VITA TOTA DISCENDA EST

Transformers: You Need Attention!

Name: JIAHUI ZHUO GAO

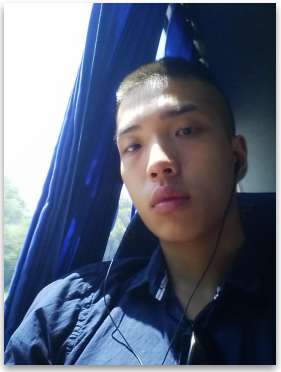
Institute: IFIC (UV - CSIC)

Date: 21/09/2026

Raymundus Lullus training program in advanced computing

VITA TOTA DISCENDA EST

About the lecturer



- **JIAHUI ZHUO GAO**
- **Ph.D., Universitat de València (IFIC-CSIC), 2026**
- **LHCb Trigger & Real-Time Analysis**
- **One of main developer & maintainer, LHCb HLT1 (Run 3)**
- **Convener, LHCb Retina DWT Simulation & Software WG**
- **2024 LHCb Early Career Scientist Award**

VITA TOTA DISCENDA EST

Transformers: You Need Attention!

35′ of fundamentals

- What is transformer?
- Does it have intelligence?
- How could we use it?

45′ hands-on

- Demonstrate the intelligence
- Example of usage

Raymundus Lullus training program in advanced computing

VITA TOTA DISCENDA EST

What is transformer?

VITA TOTA DISCENDA EST

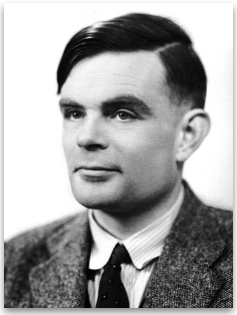
Introduction

The iPhone 4 moment

1950 — 2017 — 2022

Alan Mathison Turing

"Computing Machinery and Intelligence"



Transformer

"Attention Is All You Need"

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

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

Lukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

ChatGPT

1M users in 5 days

100M in 2 months

VITA TOTA DISCENDA EST

Transformer

The cat sat on the mat because it was

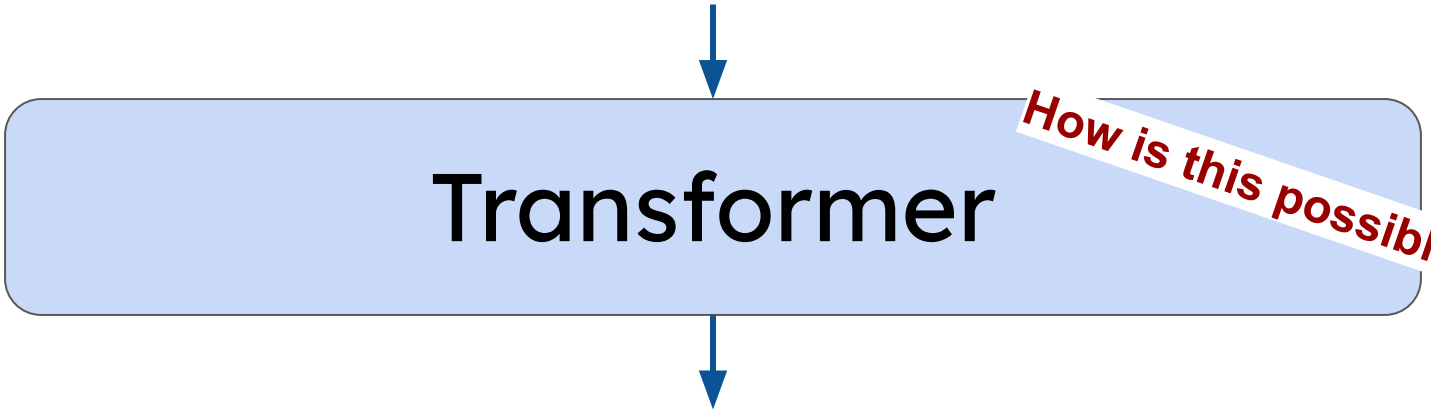
Transformer

The cat sat on the mat because it was tired

VITA TOTA DISCENDA EST

Transformer

The cat sat on the mat because it was



The cat sat on the mat because it was tired

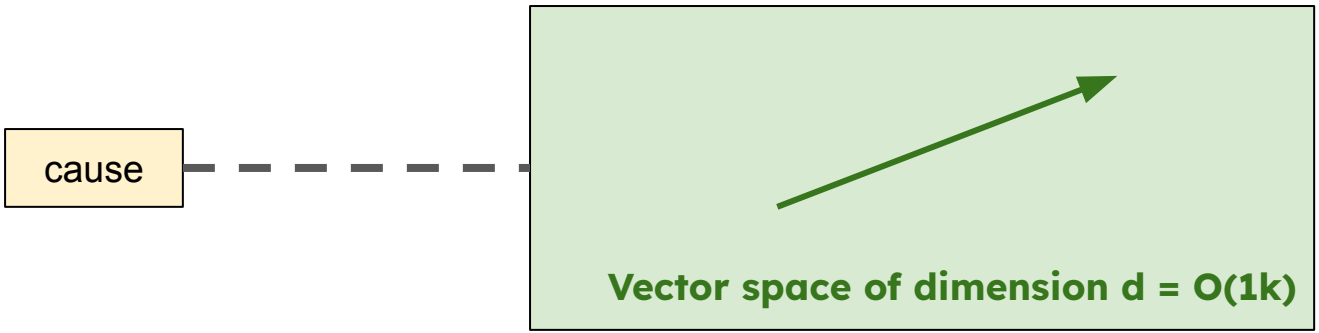
VITA TOTA DISCENDA EST

Text becomes a row of vectors

tokenize the text

The	cat	sat	on	the	mat	be	cause

embed the token into vector



VITA TOTA DISCENDA EST

Attention head

The	i-3
cat	i-2
sat	i-1
?	i

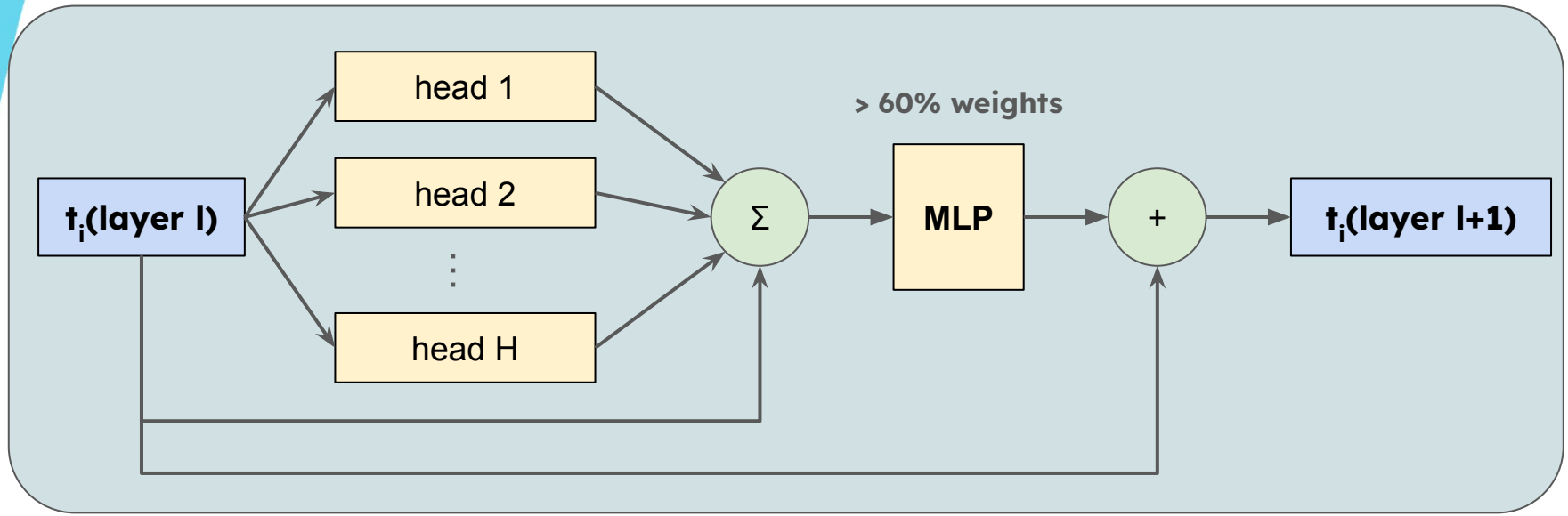
- W_Q, W_K, W_V are weights of attention head
- d_h is the dimension of attention head
- W_O is the transformation matrix

$$\mathbf{t}_i \leftarrow \mathbf{t}_i + \sum_{j=0}^{i} \mathrm{softmax}_j \left[\frac{(\mathbf{t}_i W_Q)(\mathbf{t}_j W_K)^\top}{\sqrt{d_h}} \right] (\mathbf{t}_j W_V) W_O$$

Q: query K: key V: value

VITA TOTA DISCENDA EST

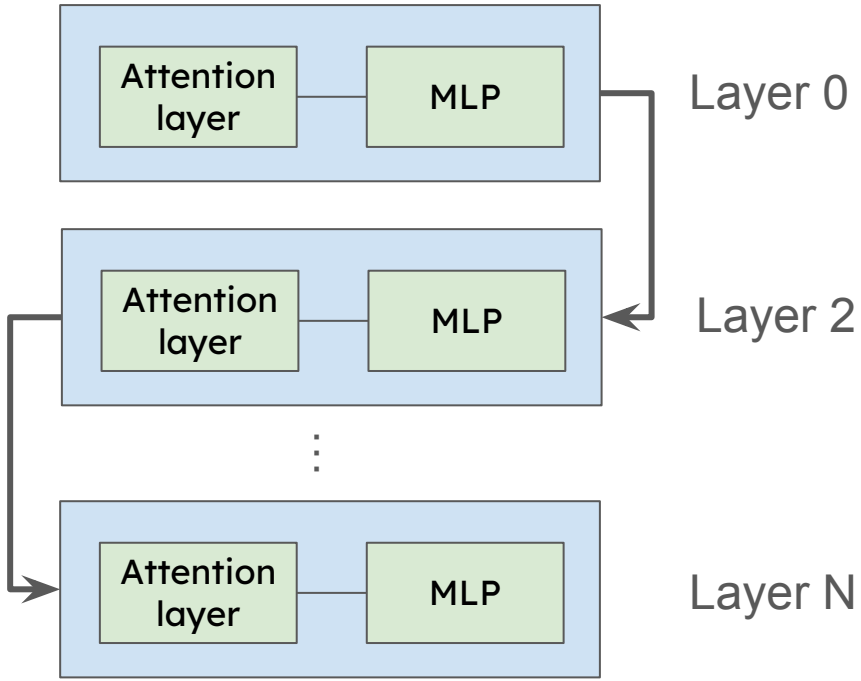
Attention layer



In high dimensions almost all directions are nearly orthogonal.

VITA TOTA DISCENDA EST

Transformer layers



Early layers

words, grammar, local context

Middle layers

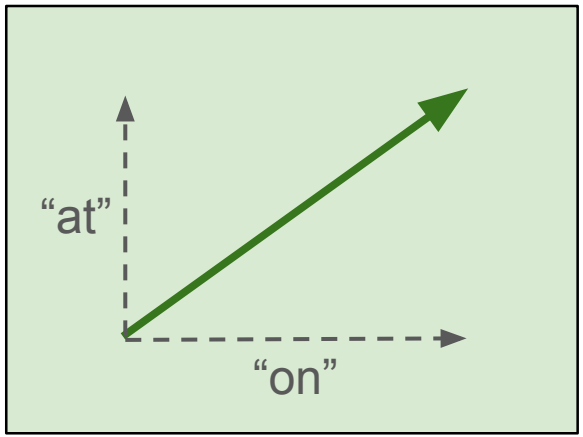
concepts, facts, reasoning

Late layers

choosing the next token

VITA TOTA DISCENDA EST

Transformer output



Its projection onto each vocabulary token gives that token's probability of coming next (after normalization).

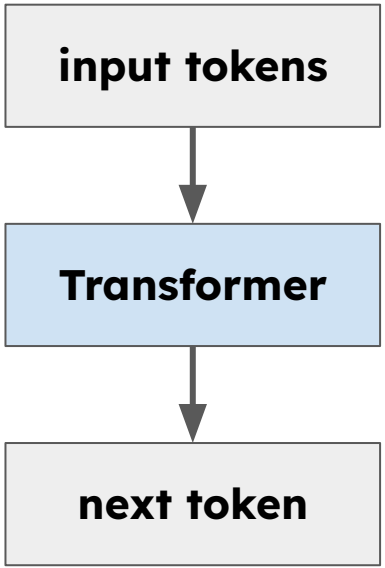
The output vector of the last layer is the prediction for the next token.

VITA TOTA DISCENDA EST

Does it have intelligence?

VITA TOTA DISCENDA EST

Transformer



That is all it does.

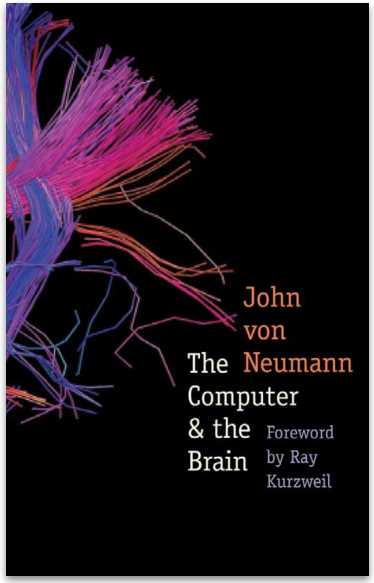
Predict the next token

Is this intelligence?

Or just a very good guess at the next token?

VITA TOTA DISCENDA EST

John von Neumann, 1958

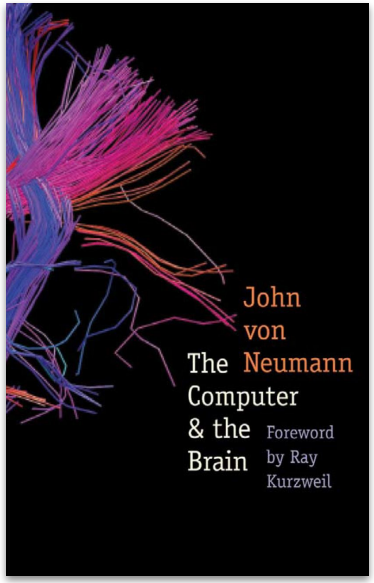


"…precisions like the ones mentioned above (10 to 12 decimals!) are altogether out of the question. The nervous system is a computing machine which manages to do its exceedingly complicated work at a rather low level of precision (2 to 3 decimals) … no known computing machine can operate reliably and significantly at such a low precision level.

… [this] leads not only to a low level of precision but also to a rather high level of reliability … if … a single pulse is lost, or even several pulses are lost … the relevant frequency, i.e. the meaning of the message, is only inessentially distorted."

VITA TOTA DISCENDA EST

John von Neumann, 1958



"The nervous system appears to be using a radically different system of notation from the ones we are familiar with in ordinary arithmetics and mathematics … the meaning is conveyed by the statistical properties of the message … a deterioration in arithmetics has been traded for an improvement in logics."

The brain's notation is statistical

So is the Transformer's.

VITA TOTA DISCENDA EST

So where is the understanding?

Early layers words, grammar, local context
Middle layers concepts, facts, reasoning
Late layers choosing the next token

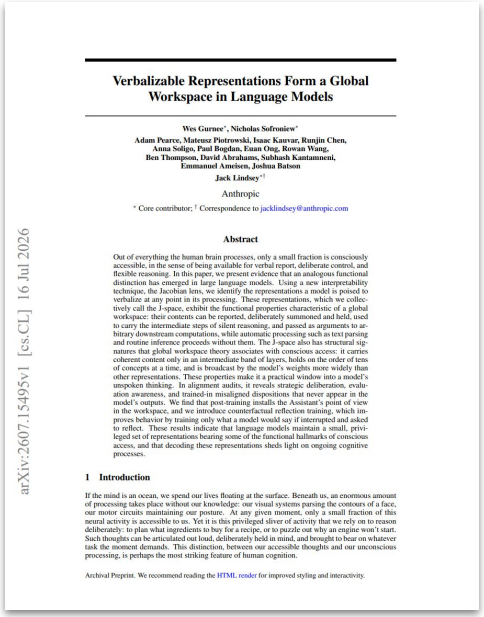
If transformer has understanding, it has to be written somewhere.

Can we read it?

Can we edit it?

VITA TOTA DISCENDA EST

j-space



arXiv:2607.15495v1 [cs.CL] 16 Jul 2026

Verbalizable Representations Form a Global Workspace in Language Models

Wes Gurnee*, Nicholas Sofroniew*
Adam Pearce, Mateusz Piotrowski, Isaac Kauvar, Runjin Chen,
Anna Soligo, Paul Bogdan, Euan Ong, Rowan Wang,
Ben Thompson, David Abrahams, Subhash Kantamneni,
Emmanuel Ameisen, Joshua Batson
Jack Lindsey*†
Anthropic
* Core contributor; † Correspondence to jacklindsey@anthropic.com

Abstract

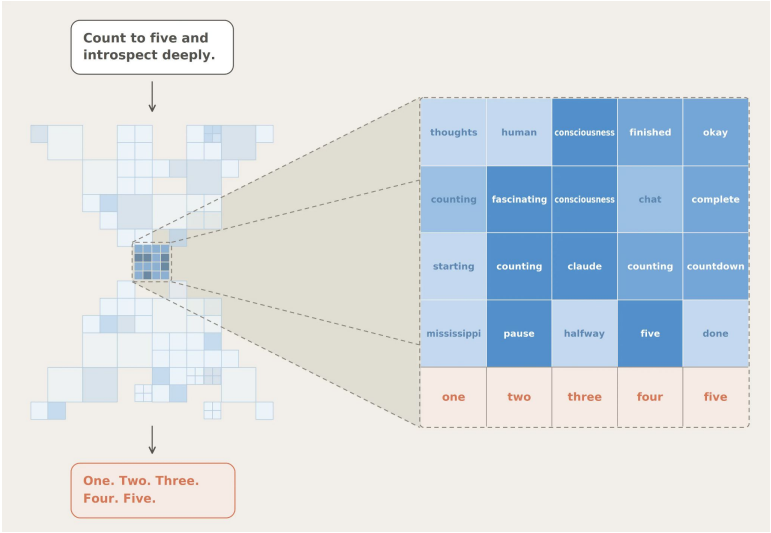
Out of everything the human brain processes, only a small fraction is consciously accessible, in the sense of being available for verbal report, deliberate control, and flexible reasoning. In this paper, we present evidence that an analogous functional distinction has emerged in large language models. Using a new interpretability technique, the Jacobian lens, we identify the representations a model is poised to verbalize at any point in its processing. These representations, which we collectively call the J-space, exhibit the functional properties characteristic of a global workspace: their contents can be reported, deliberately summoned and held, used to carry the intermediate steps of silent reasoning, and passed as arguments to arbitrary downstream computations, while automatic processing such as text parsing and routine inference proceeds without them. The J-space also has structural signatures that global workspace theory associates with conscious access: it carries coherent content only in an intermediate band of layers, holds on the order of tens of concepts at a time, and is broadcast by the model's weights more widely than other representations. These properties make it a practical window into a model's unspoken thinking. In alignment audits, it reveals strategic deliberation, evaluation awareness, and trained-in misaligned dispositions that never appear in the model's outputs. We find that post-training installs the Assistant's point of view in the workspace, and we introduce counterfactual reflection training, which improves behavior by training only what a model would say if interrupted and asked to reflect. These results indicate that language models maintain a small, privileged set of representations bearing some of the functional hallmarks of conscious access, and that decoding these representations sheds light on ongoing cognitive processes.

1 Introduction

If the mind is an ocean, we spend our lives floating at the surface. Beneath us, an enormous amount of processing takes place without our knowledge: our visual systems parsing the contours of a face, our motor circuits maintaining our posture. At any given moment, only a small fraction of this neural activity is accessible to us. Yet it is this privileged sliver of activity that we rely on to reason deliberately: to plan what ingredients to buy for a recipe, or to puzzle out why an engine won't start. Such thoughts can be articulated out loud, deliberately held in mind, and brought to bear on whatever task the moment demands. This distinction, between our accessible thoughts and our unconscious processing, is perhaps the most striking feature of human cognition.

Archival Preprint. We recommend reading the HTML render for improved styling and interactivity.

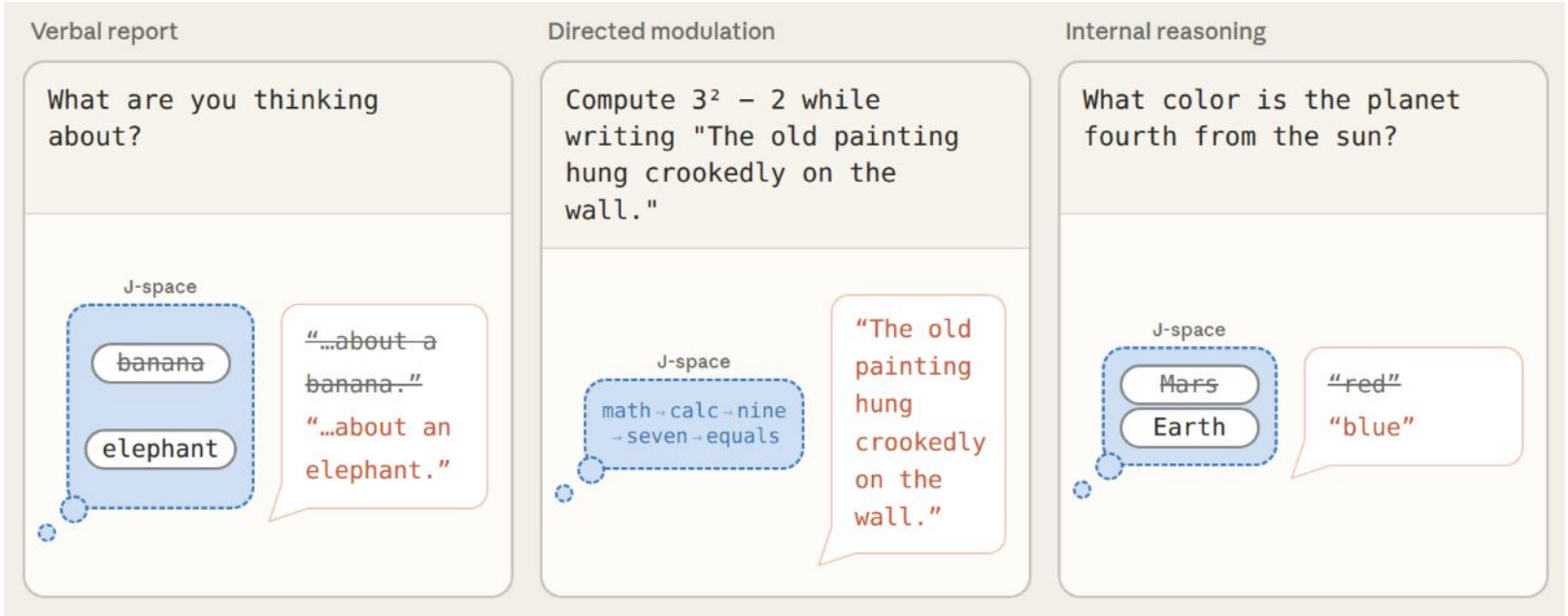
$$J_\ell = \mathbb{E}_{t,\, t' \geq t,\, \text{prompt}} \left[\frac{\partial h_{\text{final},t'}}{\partial h_{\ell,t}} \right], \qquad \text{lens}(h_\ell) = \text{softmax}(W_U \, \text{norm}(J_\ell h_\ell))$$



Raymundus Lullus training program in advanced computing

VITA TOTA DISCENDA EST

Functional roles of j-space



VITA TOTA DISCENDA EST

Functional roles of j-space

With the J-lens we can watch the model think.

VITA TOTA DISCENDA EST

How could we use it?

VITA TOTA DISCENDA EST

Vision Banana

Google DeepMind 2026-06-05

Image Generators are Generalist Vision Learners

Valentin Gabeur*, Shangbang Long*, Songyou Peng*, Paul Voigtlaender*, Shuyang Sun*, Yanan Bao*, Karen Truong*, Zhicheng Wang*, Wenlei Zhou*, Jonathan T. Barron*, Kyle Genova*, Nithish Kannen*, Sherry Ben*, Yandong Li*, Mandy Guo*, Suhas Yogin*, Yiming Gu*, Huizhong Chen*, Oliver Wang*, Saining Xie*, Howard Zhou*, Kaiming He*, Thomas Funkhouser*, Jean-Baptiste Alayrac* and Radu Soricut*
*project leads and equal contributions, *core contributors, *project advisors, *leadership sponsors

Recent works show that image and video generators exhibit zero-shot visual understanding behaviors, in a way reminiscent of how Large Language Models (LLMs) develop emergent capabilities of language understanding and reasoning from generative pretraining. While it has long been conjectured that the ability to create visual content implies an ability to understand it, there has been limited work demonstrating that generalist image generators can achieve state-of-the-art understanding capabilities on many different vision tasks. In this work, we demonstrate that image generation training serves a role similar to LLM pretraining, and lets models learn powerful and general visual representations that enable state-of-the-art performance on various vision tasks. We introduce *Vision Banana*, a generalist model built by instruction-tuning *Nano Banana Pro* (NBP) on a mixture of its original training data alongside a small amount of vision task data. By parameterizing the output space of vision tasks as RGB images, we seamlessly reframe perception as image generation. Our generalist model, *Vision Banana*, achieves state-of-the-art results on a variety of vision tasks involving both 2D and 3D understanding, beating or rivaling zero-shot domain-specialists, including *Segment Anything Model 3* on segmentation tasks, and the *Depth Anything* series on metric depth estimation. We show that these results can be achieved with lightweight instruction-tuning without sacrificing the base model's image generation capabilities. The superior results suggest that image generation pretraining is a generalist vision learner. It also shows that image generation serves as a unified and universal interface for vision tasks, similar to text generation's role in language understanding and reasoning. We could be witnessing a major paradigm shift for computer vision, where generative vision pretraining takes a central role in building Foundational Vision Models for both generation and understanding.

Project Page: vision-banana.github.io

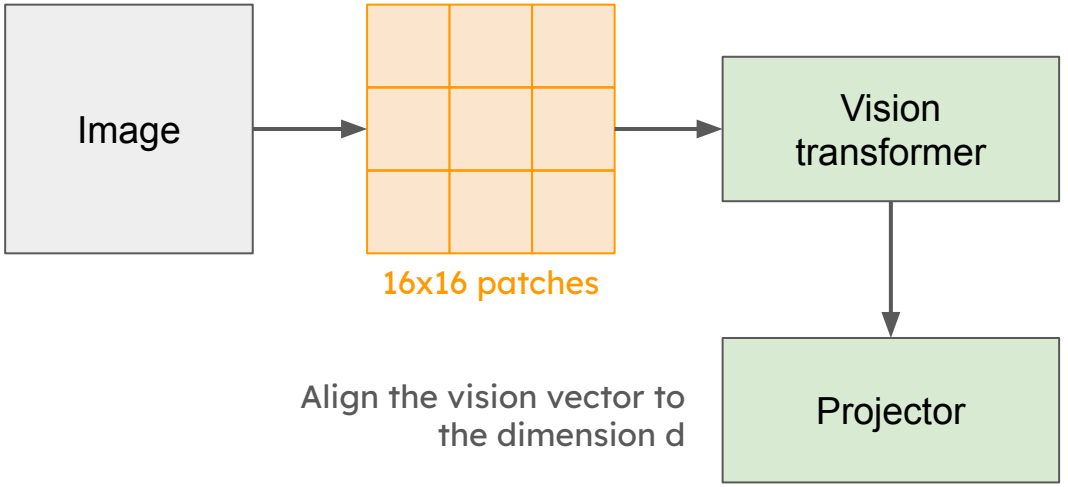
1. Introduction

In recent years, advanced image and video generation models (Black Forest Labs, 2025; ByteDance, 2026; Google, 2025a,b; Luma, 2026; OpenAI, 2026) have demonstrated unprecedented generation capabilities, synthesizing highly complex, high-fidelity visual content with precise semantic control. This remarkable capability for visual creation suggests that these models possess a deep, internalized comprehension of the visual world's underlying structures, semantics, and relationships. However, leading methods on visual representation learning in general do not belong to the family of generative modeling. Instead, they include supervised discriminative learning (Dehghani et al., 2023; Dosovitskiy et al., 2020; Krizhevsky et al., 2012), contrastive learning (Chen et al., 2020b,c; He et al., 2020; Radford et al., 2021; Tschannen et al., 2025; Zhai et al., 2023), bootstrapping (Caron et al., 2021; Grill et al., 2020), auto-encoding (Bao et al., 2021; Chen et al., 2024; He et al., 2022) among others, and their combinations (Cao et al., 2026; Oquab et al., 2023; Siméoni et al., 2025; Zhou et al., 2021). Early efforts in generative vision pretraining (Bai et al., 2024; Chen et al., 2020a) have shown promising scaling behaviors but their effectiveness has lagged behind non-generative models.

Correspondence: vision-banana@google.com

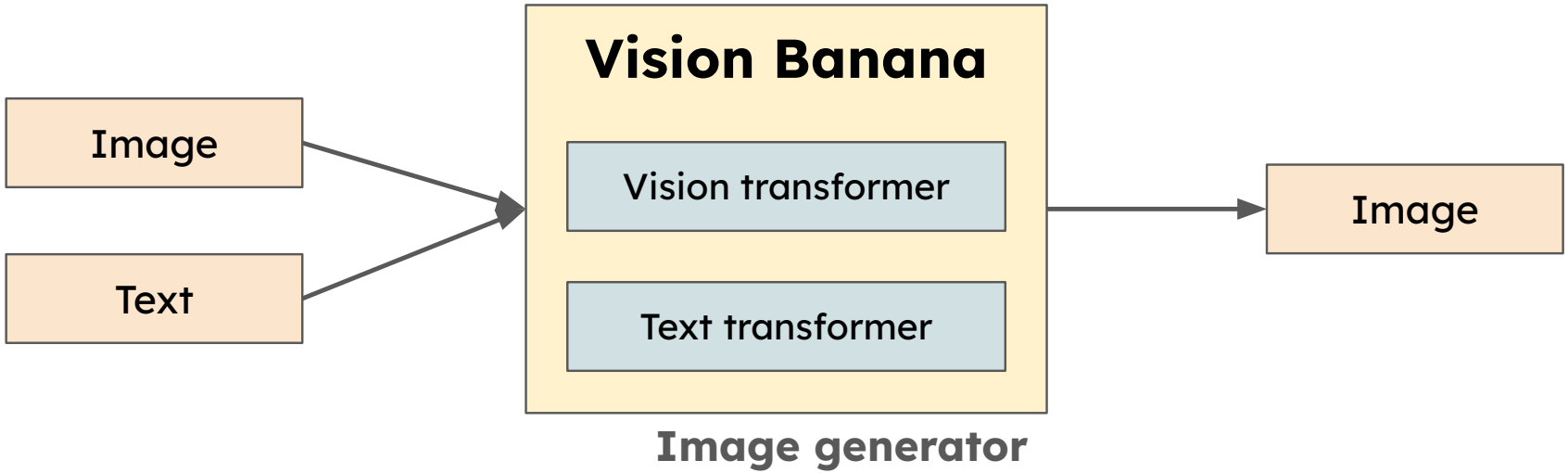
arXiv:2604.20329v3 [cs.CV] 3 Jun 2026

How images enter the transformer?



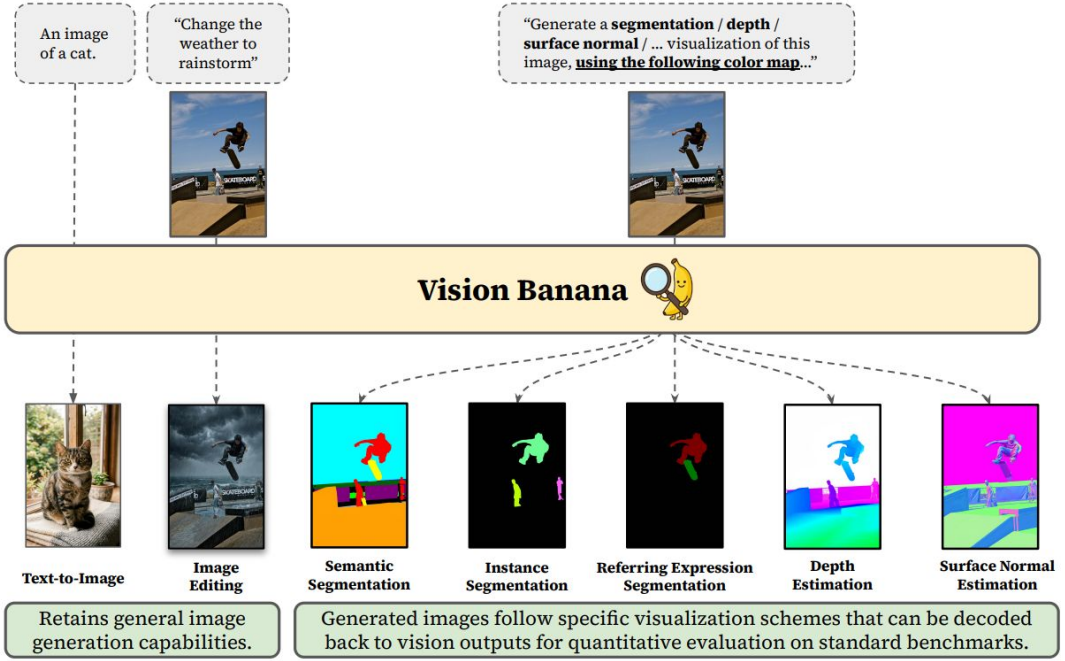
VITA TOTA DISCENDA EST

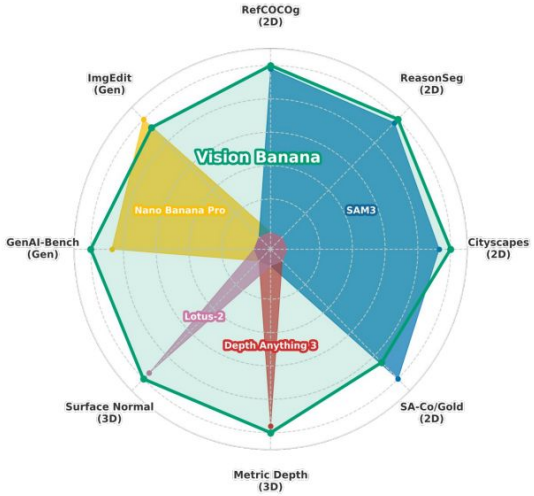
Vision Banana



VITA TOTA DISCENDA EST

Vision Banana





Vision Banana is a generalist vision model in both visual generation and understanding, surpassing or rivaling specialist models.

VITA TOTA DISCENDA EST

Give Transformer new knowledge

In the context

- No setup
- Limited by context size
- Attention gets diluted in long inputs

Fine tunning

- Knowledge goes into the weights
- Expensive, retrain on every update
- Good for style and format, unreliable for facts

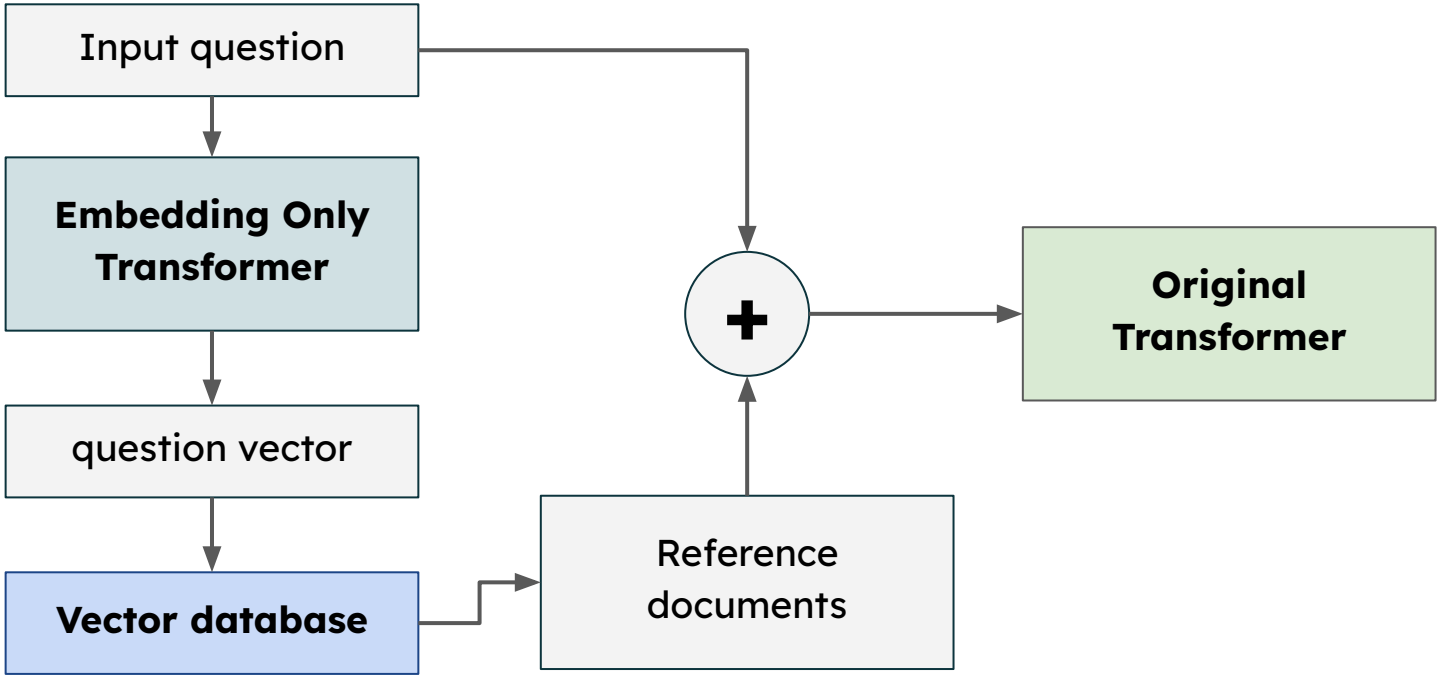
RAG

Retrieval-Augmented Generation

- Retrieve the relevant information, put them in the context
- Scales to large document sets, cheap to update

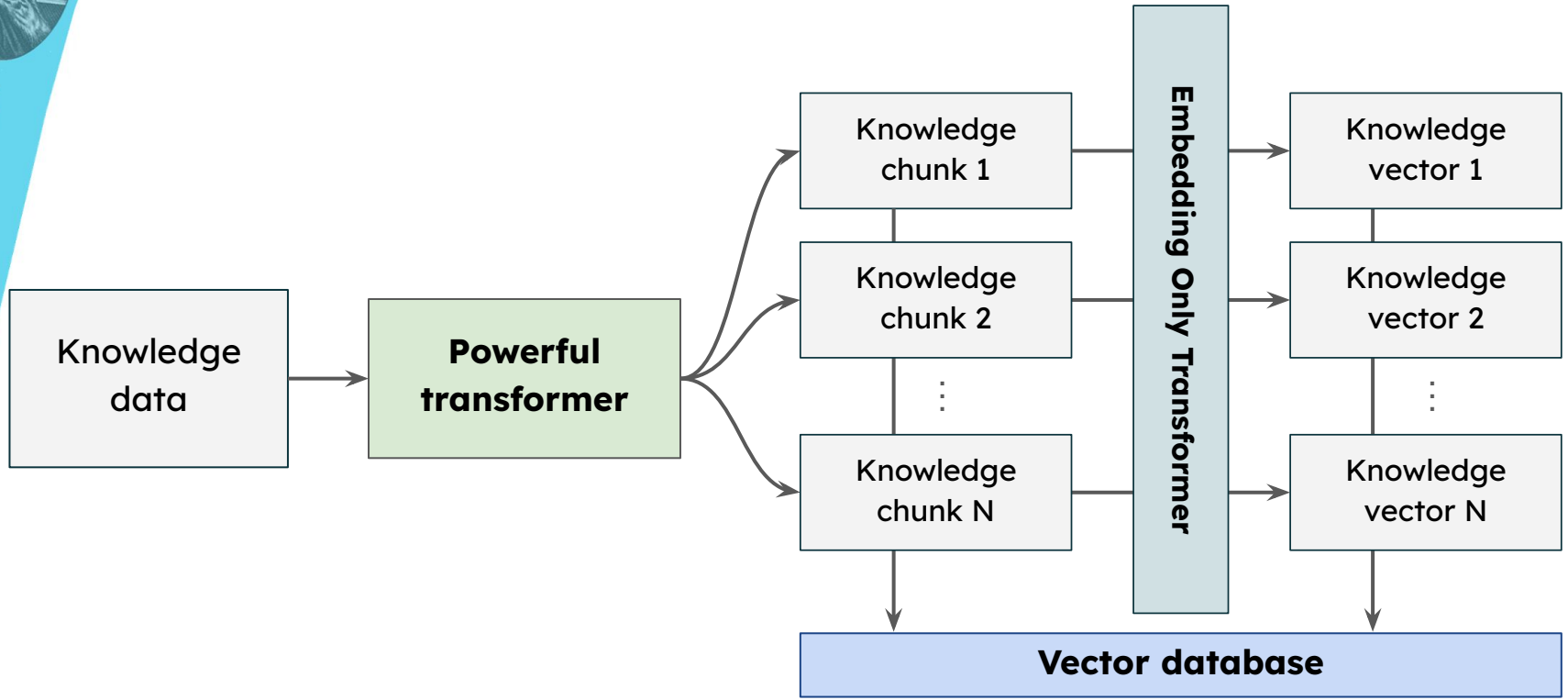
VITA TOTA DISCENDA EST

Retrieval-Augmented Generation

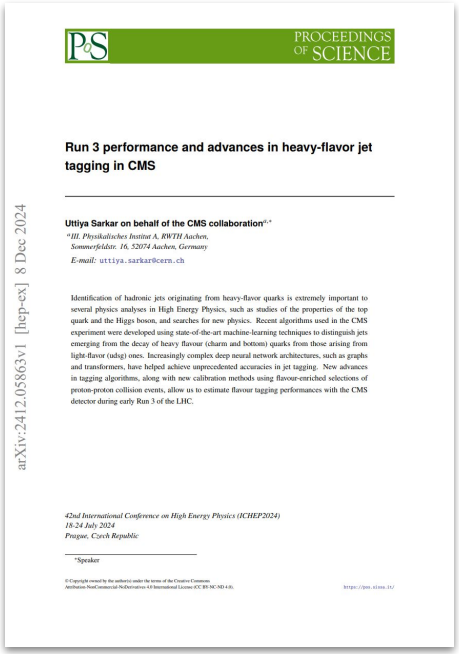


VITA TOTA DISCENDA EST

Retrieval-Augmented Generation



Transformer in HEP



PROCEEDINGS OF SCIENCE

Run 3 performance and advances in heavy-flavor jet tagging in CMS

Uttiya Sarkar on behalf of the CMS collaboration[a,*]
[a] *III. Physikalisches Institut A, RWTH Aachen, Sommerfeldstr. 16, 52074 Aachen, Germany*
E-mail: uttiya.sarkar@cern.ch

Identification of hadronic jets originating from heavy-flavor quarks is extremely important to several physics analyses in High Energy Physics, such as studies of the properties of the top quark and the Higgs boson, and searches for new physics. Recent algorithms used in the CMS experiment were developed using state-of-the-art machine-learning techniques to distinguish jets emerging from the decay of heavy flavour (charm and bottom) quarks from those arising from light-flavor (udsg) ones. Increasingly complex deep neural network architectures, such as graphs and transformers, have helped achieve unprecedented accuracies in jet tagging. New advances in tagging algorithms, along with new calibration methods using flavour-enriched selections of proton-proton collision events, allow us to estimate flavour tagging performances with the CMS detector during early Run 3 of the LHC.

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

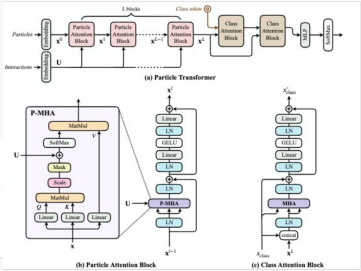
*Speaker

https://pos.sissa.it/

Transformers Inclusive Flavour Tagging

- Two separate IFT Transformer, B_s^0 and B^0. trained on $B_s^0 \to D_s^- \pi^+$ and $B_s^0 \to J/\psi\,\phi$ simulations, and $B^0 \to J/\psi\,K^{*0}$
- Event collision represented as **point clouds**. Particle attention blocks learns contextual relations among tracks and captures both local and global event information.
- One prediction per event, no combination of decisions from SS or OS taggers [*CMS-PAS-BPH-26-005*]

$$U = \begin{aligned} \Delta &= \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2}, \\ k_T &= \min\left(p_{T,a},\ p_{T,b}\right)\Delta, \\ z &= \min\left(p_{T,a},\ p_{T,b}\right) / \left(p_{T,a} + p_{T,b}\right), \\ m^2 &= (E_a + E_b)^2 - \|\mathbf{p}_a + \mathbf{p}_b\|^2. \end{aligned}$$



ParticleTransformer. Qu et al 2024. Particle Transformer for Jet Tagging.

VITA TOTA DISCENDA EST

Summary

VITA TOTA DISCENDA EST

Summary

- The Transformer is the most important AI architecture of our time
- Its interface is next-token prediction, but it is not a probability machine
- It thinks and reasons. With the J-lens we can watch it happen
- To use its intelligence, what matters is how we talk to it
  - Vision is a matter of interface: Nano Banana
  - Knowledge is a matter of interface: RAG

VITA TOTA DISCENDA EST

Thanks!

Raymundus Lullus training program in advanced computing

VITA TOTA DISCENDA EST

Hands-on session

VITA TOTA DISCENDA EST

Hands-on session

45′ hands-on

j-lens experiments

- Demonstrate the intelligence
- Try it yourself

RAG exercise

- Demonstrate how to give Transformer new knowledge

VITA TOTA DISCENDA EST

RAG exercise

LightRAG: Simple and Fast Retrieval-Augmented Generation

GITHUB TRENDING #2 Repository Of The Day

PROJECT PAGE | ARXIV 2410.05779 | STARS 40K

PYTHON 3.10 | PYPI V1.5.7

DISCORD COMMUNITY | WECHAT GROUP

CN 中文版 | US ENGLISH | JP 日本語版 | ID BAHASA INDONESIA

downloads 2M | HVTrust 84.2

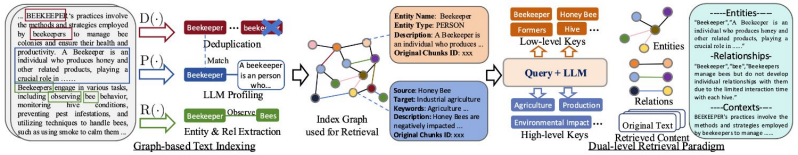


Figure 1: Overall architecture of the proposed LightRAG framework.

GitHub

github.com/hkuds/lightrag

Raymundus Lullus training program in advanced computing

VITA TOTA DISCENDA EST

Hands-on session

All the code used today is on

GitHub

`github.com/JiahuiZhuo/jlens-explorer`

`github.com/JiahuiZhuo/lightrag-pdf-kb`