From Truisms to Transformation:

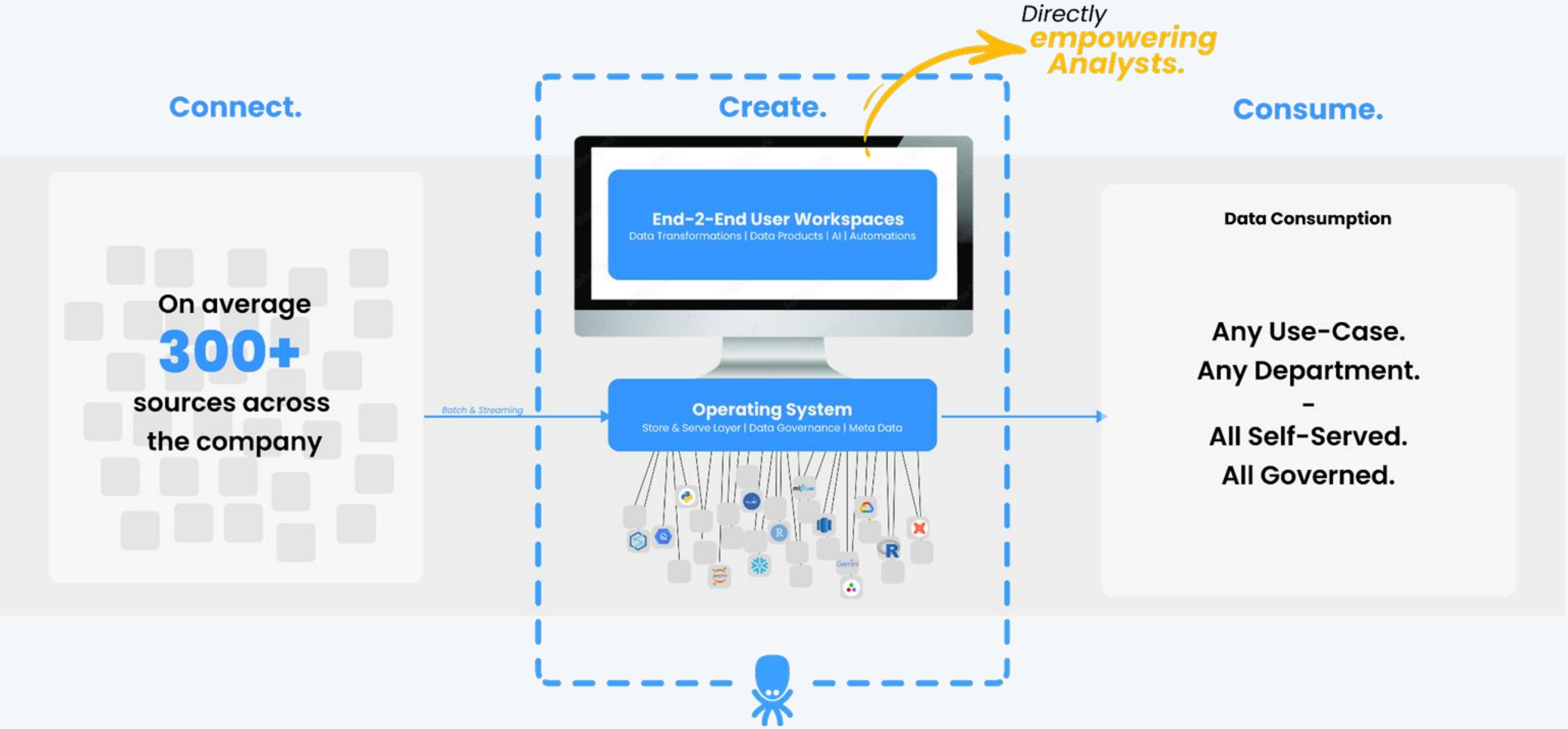
Achieving Tangible Results in AI and Data Management

Jordan Burger

Head of AI Lab, Keboola

Unlock the **full potential of enterprise data.**

Keboola



Directly *empowering Analysts.*

Connect.

On average

300+

sources across the company

Create.

End-2-End User Workspaces

Data Transformations | Data Products | AI | Automations

Operating System

Store & Serve Layer | Data Governance | Meta Data

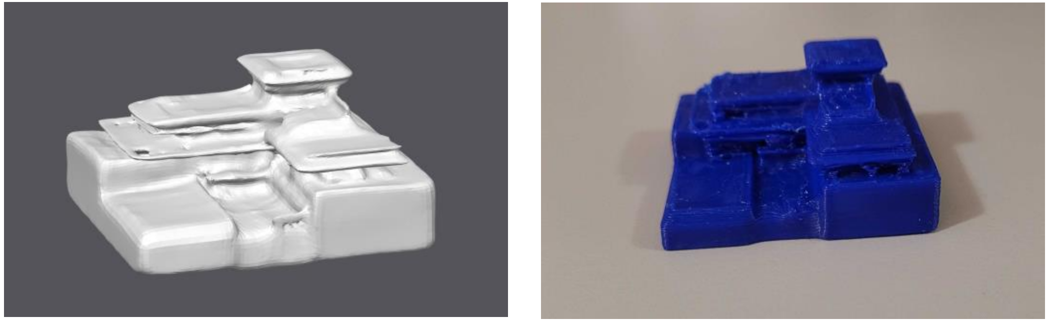
Consume.

Data Consumption

Any Use-Case.
Any Department.
-
All Self-Served.
All Governed.

About Me

- Research Lead; AI Applications
- Exploring and implementing AI technologies in data pipelines
- Addressing GenAI disruption in data and analytics



'A house in the style of Frank Lloyd Wright's Fallingwater'

Generated with Point-E from OpenAI

A broken clock is wrong...

Garbage
in

Garbage
Out

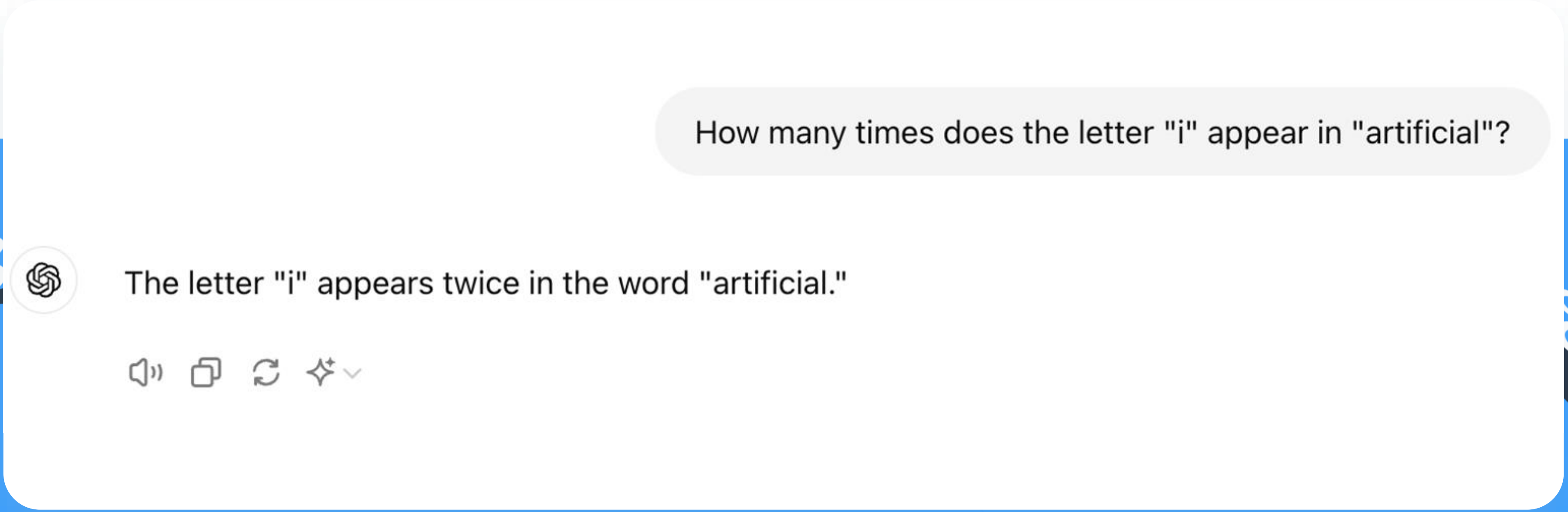
… But it's still right twice a day

Garbage in

›

Gospel Out

Struggling with Outliers;
Unable to handle genuine novelty



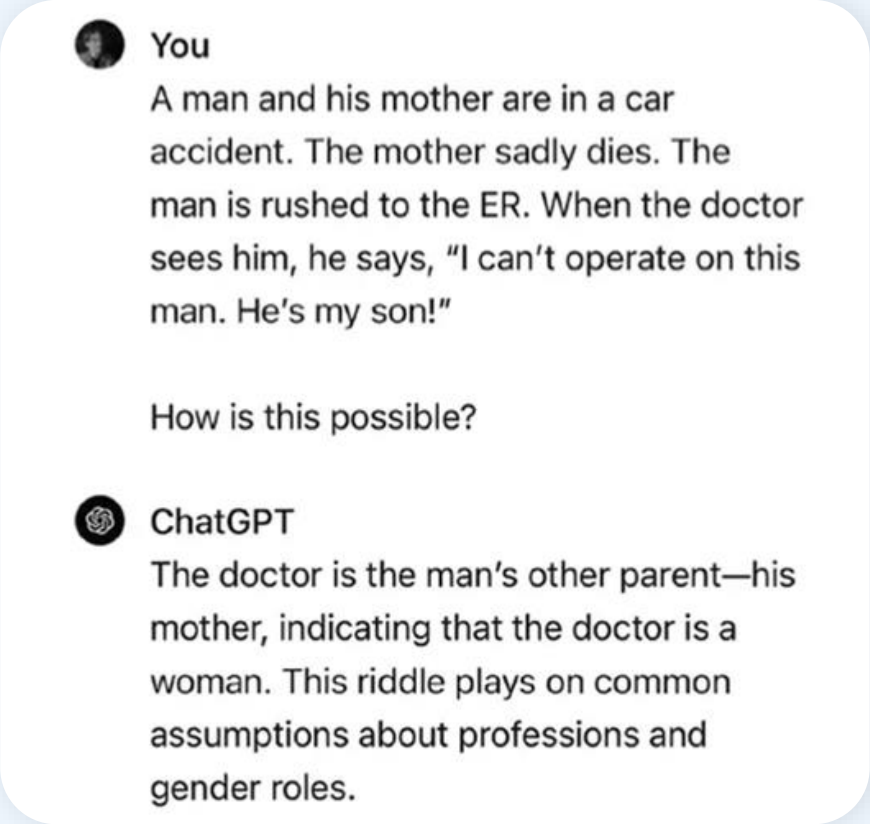
How many times does the letter "i" appear in "artificial"?

The letter "i" appears twice in the word "artificial."

Keboola

Struggling with Outliers;

Unable to handle genuine novelty



You

A man and his mother are in a car accident. The mother sadly dies. The man is rushed to the ER. When the doctor sees him, he says, “I can’t operate on this man. He’s my son!”

How is this possible?

ChatGPT

The doctor is the man’s other parent—his mother, indicating that the doctor is a woman. This riddle plays on common assumptions about professions and gender roles.

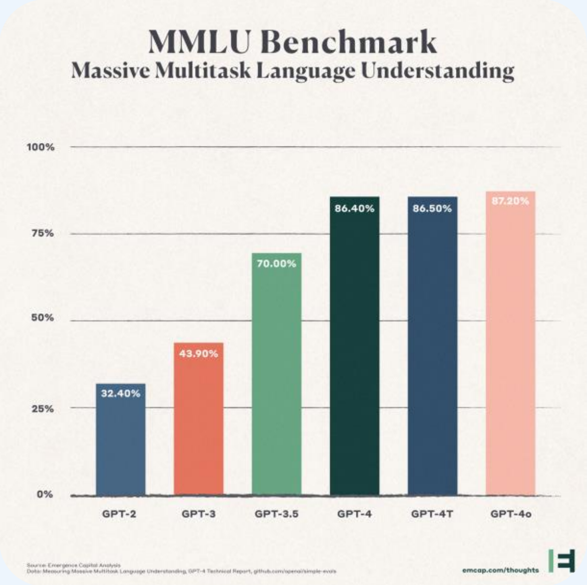
Credit: @colin_fraser

Struggling with Outliers;
Unable to handle genuine novelty

Employee Salaries	
Employee ID	Salary
12	500 CZK/Month
116	1.000.000$/year
1243	$75,000/year
242344	3.000/month
555	250.00 Kc/Year
63465	100.000 USD/month

Employee Salaries			
Employee ID	Currency	Salary	Pay Frequency
12	USD	$5,000	Annual
116	USD	$1,000,000	Annual
1243	USD	$75,000	Annual
242344	USD	$360,000	Annual
555	USD	$32,000	Annual
63465	USD	1,200,00	Annual

The Plateau was always (kind of) here

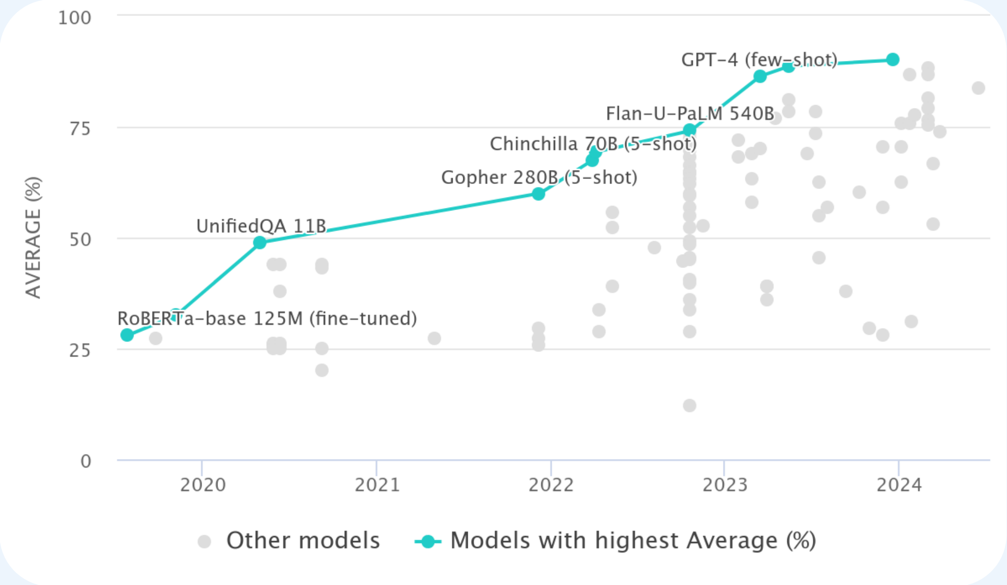


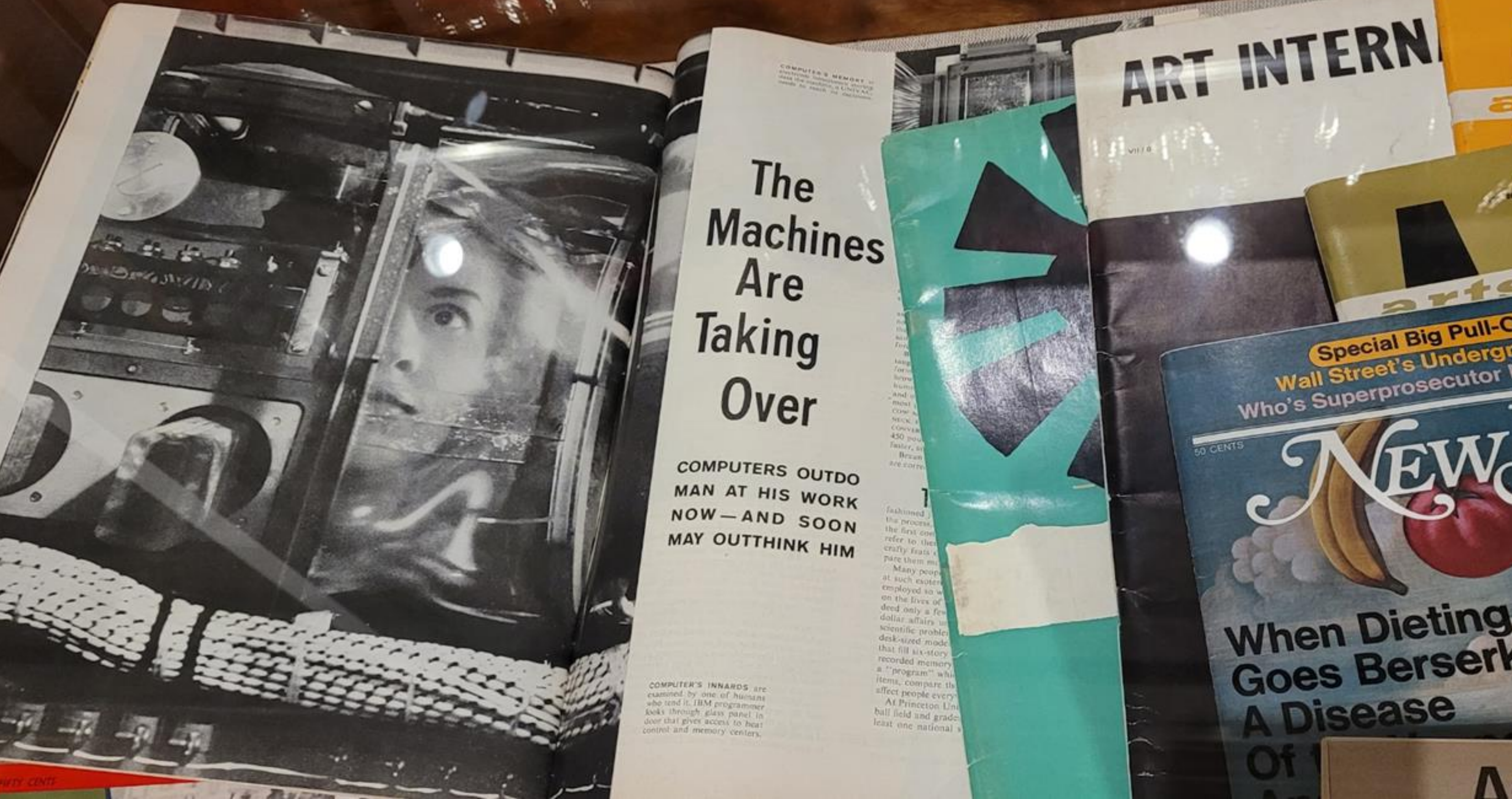
MMLU Benchmark

Massive Multitask Language Understanding

Source: Emergence Capital Analysis
Data: Measuring Massive Multitask Language Understanding, GPT-4 Technical Report, github.com/openai/simple-evals

emcap.com/thoughts





COMPUTER'S MEMORY is electronic [illegible] storing data the machine, a UNIVAC, needs to reach its decisions.

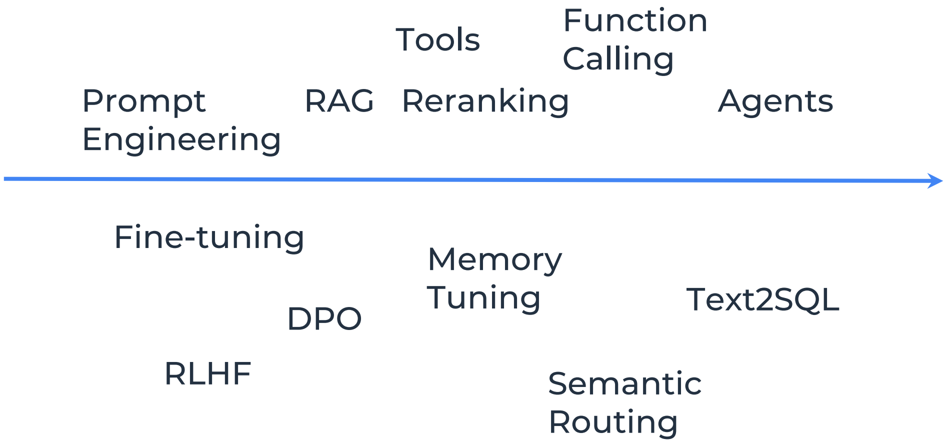
The Machines Are Taking Over

COMPUTERS OUTDO MAN AT HIS WORK NOW—AND SOON MAY OUTTHINK HIM

COMPUTER'S INNARDS are examined by one of humans who tend it. IBM programmer looks through glass panel in door that gives access to heat control and memory centers.

But where’s the value?

> **Investment in AI has reached a new high with a focus on generative AI, which, in most cases, has yet to deliver its anticipated business value.**
>
> **Gartner - Hype Cycle for Artificial Intelligence, 2024**



Tools

Function Calling

Prompt Engineering

RAG

Reranking

Agents

Fine-tuning

Memory Tuning

DPO

Text2SQL

RLHF

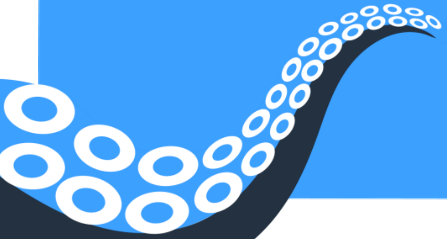
Semantic Routing

But where’s the value?

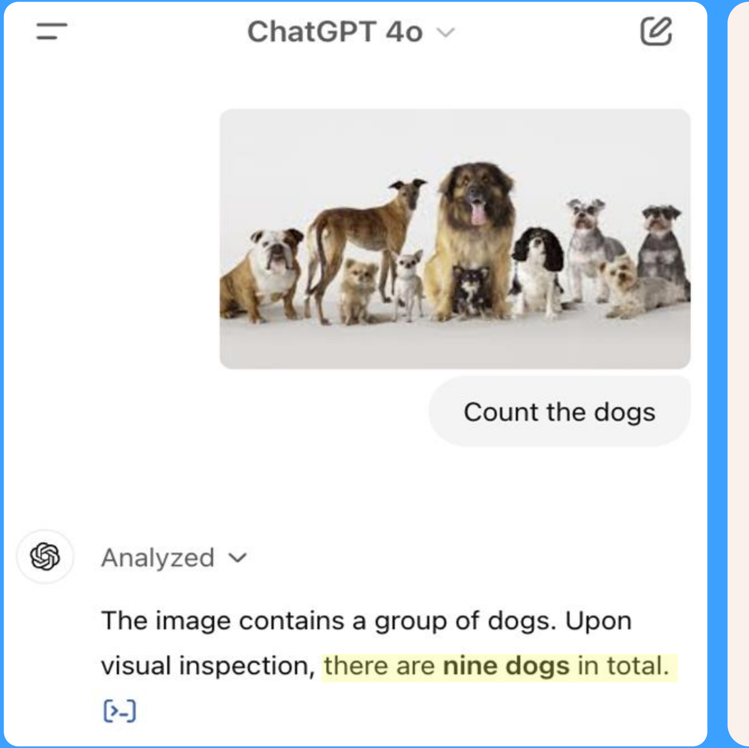
"

> **The revenue isn’t there yet, and might never come.**
>
> **Gary Marcus**

All of the applied research going on can be boiled down to one thing: “These models are cool; now how can I get them to do what I actually need?”



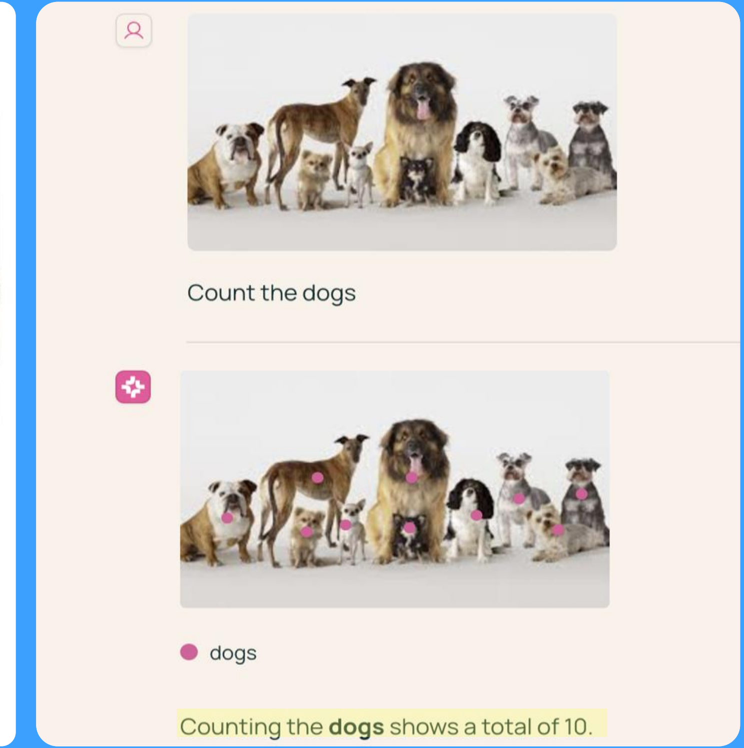
What got us here? Does it matter?



Count the dogs

Analyzed

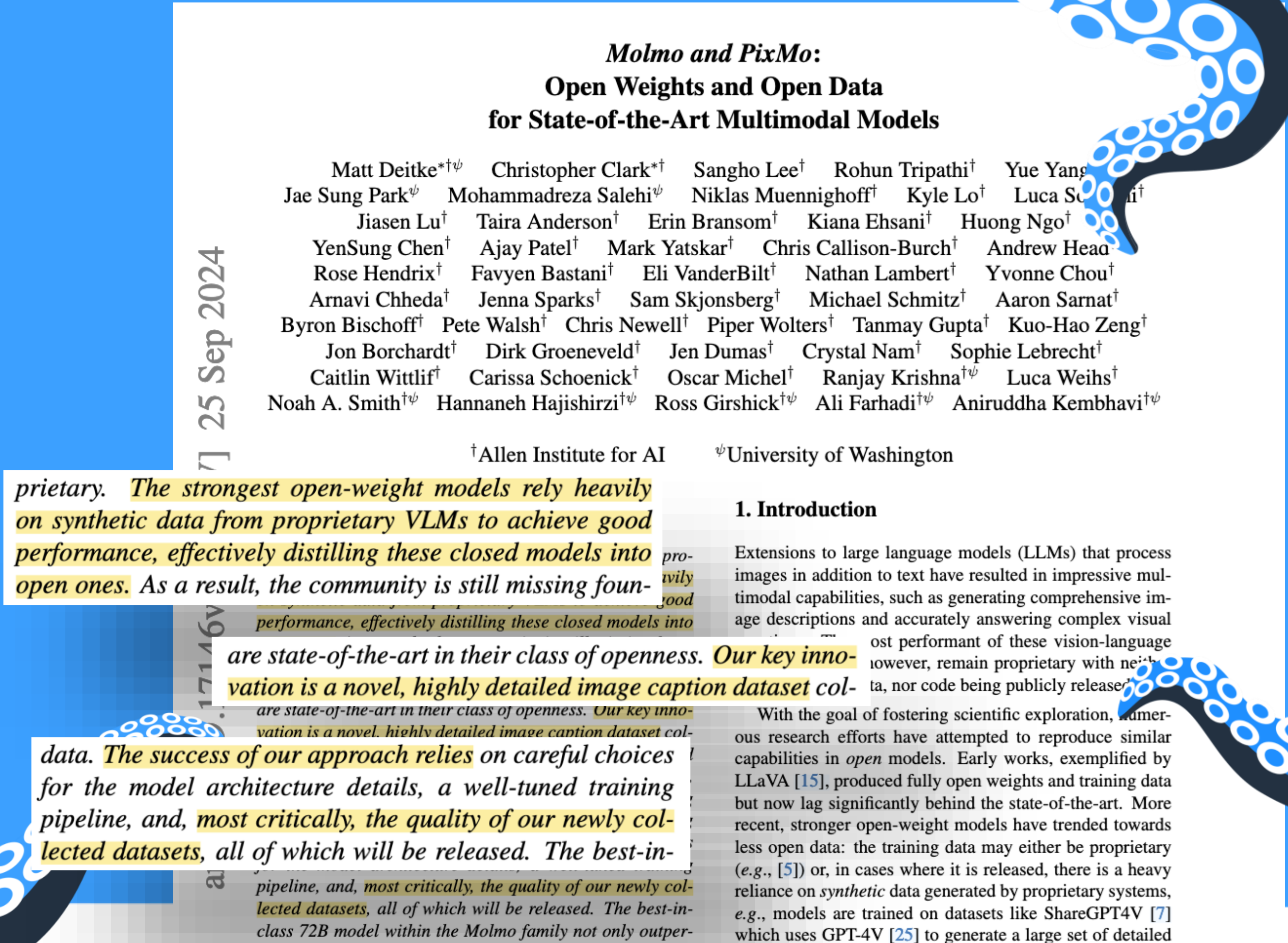
The image contains a group of dogs. Upon visual inspection, there are **nine dogs** in total.


Count the dogs

● dogs

Counting the **dogs** shows a total of 10.

Keboola



Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models

Matt Deitke[*†ψ] Christopher Clark[*†] Sangho Lee[†] Rohun Tripathi[†] Yue Yang
Jae Sung Park[ψ] Mohammadreza Salehi[ψ] Niklas Muennighoff[†] Kyle Lo[†] Luca Sc[illegible]i[†]
Jiasen Lu[†] Taira Anderson[†] Erin Bransom[†] Kiana Ehsani[†] Huong Ngo[†]
YenSung Chen[†] Ajay Patel[†] Mark Yatskar[†] Chris Callison-Burch[†] Andrew Head
Rose Hendrix[†] Favyen Bastani[†] Eli VanderBilt[†] Nathan Lambert[†] Yvonne Chou[†]
Arnavi Chheda[†] Jenna Sparks[†] Sam Skjonsberg[†] Michael Schmitz[†] Aaron Sarnat[†]
Byron Bischoff[†] Pete Walsh[†] Chris Newell[†] Piper Wolters[†] Tanmay Gupta[†] Kuo-Hao Zeng[†]
Jon Borchardt[†] Dirk Groeneveld[†] Jen Dumas[†] Crystal Nam[†] Sophie Lebrecht[†]
Caitlin Wittlif[†] Carissa Schoenick[†] Oscar Michel[†] Ranjay Krishna[†ψ] Luca Weihs[†]
Noah A. Smith[†ψ] Hannaneh Hajishirzi[†ψ] Ross Girshick[†ψ] Ali Farhadi[†ψ] Aniruddha Kembhavi[†ψ]

[†]Allen Institute for AI [ψ]University of Washington

25 Sep 2024

prietary. The strongest open-weight models rely heavily on synthetic data from proprietary VLMs to achieve good performance, effectively distilling these closed models into open ones. As a result, the community is still missing foun-

are state-of-the-art in their class of openness. Our key innovation is a novel, highly detailed image caption dataset col-

data. The success of our approach relies on careful choices for the model architecture details, a well-tuned training pipeline, and, most critically, the quality of our newly collected datasets, all of which will be released. The best-in-

pro- vily good performance, effectively distilling these closed models into

are state-of-the-art in their class of openness. Our key innovation is a novel, highly detailed image caption dataset col-

pipeline, and, most critically, the quality of our newly collected datasets, all of which will be released. The best-in-

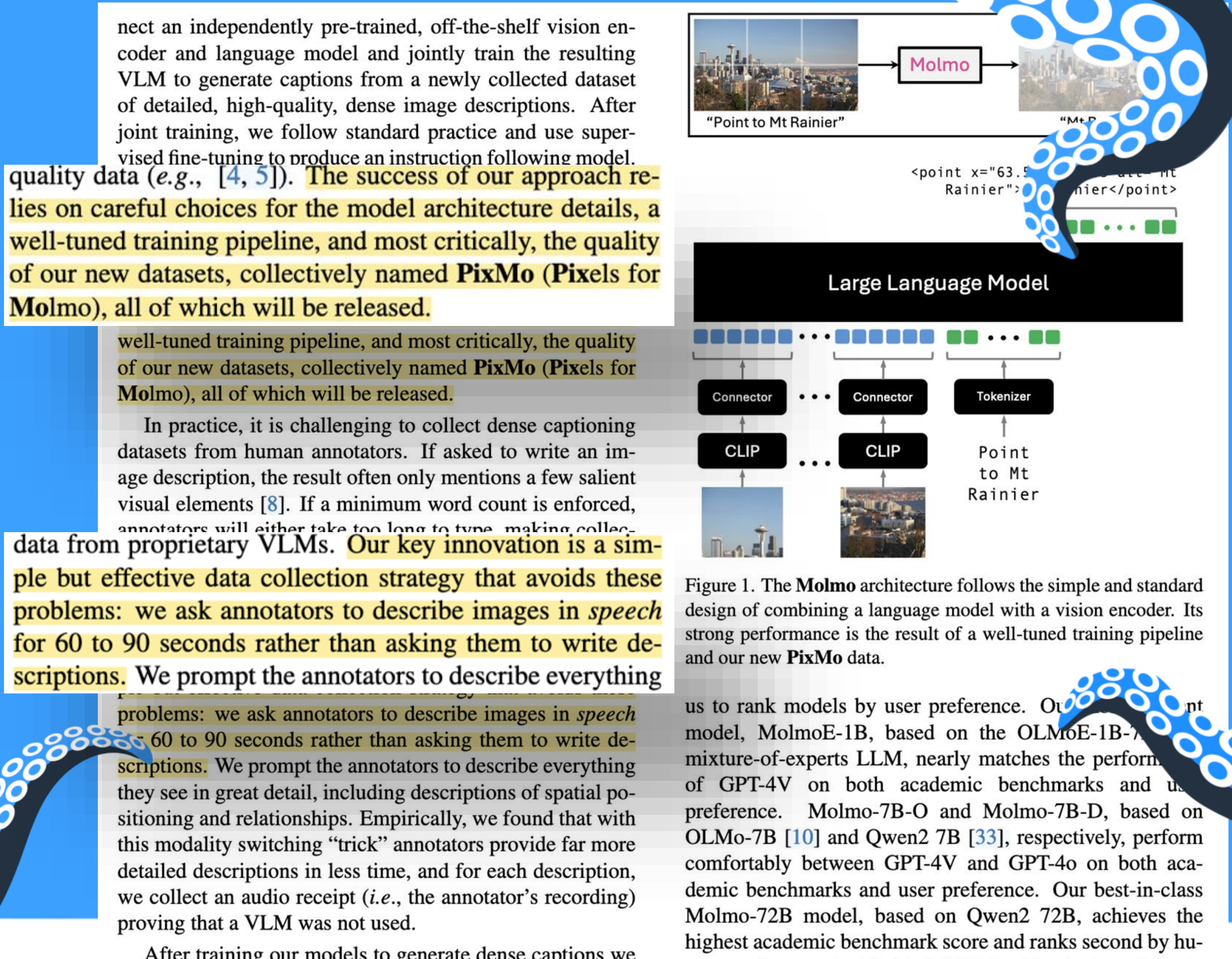
class 72B model within the Molmo family not only outper-

1. Introduction

Extensions to large language models (LLMs) that process images in addition to text have resulted in impressive multimodal capabilities, such as generating comprehensive image descriptions and accurately answering complex visual [illegible] ost performant of these vision-language [illegible] owever, remain proprietary with nei[illegible] ta, nor code being publicly release[illegible]

With the goal of fostering scientific exploration, numerous research efforts have attempted to reproduce similar capabilities in *open* models. Early works, exemplified by LLaVA [15], produced fully open weights and training data but now lag significantly behind the state-of-the-art. More recent, stronger open-weight models have trended towards less open data: the training data may either be proprietary (*e.g.*, [5]) or, in cases where it is released, there is a heavy reliance on *synthetic* data generated by proprietary systems, *e.g.*, models are trained on datasets like ShareGPT4V [7] which uses GPT-4V [25] to generate a large set of detailed

Deitke et al., “Molmo and PixMo.”



nect an independently pre-trained, off-the-shelf vision encoder and language model and jointly train the resulting VLM to generate captions from a newly collected dataset of detailed, high-quality, dense image descriptions. After joint training, we follow standard practice and use supervised fine-tuning to produce an instruction following model.

quality data (*e.g.*, [4, 5]). The success of our approach relies on careful choices for the model architecture details, a well-tuned training pipeline, and most critically, the quality of our new datasets, collectively named **PixMo** (**Pix**els for **Mo**lmo), all of which will be released.

well-tuned training pipeline, and most critically, the quality of our new datasets, collectively named **PixMo** (**Pix**els for **Mo**lmo), all of which will be released.

In practice, it is challenging to collect dense captioning datasets from human annotators. If asked to write an image description, the result often only mentions a few salient visual elements [8]. If a minimum word count is enforced, annotators will either take too long to type, making collec-

data from proprietary VLMs. Our key innovation is a simple but effective data collection strategy that avoids these problems: we ask annotators to describe images in *speech* for 60 to 90 seconds rather than asking them to write descriptions. We prompt the annotators to describe everything

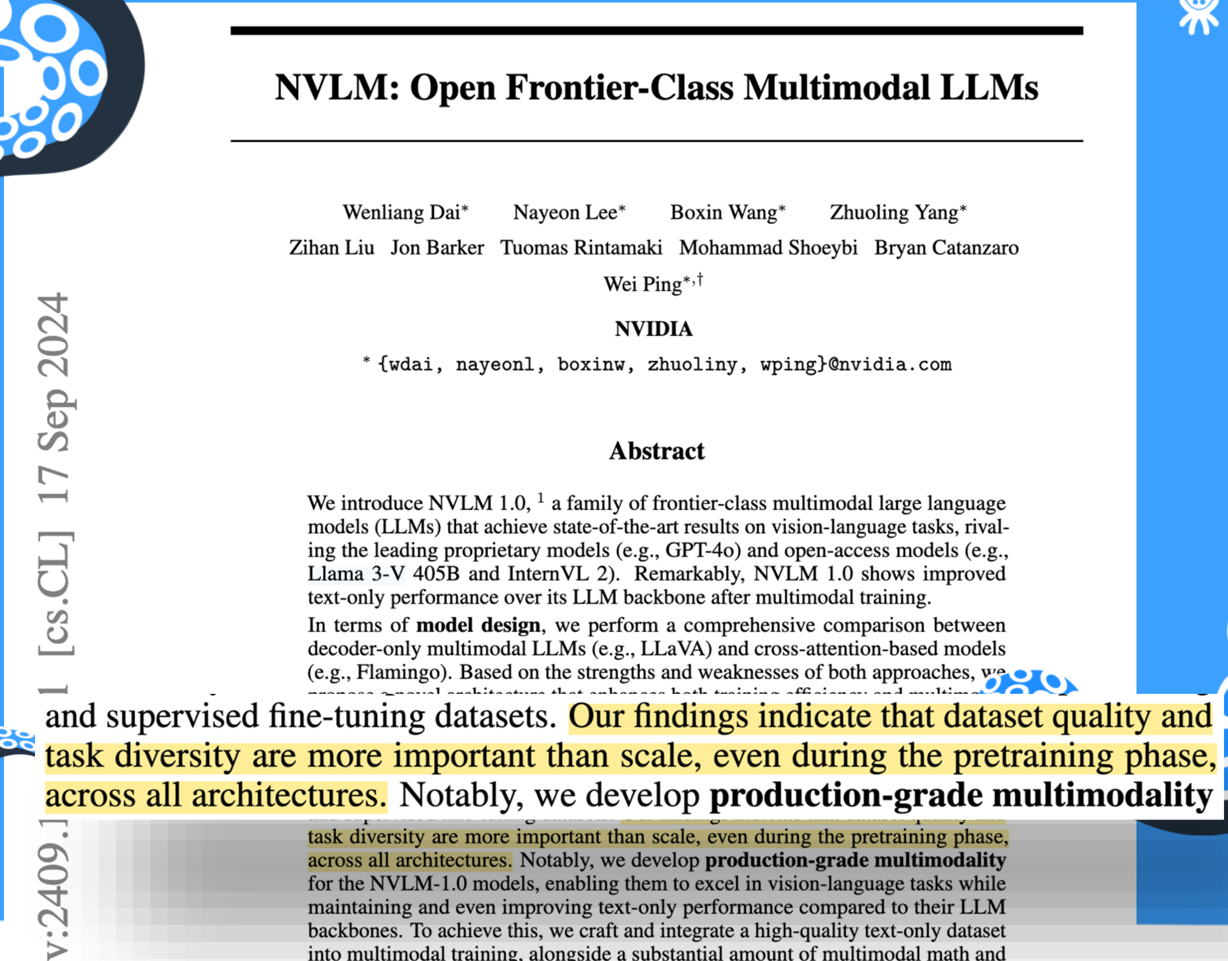
problems: we ask annotators to describe images in *speech* 60 to 90 seconds rather than asking them to write descriptions. We prompt the annotators to describe everything

they see in great detail, including descriptions of spatial positioning and relationships. Empirically, we found that with this modality switching “trick” annotators provide far more detailed descriptions in less time, and for each description, we collect an audio receipt (*i.e.*, the annotator’s recording) proving that a VLM was not used.

After training our models to generate dense captions we

Figure 1. The **Molmo** architecture follows the simple and standard design of combining a language model with a vision encoder. Its strong performance is the result of a well-tuned training pipeline and our new **PixMo** data.

us to rank models by user preference. Ou nt model, MolmoE-1B, based on the OLMoE-1B-7 mixture-of-experts LLM, nearly matches the perform of GPT-4V on both academic benchmarks and u preference. Molmo-7B-O and Molmo-7B-D, based on OLMo-7B [10] and Qwen2 7B [33], respectively, perform comfortably between GPT-4V and GPT-4o on both academic benchmarks and user preference. Our best-in-class Molmo-72B model, based on Qwen2 72B, achieves the highest academic benchmark score and ranks second by hu-



NVLM: Open Frontier-Class Multimodal LLMs

Wenliang Dai* Nayeon Lee* Boxin Wang* Zhuoling Yang*
Zihan Liu Jon Barker Tuomas Rintamaki Mohammad Shoeybi Bryan Catanzaro
Wei Ping*,†

NVIDIA

* `{wdai, nayeonl, boxinw, zhuoliny, wping}@nvidia.com`

v:2409.1
1 [cs.CL] 17 Sep 2024

Abstract

We introduce NVLM 1.0, [1] a family of frontier-class multimodal large language models (LLMs) that achieve state-of-the-art results on vision-language tasks, rivaling the leading proprietary models (e.g., GPT-4o) and open-access models (e.g., Llama 3-V 405B and InternVL 2). Remarkably, NVLM 1.0 shows improved text-only performance over its LLM backbone after multimodal training.

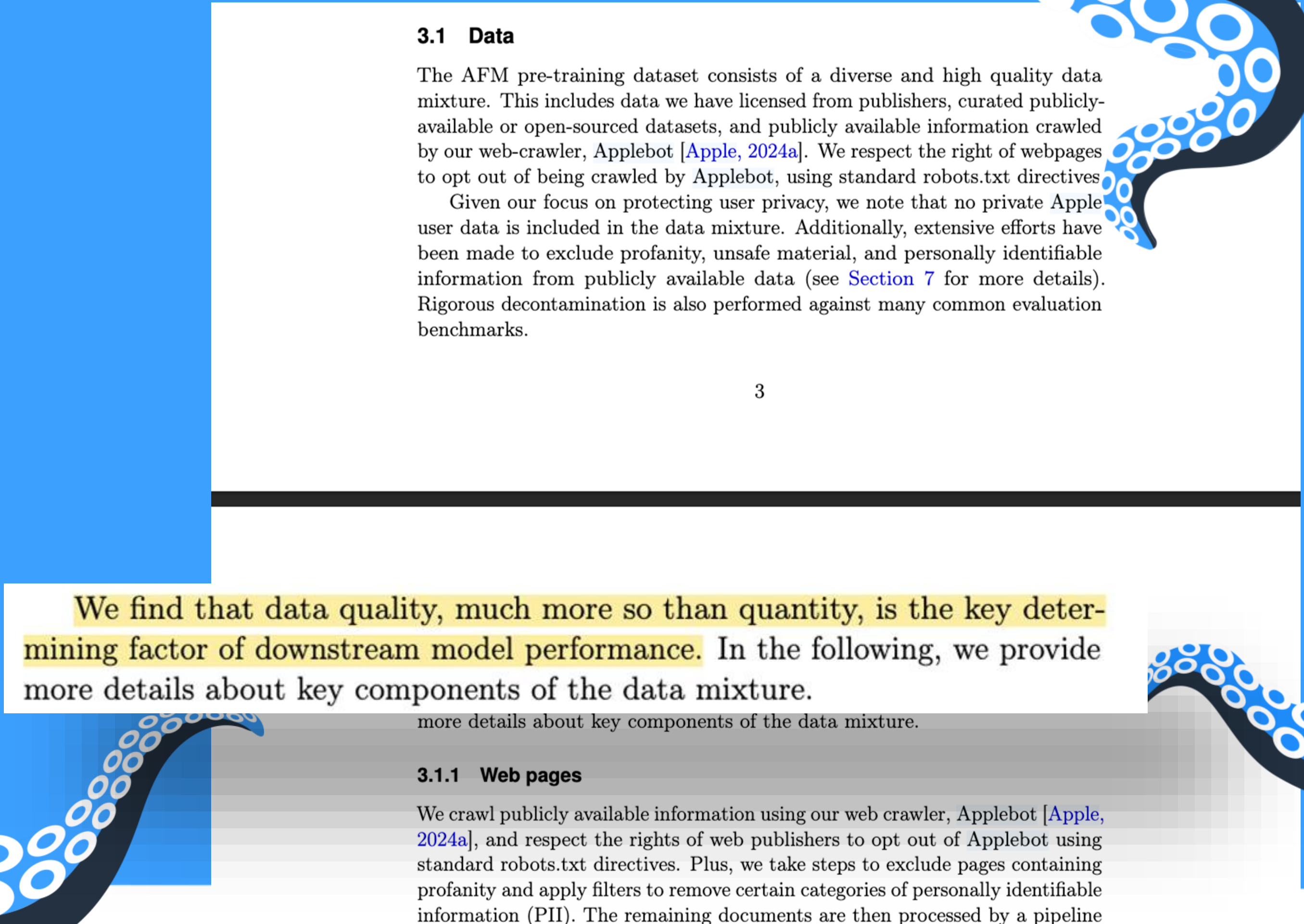
In terms of **model design**, we perform a comprehensive comparison between decoder-only multimodal LLMs (e.g., LLaVA) and cross-attention-based models (e.g., Flamingo). Based on the strengths and weaknesses of both approaches, we propose a novel architecture that enhances both training efficiency and multi

and supervised fine-tuning datasets. Our findings indicate that dataset quality and task diversity are more important than scale, even during the pretraining phase, across all architectures. Notably, we develop **production-grade multimodality**

task diversity are more important than scale, even during the pretraining phase, across all architectures. Notably, we develop **production-grade multimodality**

for the NVLM-1.0 models, enabling them to excel in vision-language tasks while maintaining and even improving text-only performance compared to their LLM backbones. To achieve this, we craft and integrate a high-quality text-only dataset into multimodal training, alongside a substantial amount of multimodal math and

Source: arXiv:2407.21075v1 [cs.AI] 29 Jul 2024



3.1 Data

The AFM pre-training dataset consists of a diverse and high quality data mixture. This includes data we have licensed from publishers, curated publicly-available or open-sourced datasets, and publicly available information crawled by our web-crawler, Applebot [Apple, 2024a]. We respect the right of webpages to opt out of being crawled by Applebot, using standard robots.txt directives.

Given our focus on protecting user privacy, we note that no private Apple user data is included in the data mixture. Additionally, extensive efforts have been made to exclude profanity, unsafe material, and personally identifiable information from publicly available data (see Section 7 for more details). Rigorous decontamination is also performed against many common evaluation benchmarks.

We find that data quality, much more so than quantity, is the key determining factor of downstream model performance. In the following, we provide more details about key components of the data mixture.

more details about key components of the data mixture.

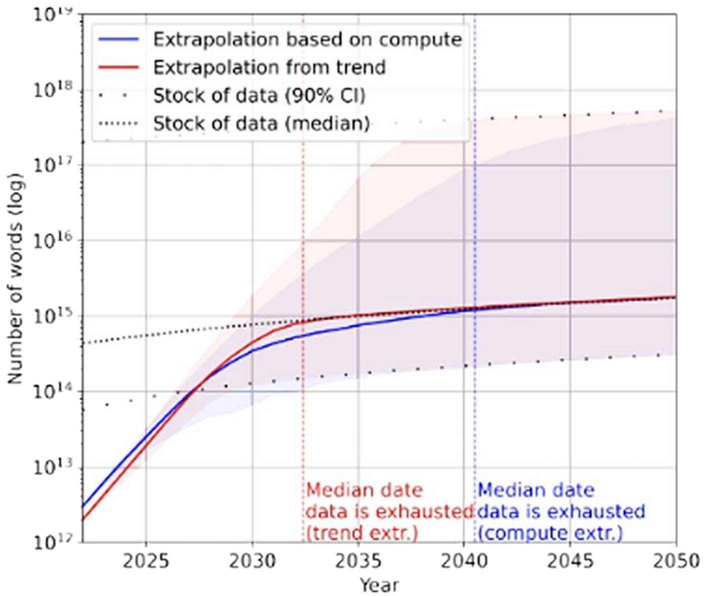
3.1.1 Web pages

We crawl publicly available information using our web crawler, Applebot [Apple, 2024a], and respect the rights of web publishers to opt out of Applebot using standard robots.txt directives. Plus, we take steps to exclude pages containing profanity and apply filters to remove certain categories of personally identifiable information (PII). The remaining documents are then processed by a pipeline

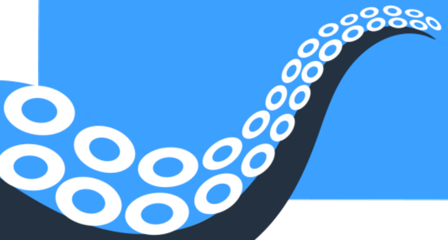
Keboola

Data is Central to AI Training Foundation Models

We will run out of data? An analysis of the limits of scaling datasets in Machine Learning



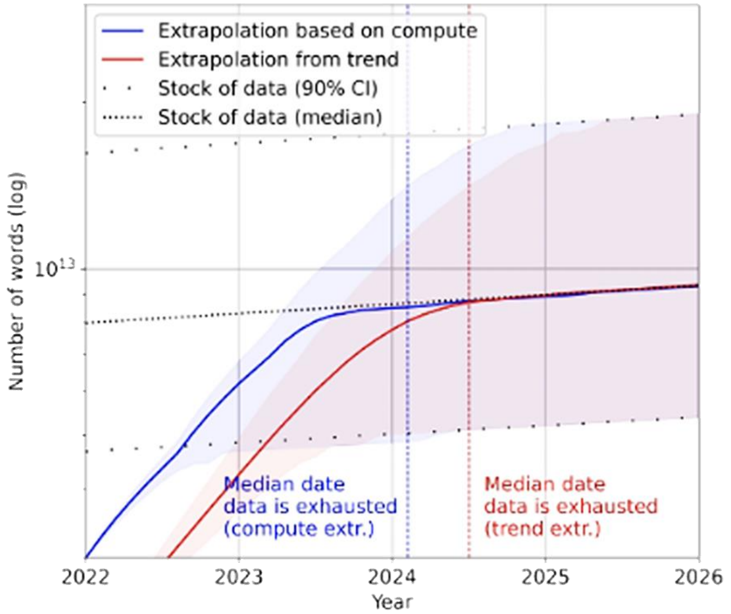
Projections of Low-Quality Data



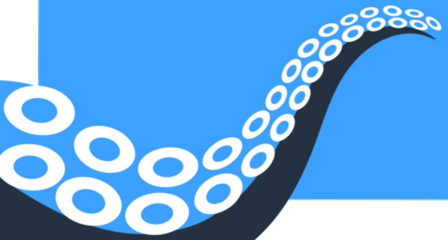
Keboola

Data is Central to AI Training Foundation Models

Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning



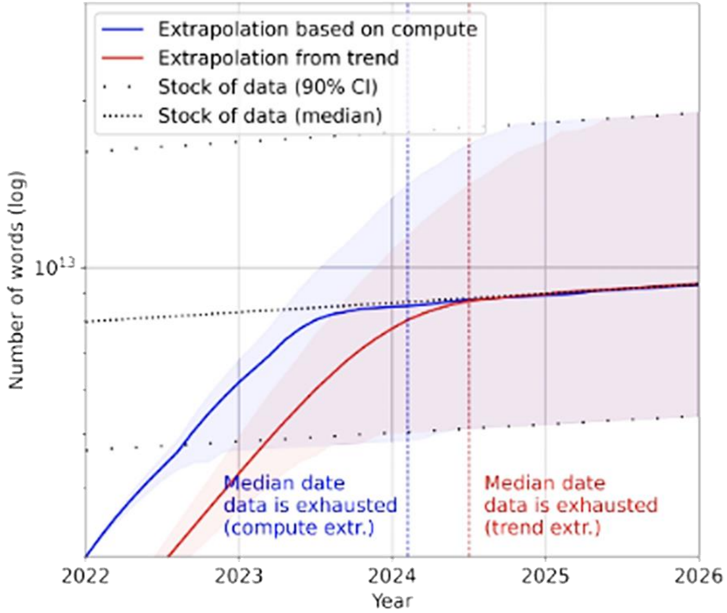
Projections of High-Quality Data



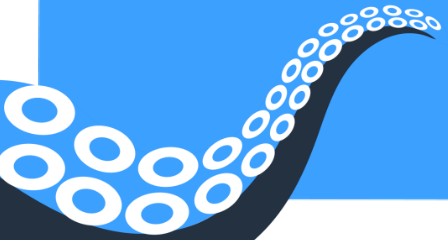
Keboola

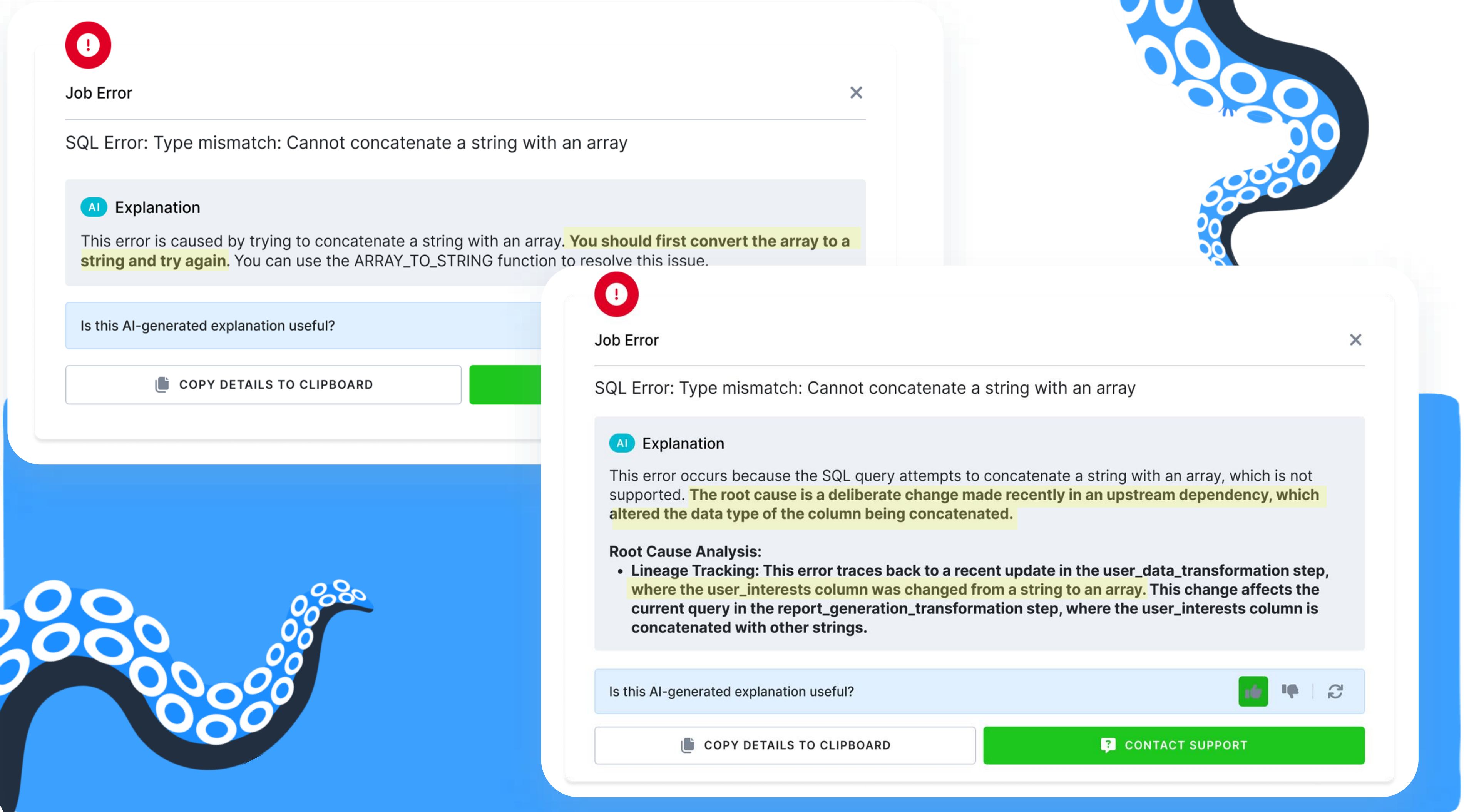
What are you doing to lengthen this curve in *your* domain?

Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning



Projections of High-Quality Data





Job Error

SQL Error: Type mismatch: Cannot concatenate a string with an array

AI Explanation

This error is caused by trying to concatenate a string with an array. You should first convert the array to a string and try again. You can use the ARRAY_TO_STRING function to resolve this issue.

Is this AI-generated explanation useful?

Job Error

SQL Error: Type mismatch: Cannot concatenate a string with an array

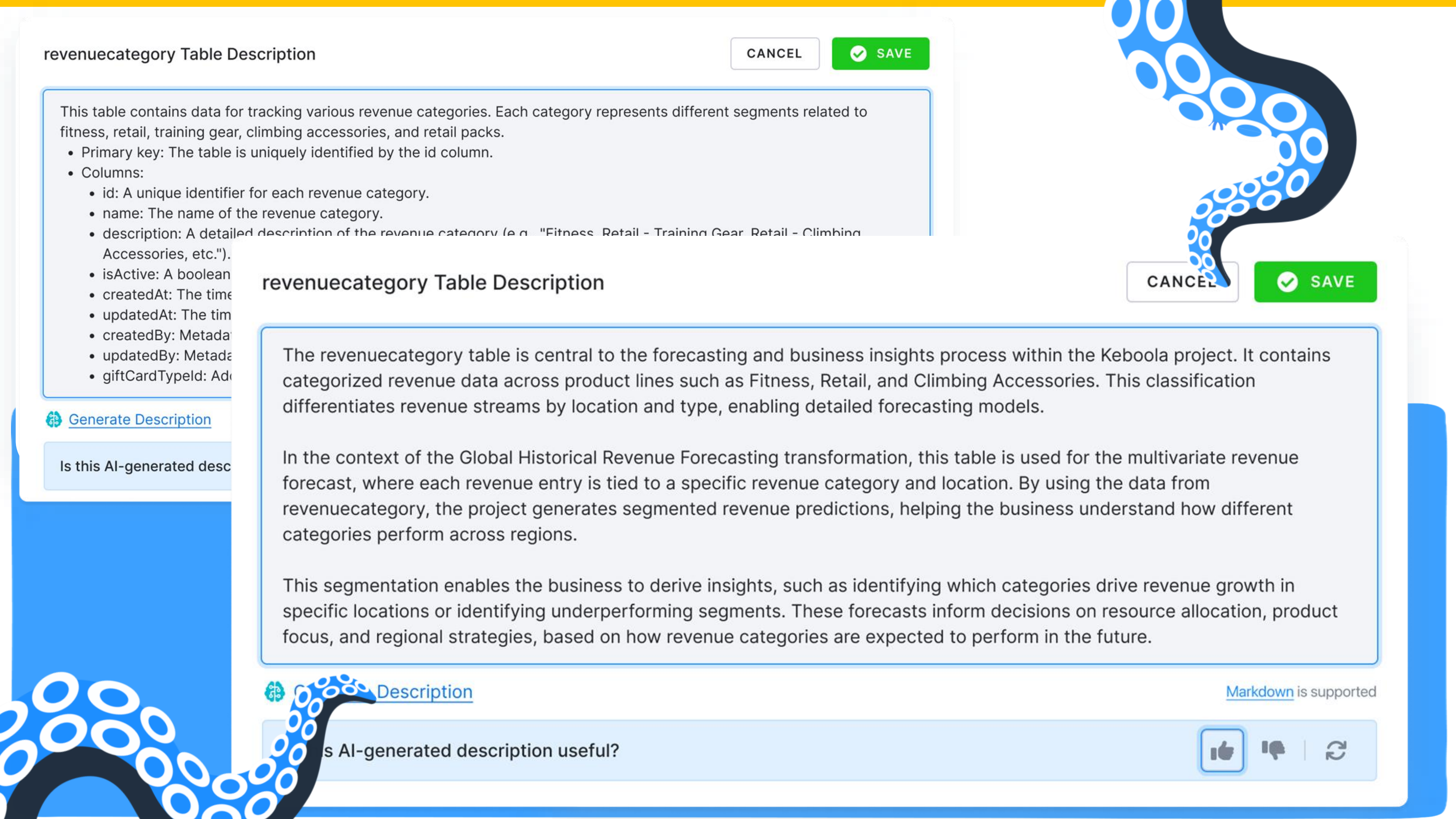
AI Explanation

This error occurs because the SQL query attempts to concatenate a string with an array, which is not supported. The root cause is a deliberate change made recently in an upstream dependency, which altered the data type of the column being concatenated.

Root Cause Analysis:

- **Lineage Tracking: This error traces back to a recent update in the user_data_transformation step, where the user_interests column was changed from a string to an array. This change affects the current query in the report_generation_transformation step, where the user_interests column is concatenated with other strings.**

Is this AI-generated explanation useful?



revenuecategory Table Description

This table contains data for tracking various revenue categories. Each category represents different segments related to fitness, retail, training gear, climbing accessories, and retail packs.

- Primary key: The table is uniquely identified by the id column.
- Columns:
  - id: A unique identifier for each revenue category.
  - name: The name of the revenue category.
  - description: A detailed description of the revenue category (e.g., "Fitness, Retail - Training Gear, Retail - Climbing Accessories, etc.").
  - isActive: A boolean
  - createdAt: The time
  - updatedAt: The tim
  - createdBy: Metada
  - updatedBy: Metada
  - giftCardTypeId: Ad

Is this AI-generated desc

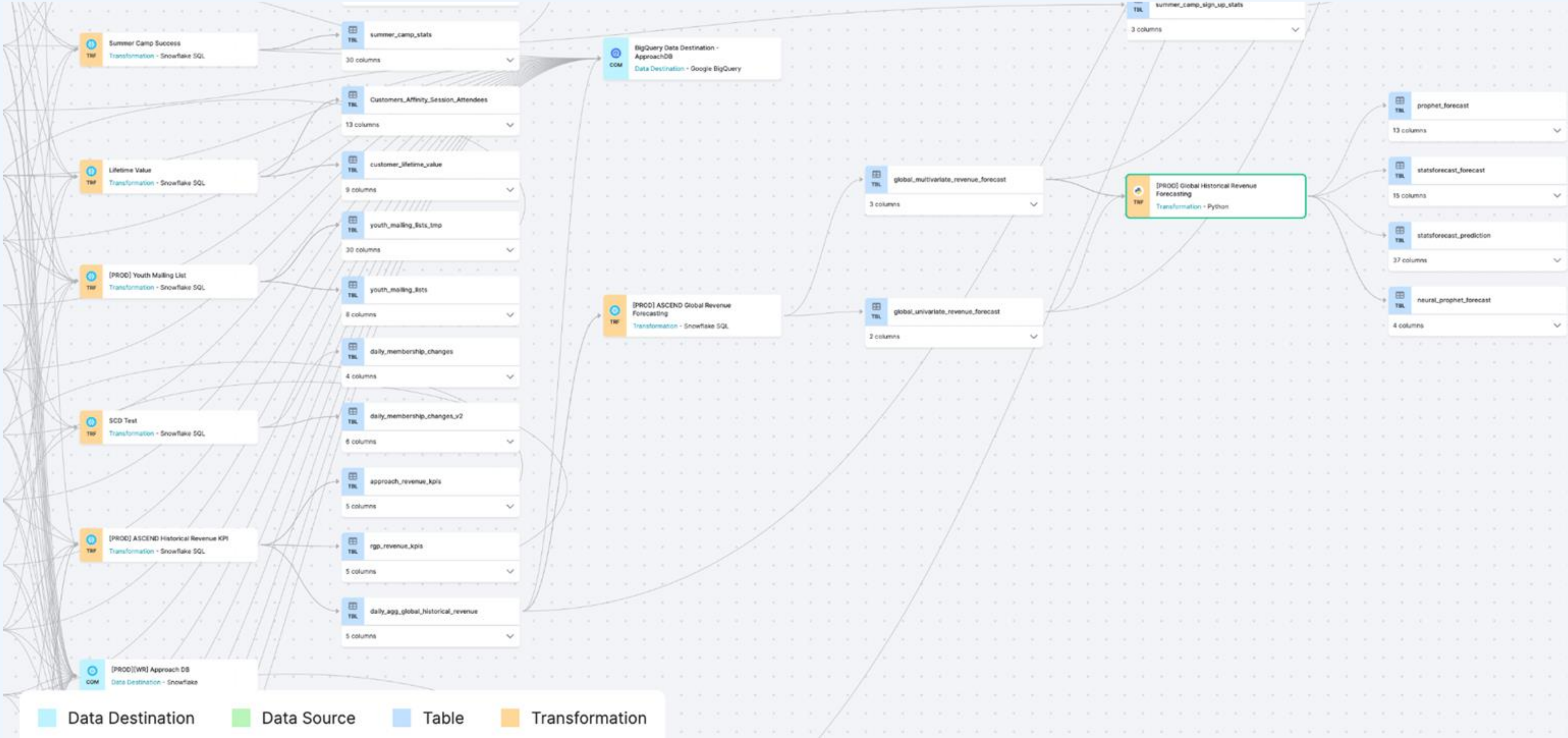
revenuecategory Table Description

The revenuecategory table is central to the forecasting and business insights process within the Keboola project. It contains categorized revenue data across product lines such as Fitness, Retail, and Climbing Accessories. This classification differentiates revenue streams by location and type, enabling detailed forecasting models.

In the context of the Global Historical Revenue Forecasting transformation, this table is used for the multivariate revenue forecast, where each revenue entry is tied to a specific revenue category and location. By using the data from revenuecategory, the project generates segmented revenue predictions, helping the business understand how different categories perform across regions.

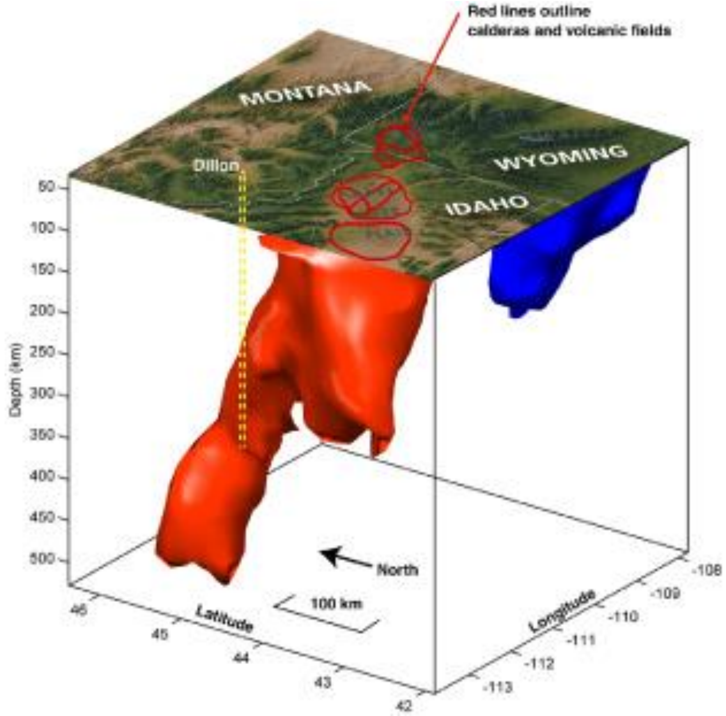
This segmentation enables the business to derive insights, such as identifying which categories drive revenue growth in specific locations or identifying underperforming segments. These forecasts inform decisions on resource allocation, product focus, and regional strategies, based on how revenue categories are expected to perform in the future.

s AI-generated description useful?


TRF
Summer Camp Success
Transformation - Snowflake SQL
TBL
summer_camp_stats
30 columns
COM
BigQuery Data Destination - ApproachDB
Data Destination - Google BigQuery
TBL
summer_camp_sign_up_stats
3 columns
TBL
Customers_Affinity_Session_Attendees
13 columns
TBL
prophet_forecast
13 columns
TRF
Lifetime Value
Transformation - Snowflake SQL
TBL
customer_lifetime_value
9 columns
TBL
global_multivariate_revenue_forecast
3 columns
TRF
[PROD] Global Historical Revenue Forecasting
Transformation - Python
TBL
statsforecast_forecast
15 columns
TBL
youth_mailing_lists_tmp
30 columns
TBL
statsforecast_prediction
37 columns
TRF
[PROD] Youth Mailing List
Transformation - Snowflake SQL
TBL
youth_mailing_lists
8 columns
TRF
[PROD] ASCEND Global Revenue Forecasting
Transformation - Snowflake SQL
TBL
global_univariate_revenue_forecast
2 columns
TBL
neural_prophet_forecast
4 columns
TBL
daily_membership_changes
4 columns
TRF
SCD Test
Transformation - Snowflake SQL
TBL
daily_membership_changes_v2
6 columns
TBL
approach_revenue_kpis
5 columns
TRF
[PROD] ASCEND Historical Revenue KPI
Transformation - Snowflake SQL
TBL
rgp_revenue_kpis
5 columns
TBL
daily_agg_global_historical_revenue
5 columns
COM
[PROD][WR] Approach DB
Data Destination - Snowflake

🟦 Data Destination 🟩 Data Source 🟦 Table 🟧 Transformation

Knowing the trajectory of AI doesn’t help to predict the landscape that it will shape



Building foundations for AI is a delicate balance

We need to demand more from our tools and vendors

- What is the data for? Does it represent what we want?

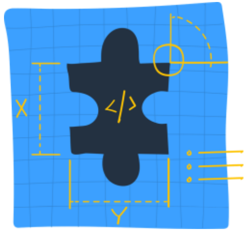
We need to demand more from our tools and vendors

- What is the data for? Does it represent what we want?
- What data is actually being used? How is it being processed?

We need to demand more from our tools and vendors

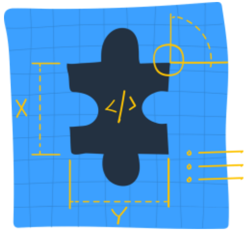
- What is the data for? Does it represent what we want?
- What data is actually being used? How is it being processed?
- How do we validate that the data and models do the things that we want it to do?

Lessons from the Trenches



Be Pragmatic

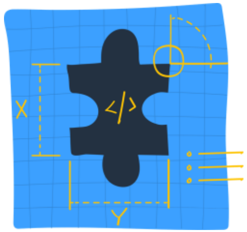
Lessons from the Trenches



Be Pragmatic

Forge Partnerships

Lessons from the Trenches



Be Pragmatic

Forge Partnerships

Curate your Knowledge