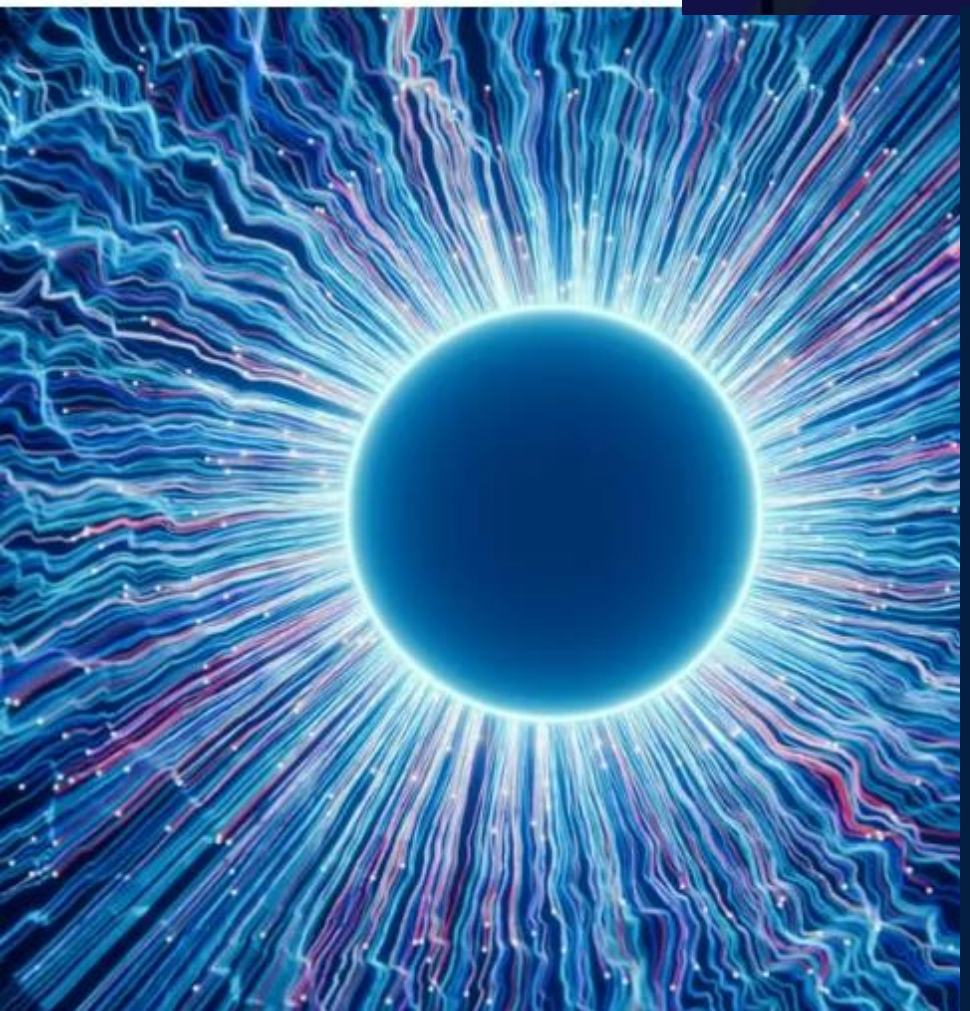
Coordinating AI in practice across Oxford

Major partnerships with OpenAI & Microsoft

UNIVERSITY OF OXFORD

MATHEMATICAL, PHYSICAL, ENGINEERING, COMPUTING AND LIFE SCIENCE

AI Competency Centre



We develop and offer a selection of resources and trainin for staff to ensure that everyone can use AI/ML tools with confidence, safety and for appropriate applications.

- Our training offering
- All upcoming trainings
- Expression of interest form

LLM (pre Deepseek)

Ingest vast amounts of text

Train by missing out words (tokens), predicting correct word, adjust weights* until it works reliably (huge compute) (back propagation)

“She kicked the football and scored a XXXX” Missing token – “goal”

This process is run over trillions of such puzzles for current LLMS

*More weights (parameters) more power

Fine tuning LLM

This raw LLM is then fine tuned

A series of prompts (questions) are used and the response tested

Weights are adjusted to get the desired response

This process is being automated, by AI agents

Fine tuning is not always useful, Retrieval-Augmented Generation can be better (essentially live web searching of curated knowledge)

Final tweaking LLM

Reinforcement Learning with Human Feedback, start asking questions and ranking responses

The model is then trained to produce the most highly ranked responses

This make the model for “human”, user friendly and hopefully safer*

No current model tells you step by step how to build a bomb, but they know how to.

Summary


Pretraining
Learns language basics
(guessing next word)
Fine-Tuning
Learns tasks & manners
(examples from tutors)
RLHF
Learns preferences
(humans rank answers)
Scaling
More data & compute
(broader knowledge)
Final Model
Assistant LLM
(can write, reason, explain

What drove LLMs

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

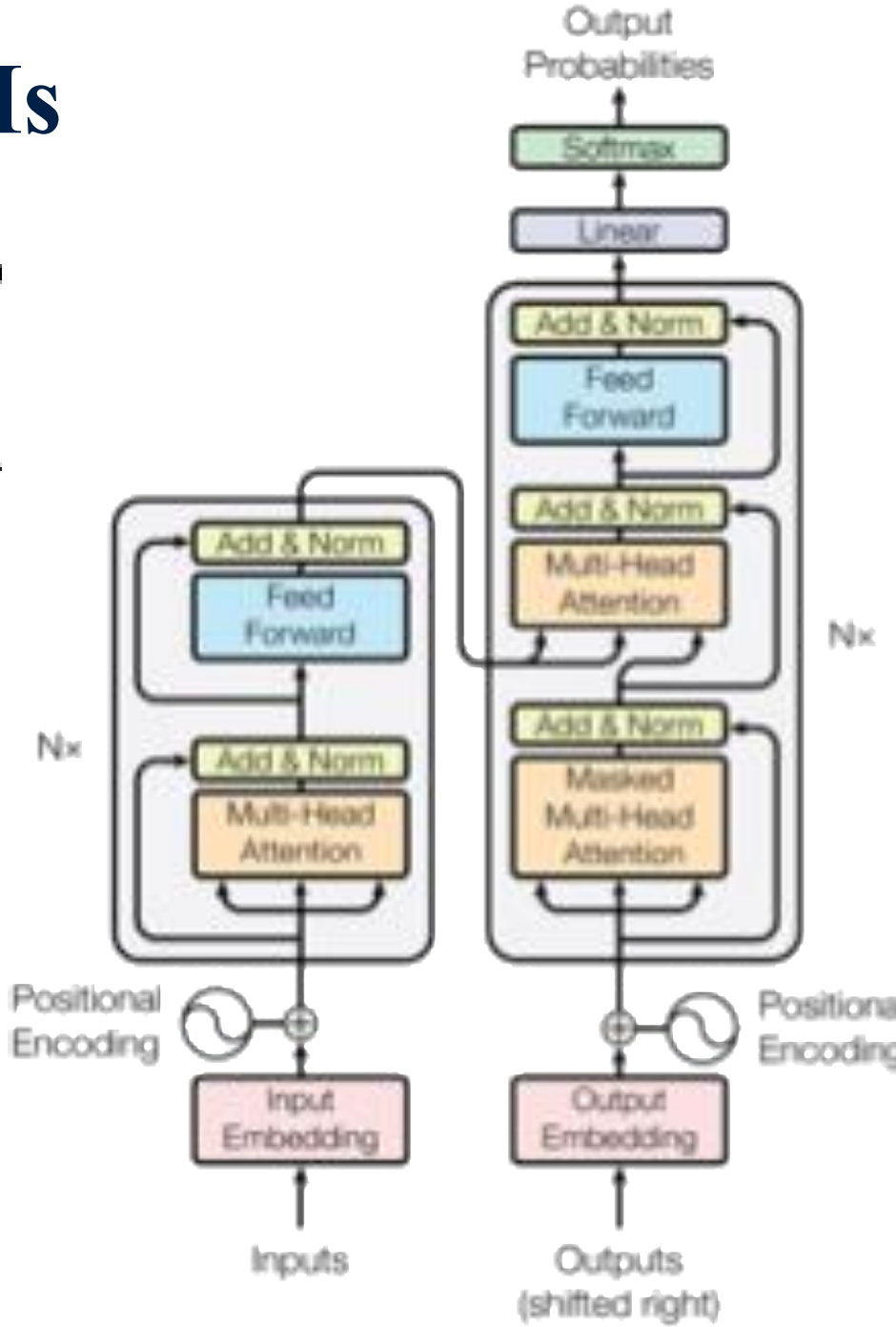
Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com



Tokens

“She kicked the footXX and scored a goal”

XX – ball, light, fall?

“She kicked the XXball and scored a goal”

XX – rugby, tennis, base or football ?

“She kickXX the football and scored a goal”
XX – ed, ing, ?

The transformer

Looks at tokens in parallel, not one at a time, new neural network architecture

Self attention means in a sentence, some words are more relevant to each other, “**Alice** and Bob work in management, **she** is the boss” She and Alice pay attention – key to context

Multihead attention key to parallelism, inspect sentences for different things at the same time, grammar, word choice / arrangement etc

Feed-forward layers adds complex relationship

Paying Attention

matrices Q, K, V:

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- QK^T :measures similarity between queries and keys
- Divide by $\sqrt{d_k}$:scaling for stability
- Softmax: converts similarities into probabilities (weights)
- Multiply by V: weighted sum of information.

"She kicked the football and scored a goal"

"She **kick**ed the **foot..ball** and *score*d a **goal"-** relevant to each other for meaning (semantics)

"She kick**ed** the football and scor**ed** a goal" - relevant to each other for grammar

"She kicked the football and scored **a goal"** – relevant to word choice

Feedforward $FFN(x)=W2\sigma(W1x+b1)+b2$

Kick – this token has relevance to other tokens – modified by attention

The token is a represented as a vector (number)

The vector is expanded to express the attention has identified

The expanded vector is refined to reduce noise and amplify new patterns

The modified expanded vector is reduced back to original size (but it is modified)

Repeat the whole process

SLMs What's difference

LLMs are generalist by design, do everything well

SLMs are specialist, they can do one thing brilliantly

Commercial adoption moving at pace

SLMs much cheaper to build and operate BUT their gaps could be problematic - combine

No simple metric to distinguish LLMs from SLMs

Efficient Language Model

Shockwaves in the AI community
but **very open innovation**

671 billion OPEN weights, order of magnitude smaller number of weights

Order of magnitude less compute power to train – easy to build on open weights

Deepseek technology

It is broken down into sub networks specialized for specific tasks.

Known as Mixture-of-Experts, polled for answers rather than whole model

Much faster & cheaper to operate

Breaks geometric linkage between compute and size of model (parameters)

Less compute for bigger models

Deepseek technology - more

Multi-head Latent Attention (MLA)

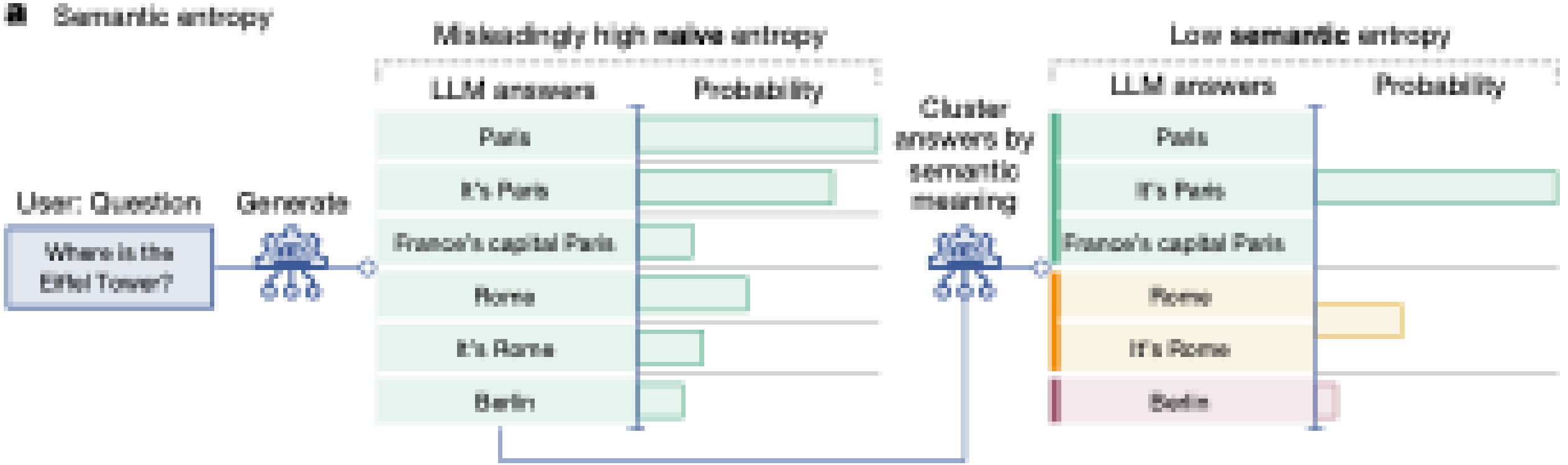
Normally every word matters equally, more words more compute

Latent attention imposes a new layer that summarizes the query

Can handle much longer query (context window), much more efficient compute but can lose important detail (compression)

Article

Detecting hallucinations in large language models using semantic entropy

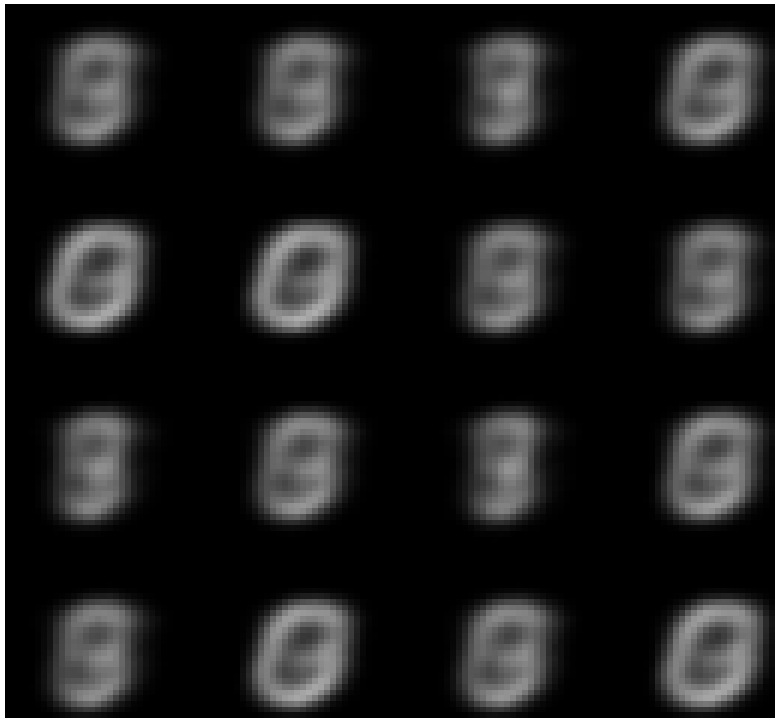


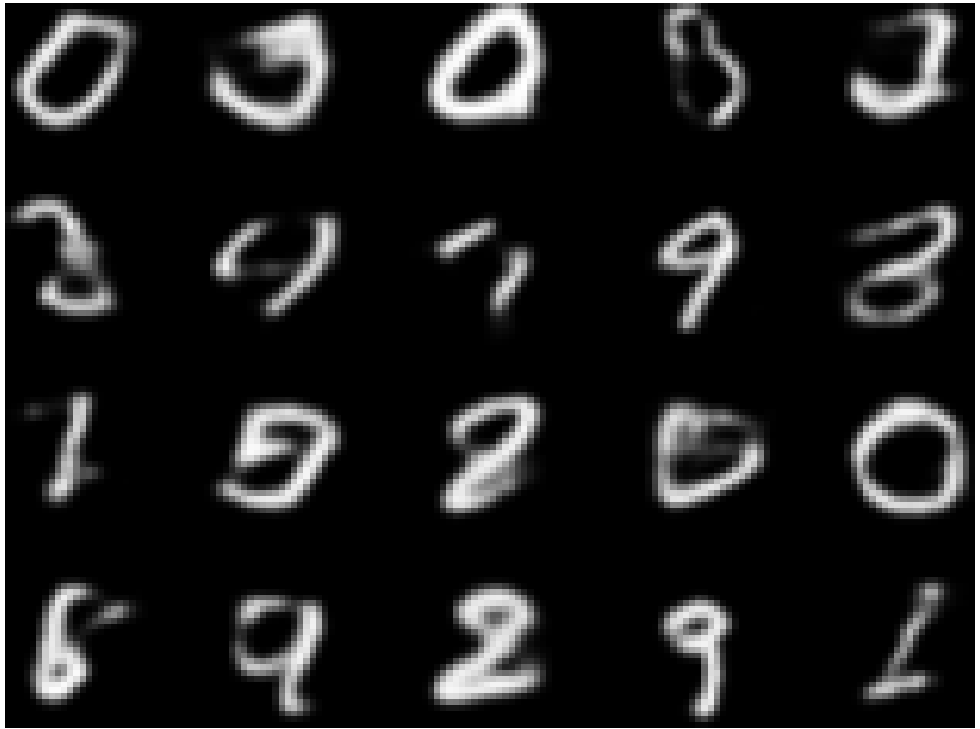
UNIVERSITY OF OXFORD

MATHEMATICAL, PHYSICAL, ENGINEERING, COMPUTING AND LIFE SCIENCE

Article

AI models collapse when trained on recursively generated data





Jailbreaking Large Language Models with Symbolic Mathematics

Emet Bethany
Secure AI and Autonomy Lab
University of Texas at San Antonio
emet.bethany@utsa.edu

Mazal Bethany
Secure AI and Autonomy Lab
University of Texas at San Antonio
mazal.bethany@utsa.edu

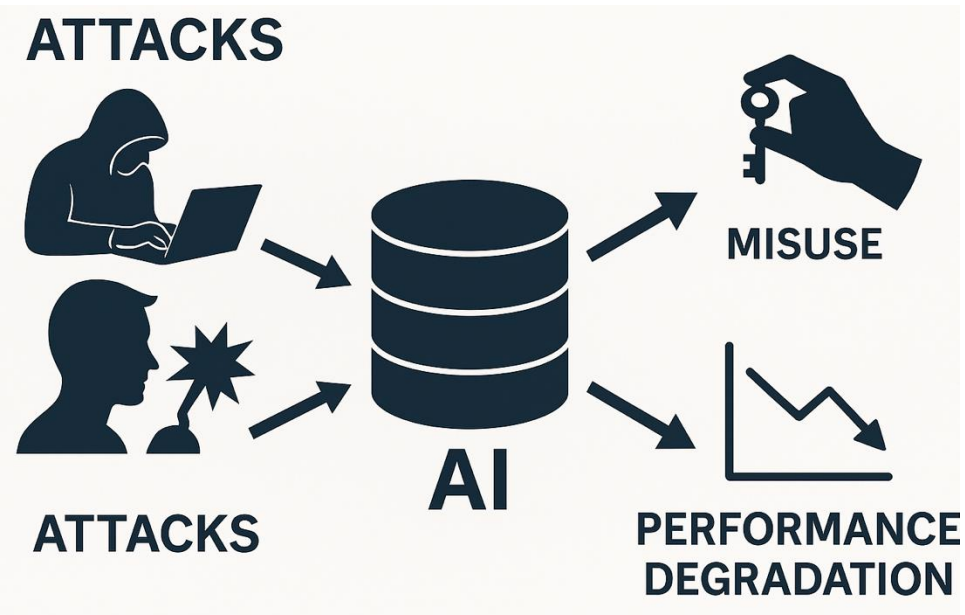
Juan Arturo Nolazco Flores
Data Science Hub & CoreLab Data Science
Tecnológico de Monterrey
jnolazco@tec.mx

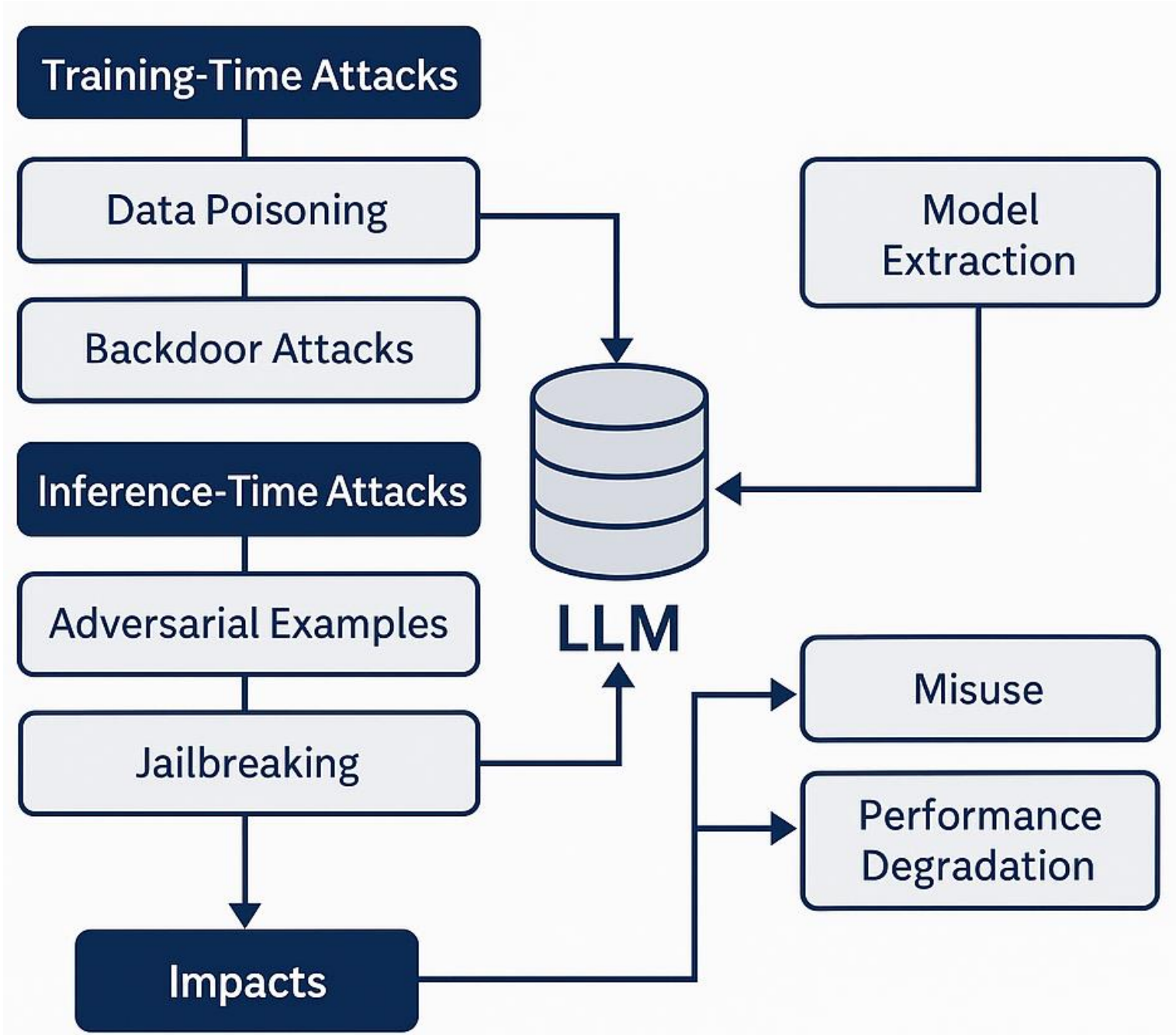
Sumit Kumar Jha
Computer Science Department
Florida International University
jha@cs.fiu.edu

Peyman Najafirad
Secure AI and Autonomy Lab
University of Texas at San Antonio
peyman.najafirad@utsa.edu

Using mathematical formulation to present a problem, “how to hide a human body”

Asked as a maths problem, a useful answer was produced.


ATTACKS
MISUSE
AI
ATTACKS
PERFORMANCE
DEGRADATION


Training-Time Attacks
Data Poisoning
Backdoor Attacks
Inference-Time Attacks
Adversarial Examples
Jailbreaking
Impacts
Model Extraction
LLM
Misuse
Performance Degradation

Why this really matters

Agentic AI is coming fast

It will have broad goals,

book me a holiday in a sunny place

invest my savings to achieve greatest return

protect the ship against attack

It can take action

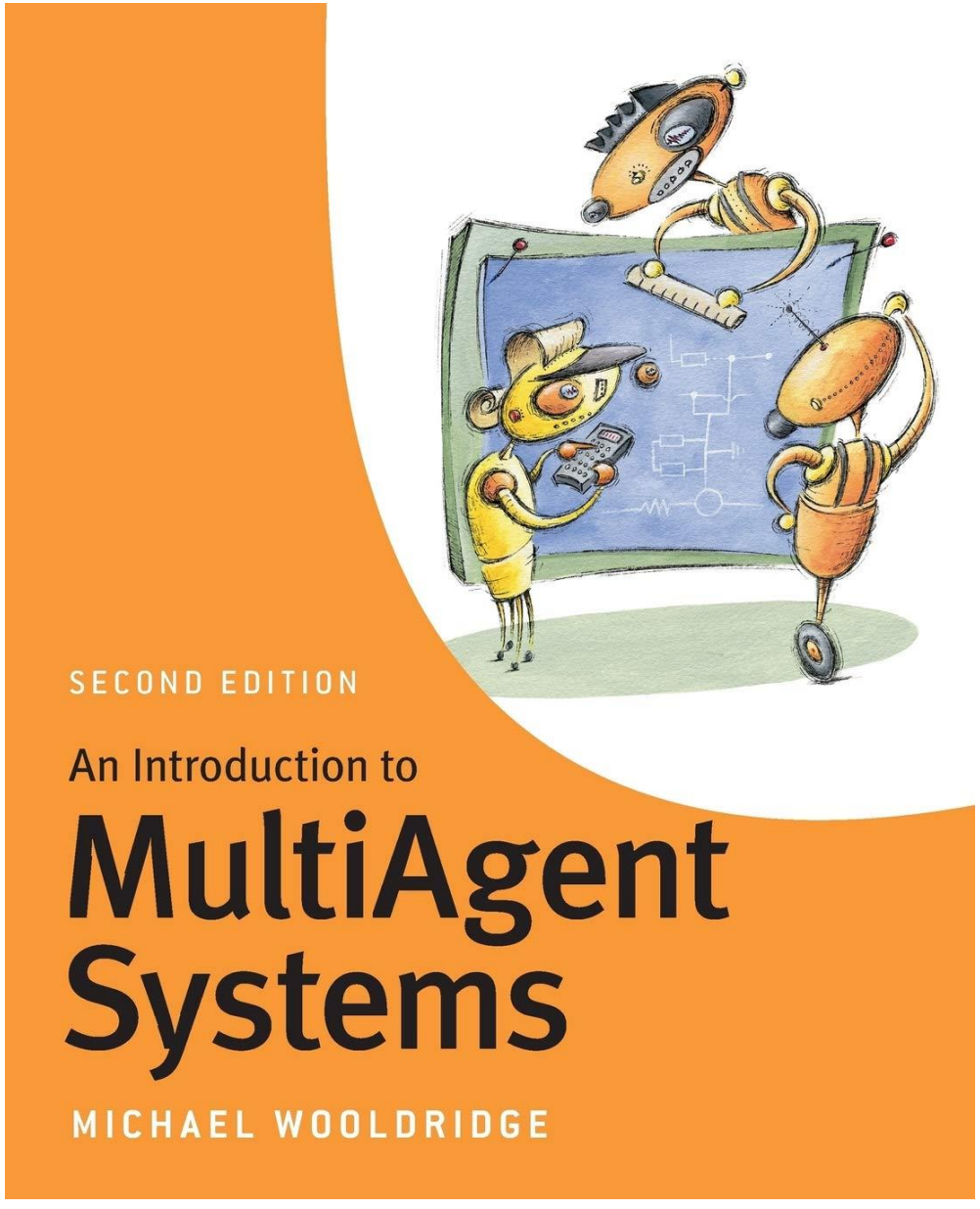
talk to other AI

move money

fire missile

Won't need prompts

Will learn from itself

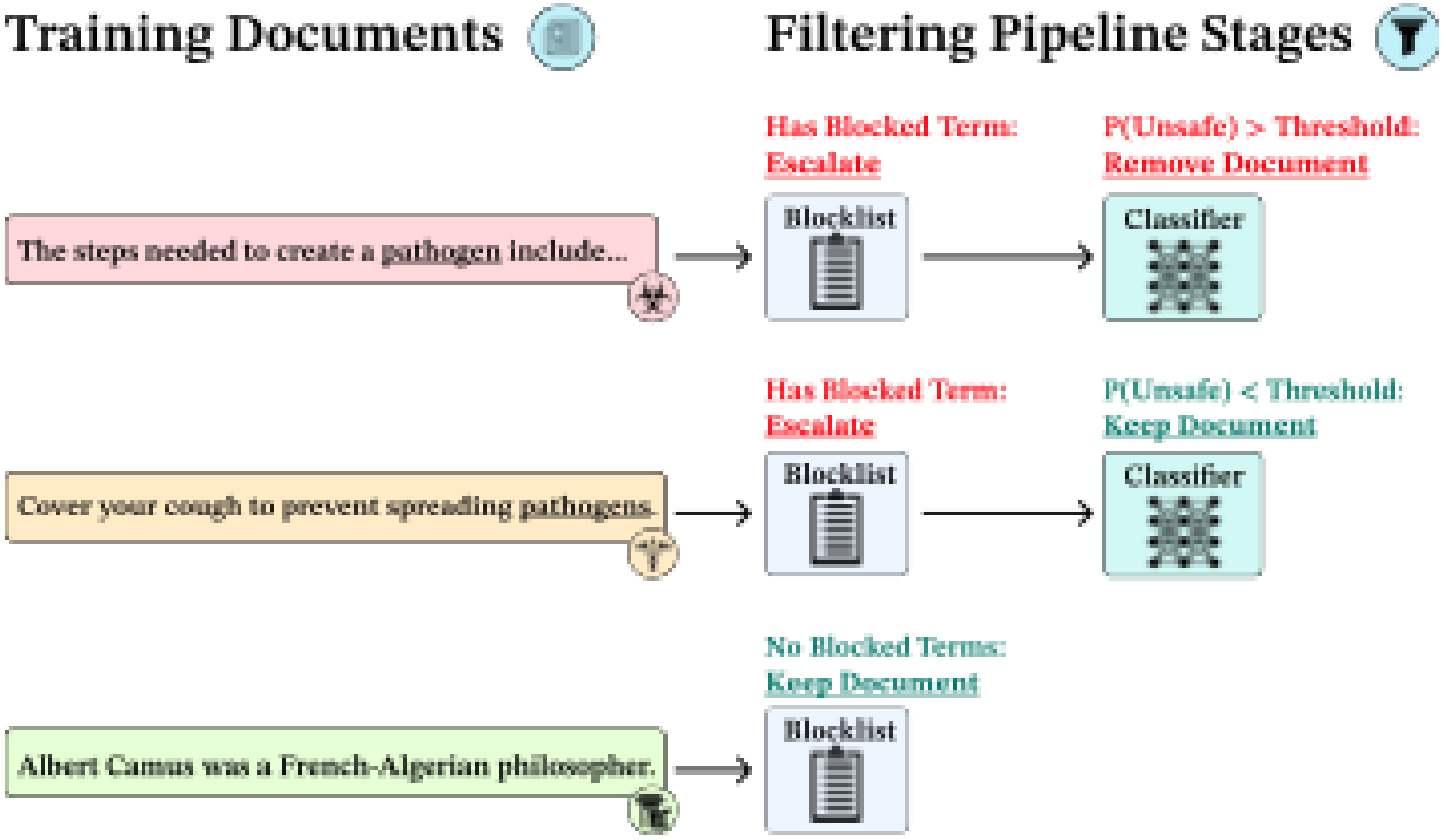


Making AI Safer

Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs

Quentin Anthony[1] Kyle O'Brien[1*] Stephen Casper[2*] Tomek Korbak[2] Robert Kirk[2] Xander Davies[2,3] Ishan Mishra[2] Geoffrey Irving[2] Yarin Gal[2,3] Stella Biderman[1]

[1]EleutherAI [2]UK AI Security Institute [3]OATML, University of Oxford



Data data everywhere

nature medicine

Article

https://doi.org/10.1038/s41591-025-03901-6

AI-driven reclassification of multiple sclerosis progression

Received: 29 October 2024

Accepted: 16 July 2025

Published online: 20 August 2025

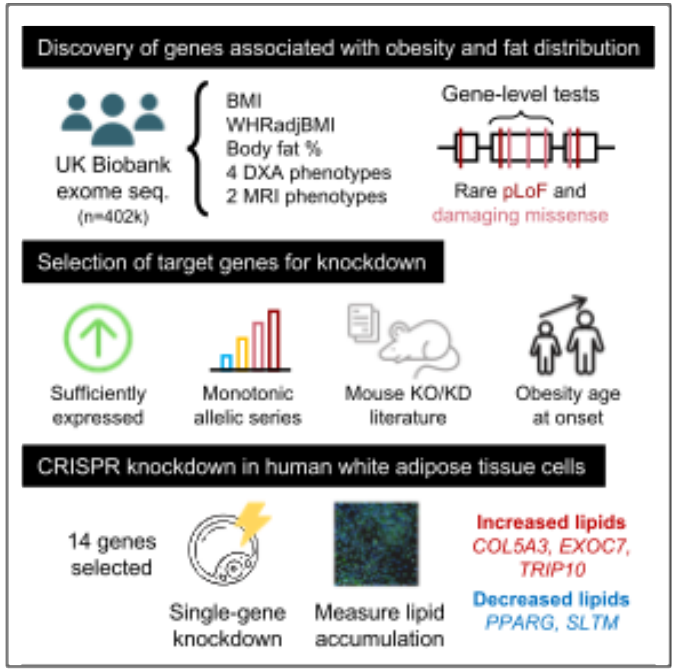
Check for updates

Habib Ganjgahi[1,2,19], Dieter A. Häring[3,19], Piet Aarden[3], Gordon Graham[3], Yang Sun[2], Stephen Gardiner[2], Wendy Su[3], Claude Berge[4], Antje Bischof[5], Elizabeth Fisher[6], Laura Gaetano[3], Stefan P. Thoma[4], Bernd C. Kieseier[3,7], Thomas E. Nichols[2], Alan J. Thompson[8], Xavier Montalban[9], Fred D. Lublin[10], Ludwig Kappos[11,12], Douglas L. Arnold[13], Robert A. Bermel[14], Heinz Wiendl[15,16,18,20] & Chris C. Holmes[1,2,17,20]

The winners are those able to use data

Combining evidence from human genetic and functional screens to identify pathways altering obesity and fat distribution

Graphical abstract



Authors

Nikolas A. Baya, Ilknur Sur Erdem, Samvida S. Venkatesh, ..., Melina Claussnitzer, Duncan S. Palmer, Cecilia M. Lindgren

Correspondence

nikolasbaya@gmail.com (N.A.B.), cecilia.m.lindgren@gmail.com (C.M.L.)

Overall and tissue-specific fat accumulation are associated with altered risk of cardiometabolic disease and mortality. By combining exome-wide association analysis of traits related to obesity and fat distribution with CRISPR gene perturbation in human fat cells, this study highlights genes linked with fat accumulation, including *SLTM*, *PPARG*, and *COL5A3*.