

Demystifying Artificial Intelligence for Data Preparation

Chengliang Chai (Beijing Institute of Technology, China)

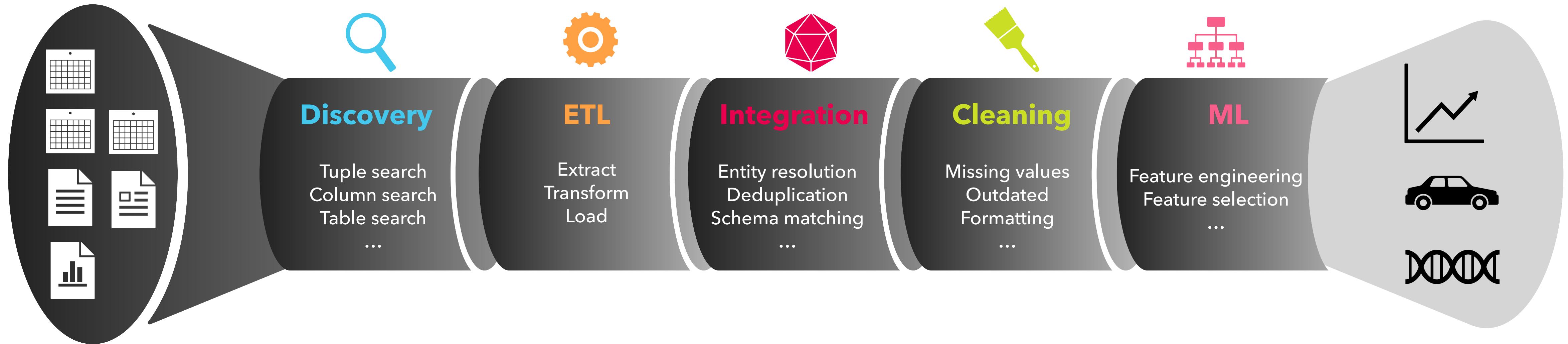
Nan Tang (QCRI, Qatar)

Ju Fan (Renmin University of China)

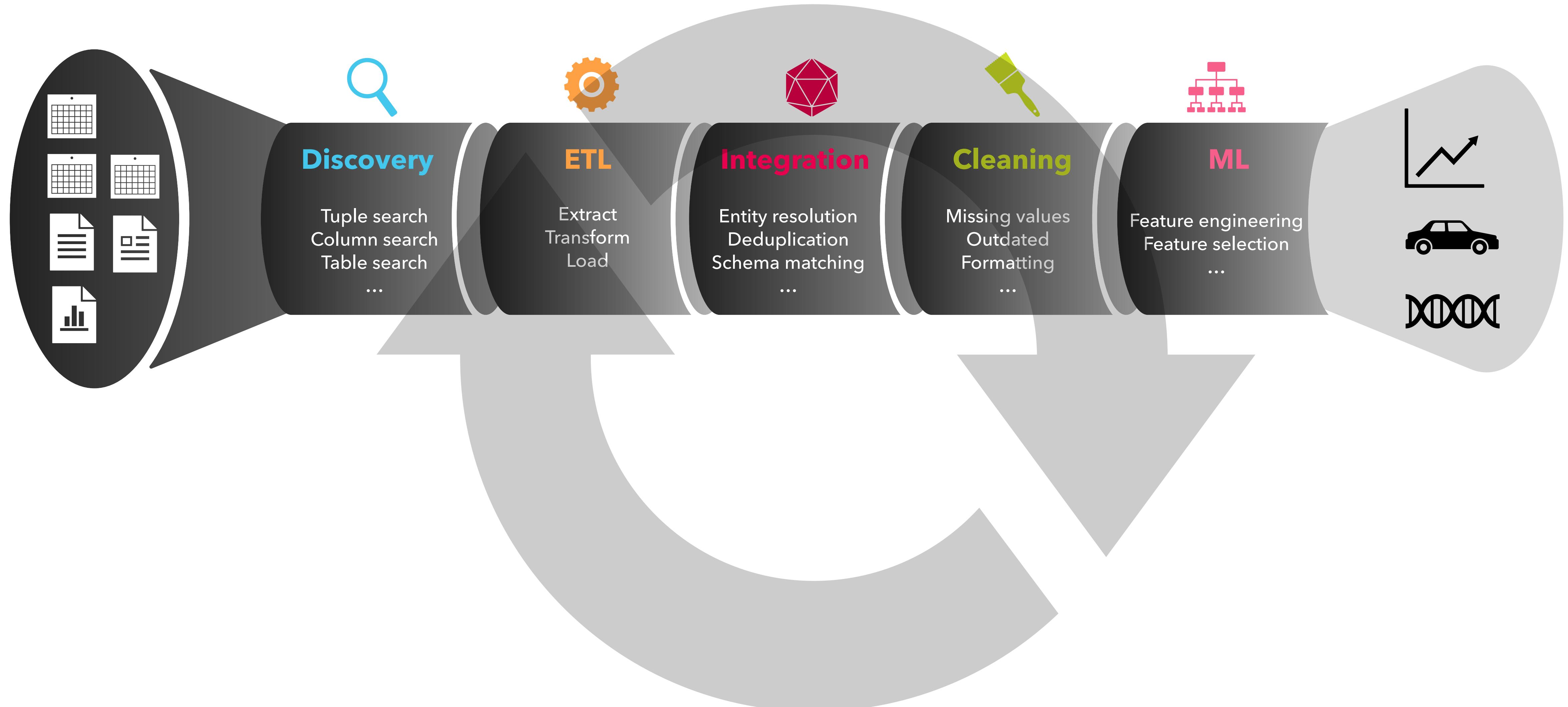
Yuyu Luo (Tsinghua University, China)

2023.06.18 @ SIGMOD 2023, Seattle, WA, USA

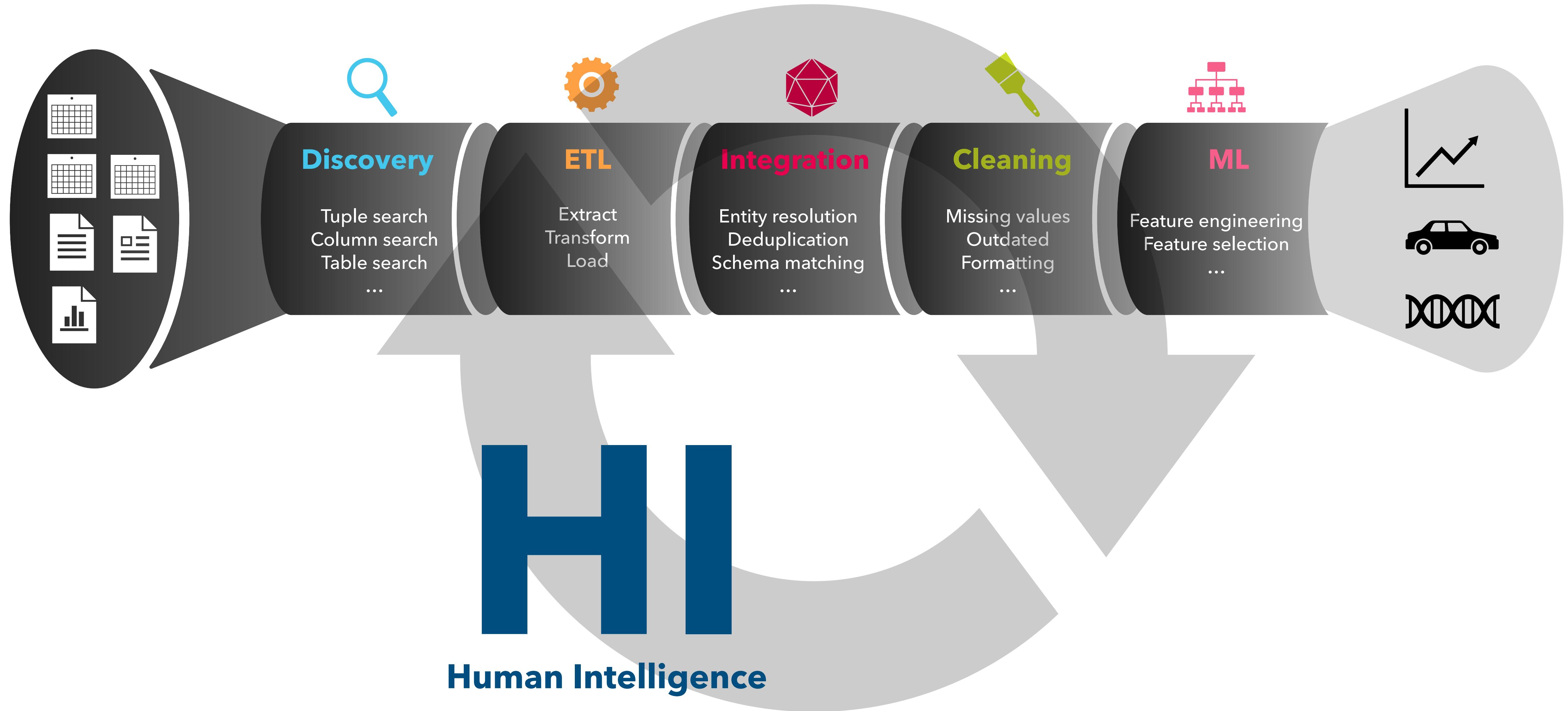
Artificial Intelligence for Data Preparation



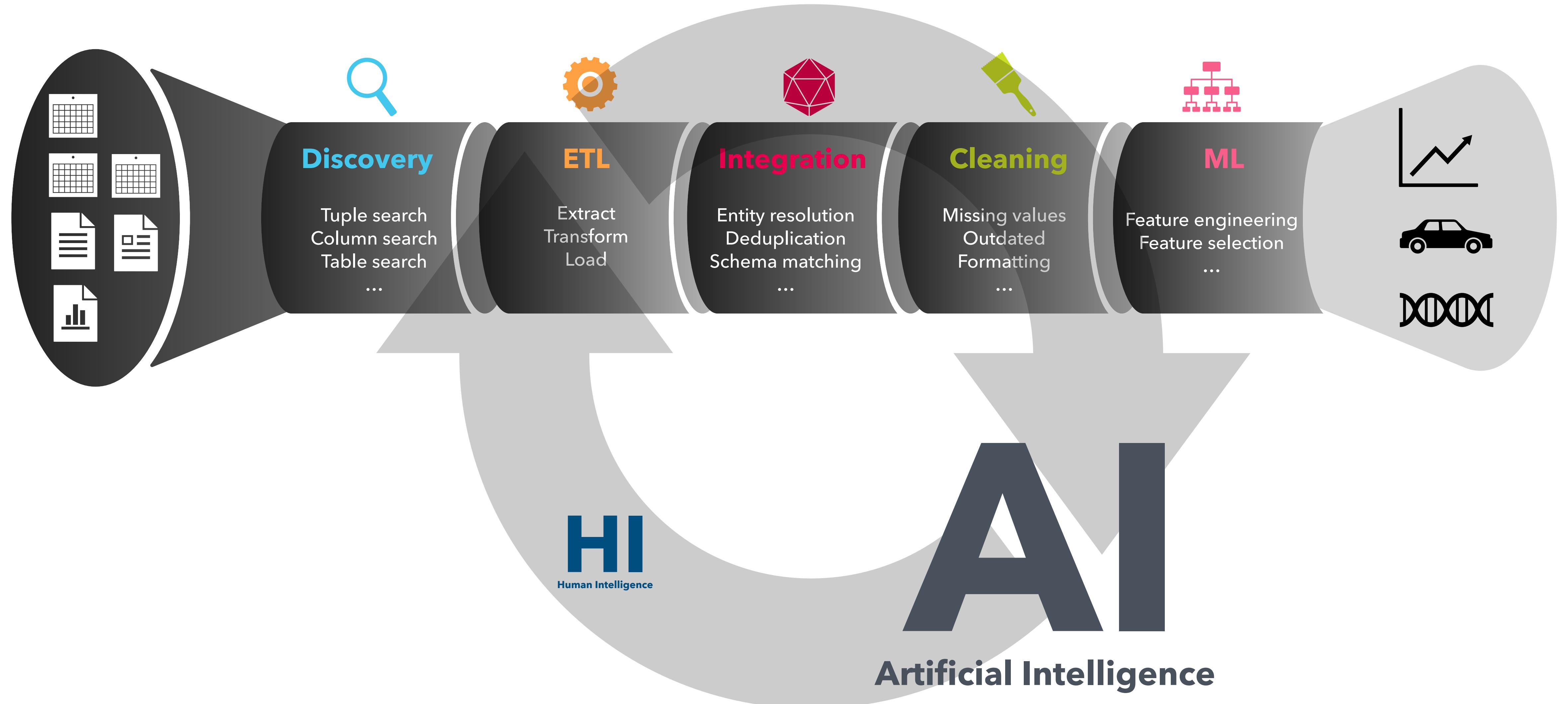
Artificial Intelligence for Data Preparation



Artificial Intelligence for Data Preparation



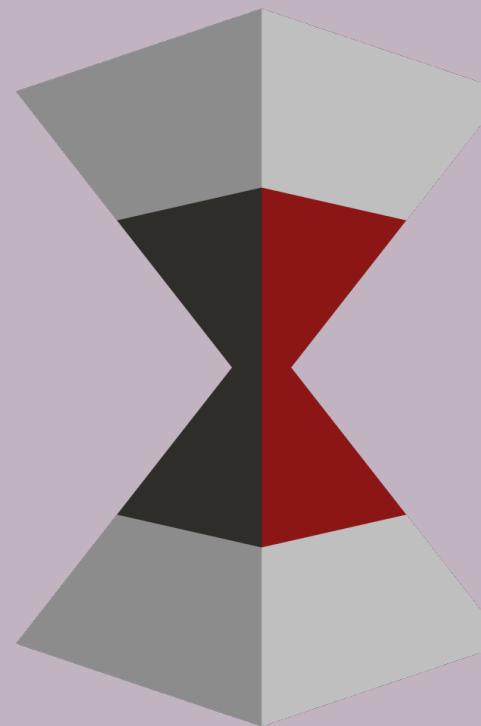
Artificial Intelligence for Data Preparation



Artificial Intelligence for Data Preparation



Stanford University
Human-Centered
Artificial Intelligence



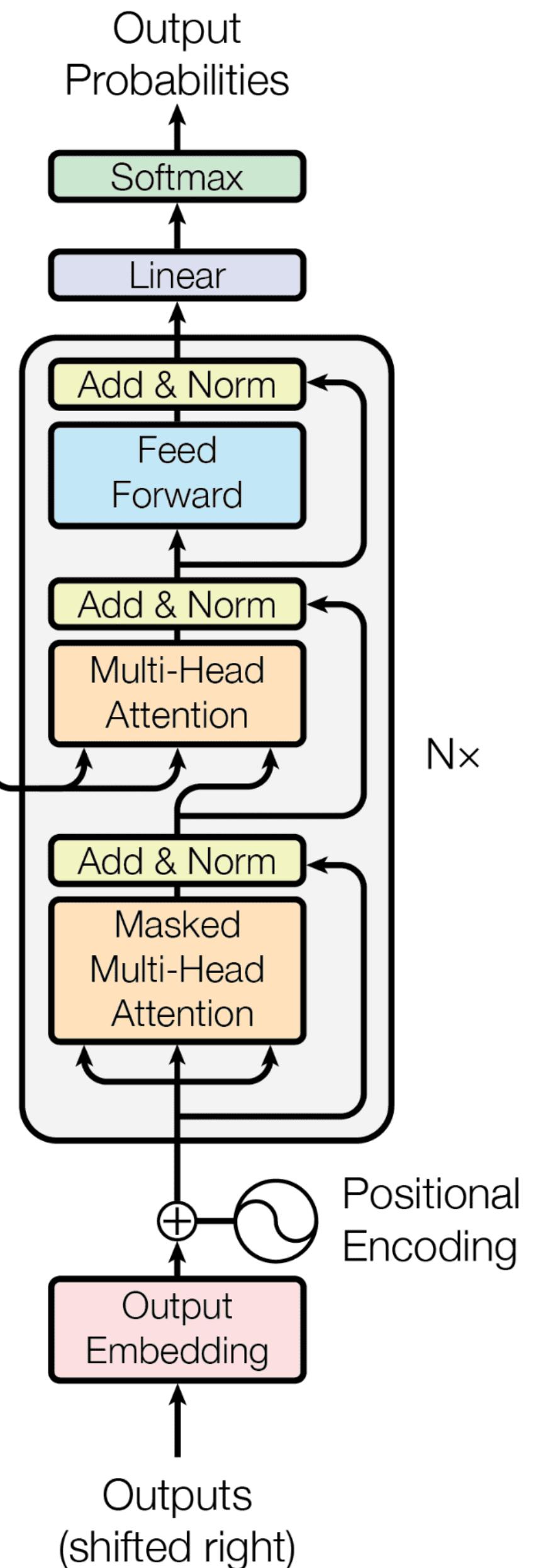
**Center for
Research on
Foundation
Models**

On the Opportunities and Risks of Foundation Models

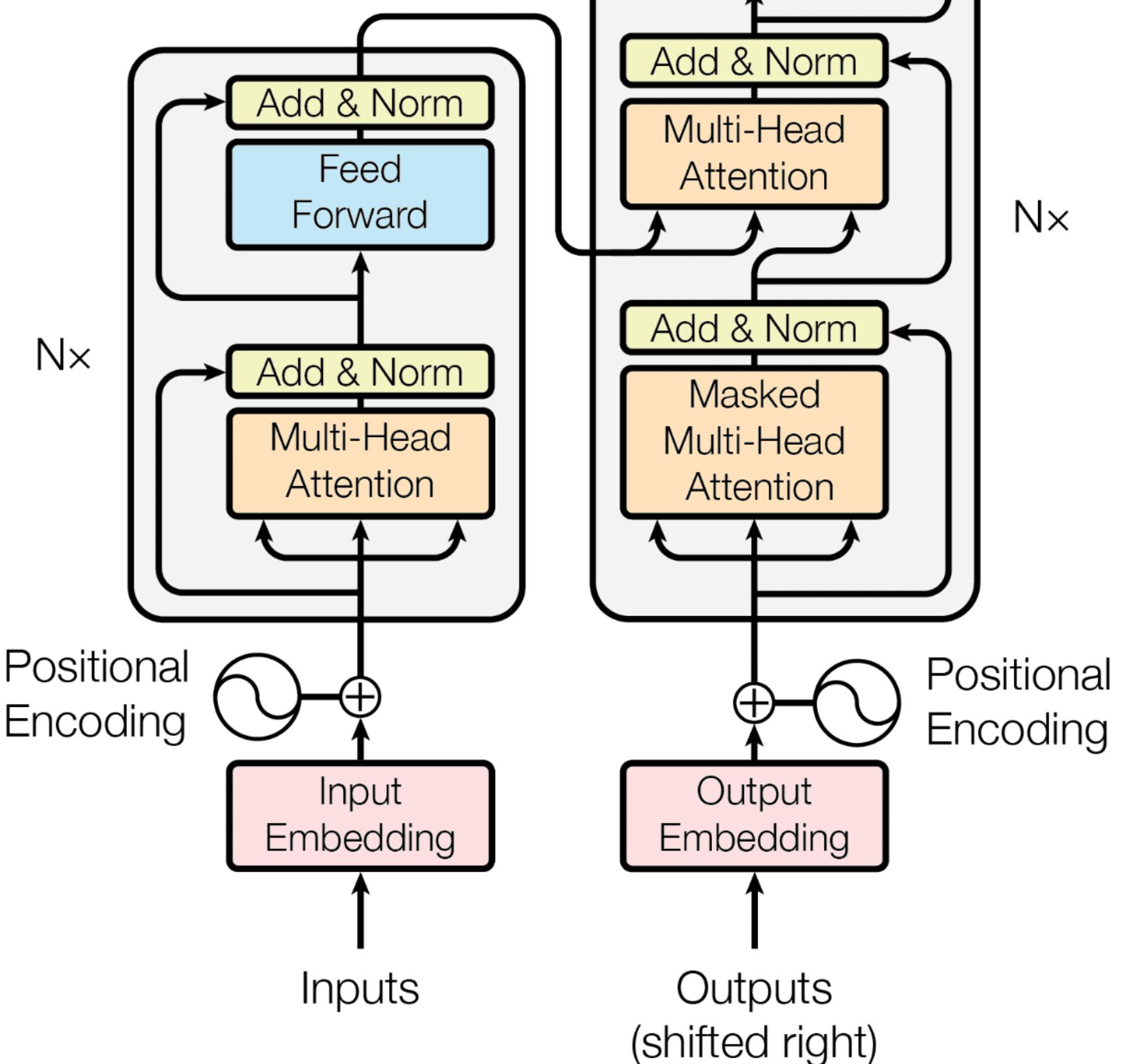
Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*¹

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

Decoder



Encoder



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

Abstract

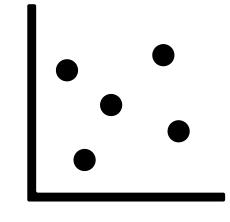
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, 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.

Using Foundation Models



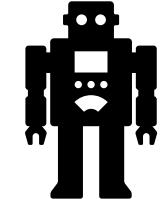
Fine Tuning (adaptation)

- Load foundation model
- Small train data
- State-of-the-art results



Embeddings (representations)

- Encode object into vector
- Similarity search



In-Context Learning

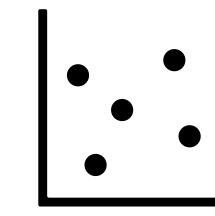
- Simply ask (zero-shot)
- Provide examples (few-shot)

Using Foundation Models



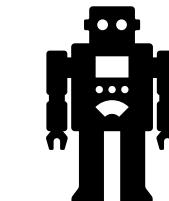
Fine Tuning
(adaptation)

- Load foundation model
- Small train data
- State-of-the-art results



Embeddings
(representations)

- Encode object into vector
- Similarity search



In-Context
Learning

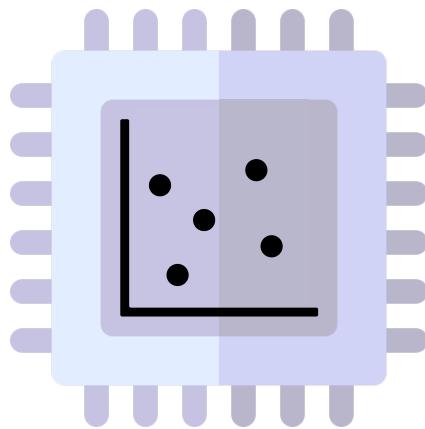
- Simply ask (zero-shot)
- Provide examples (few-shot)

Using Foundation Models



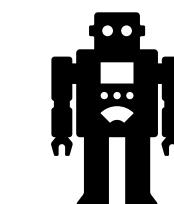
Fine Tuning (adaptation)

- Load foundation model
- Small train data
- State-of-the-art results



Embeddings (representations)

- Encode object into vector
- Similarity search



In-Context Learning

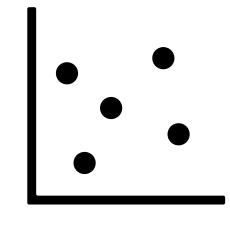
- Simply ask (zero-shot)
- Provide examples (few-shot)

Using Foundation Models



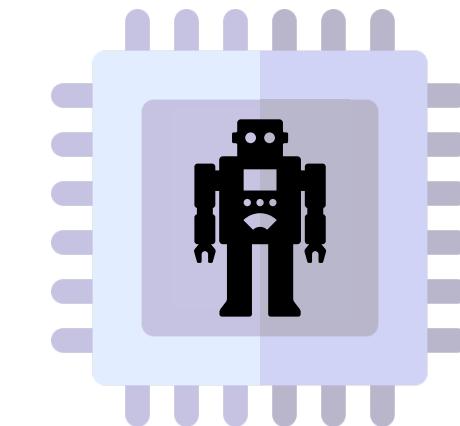
Fine Tuning
(adaptation)

- Load foundation model
- Small train data
- State-of-the-art results



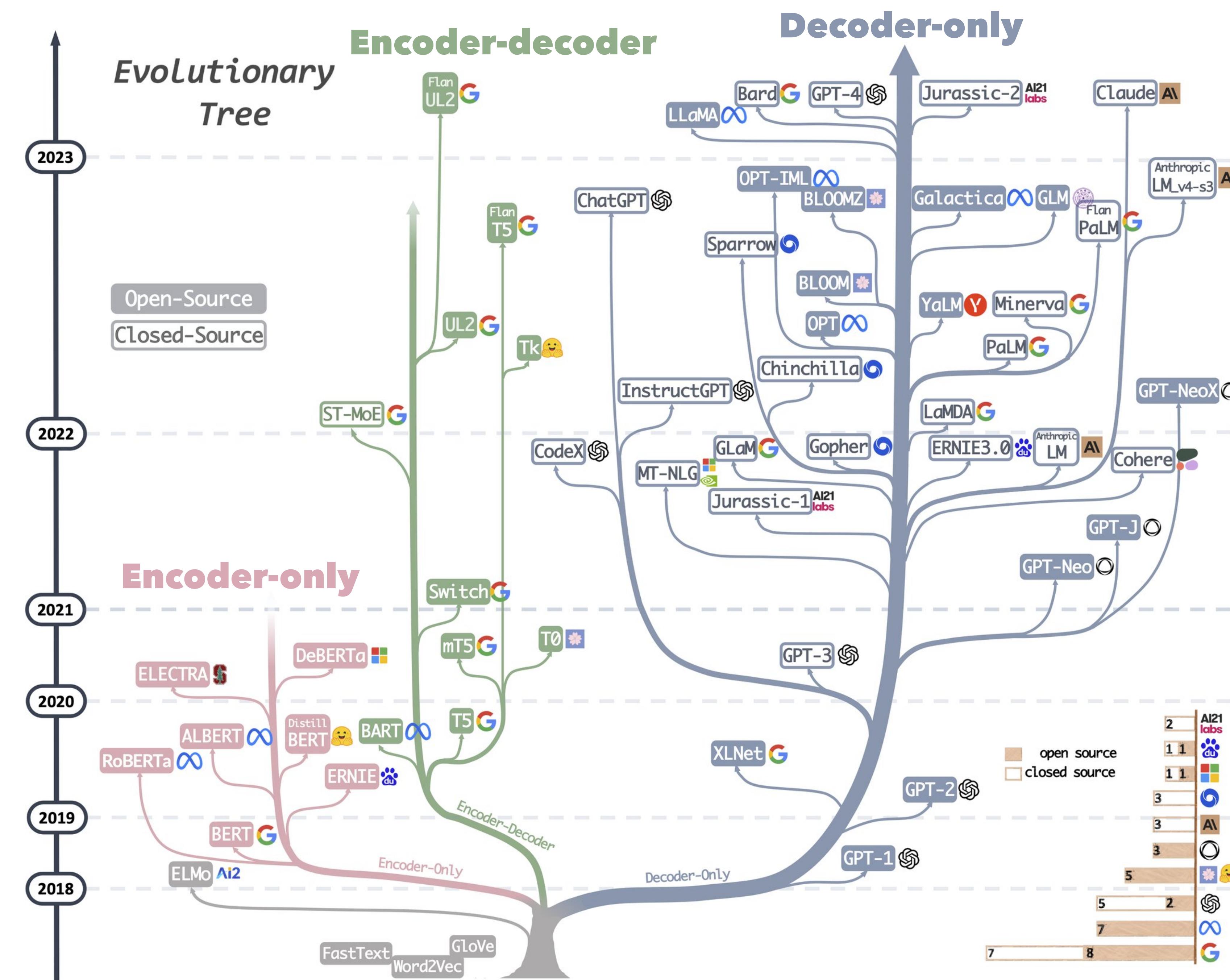
Embeddings
(representations)

- Encode object into vector
- Similarity search



In-Context
Learning

- Simply ask (zero-shot)
- Provide examples (few-shot)



A survey of LLMs with a practical guide and evolutionary tree

Tutorial Outline

PART I

Foundation models: In-context learning for single tasks

PART II

Fine-tuning language models for single tasks

PART III

AI-assisted data preparation pipeline orchestration
with meta-learning, genetic programming, reinforcement learning

PART I: Foundation Models: In-context Learning for Data Preparation Single Tasks

- **Early successes**
- Limitations
- Research Opportunities

Can Foundation Models Wrangle Your Data?

Avanika Narayan, Ines Chami†, Laurel Orr, Simran Arora, Christopher Ré
Stanford University and †Numbers Station
{avanika, lorr1, chrismre, simarora}@cs.stanford.edu, ines.chami@numbersstation.ai

Can Foundation Models Wrangle Your Data?

Avanika Narayan, Ines Chami†, Laurel Orr, Simran Arora, Christopher Ré
Stanford University and †Numbers Station
{avanika, lorr1, chrismre, simarora}@cs.stanford.edu, ines.chami@numbersstation.ai

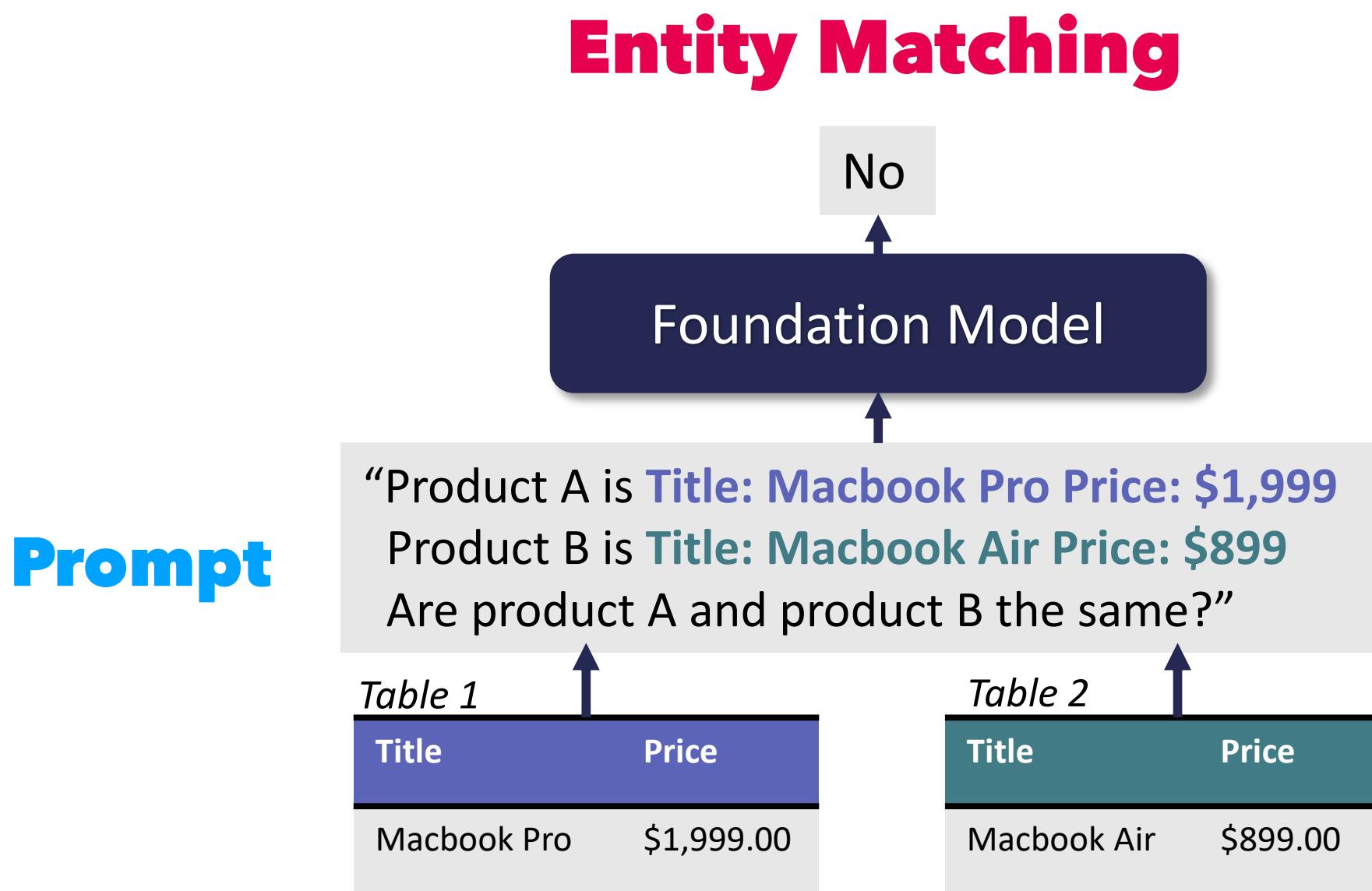


Figure 1: A large FM can address an entity matching task using prompting. Rows are serialized into text and passed to the FM with the question “Are products A and B the same?”. The FM then generates a string “Yes” or “No” as the answer.

Can Foundation Models Wrangle Your Data?

Avanika Narayan, Ines Chami†, Laurel Orr, Simran Arora, Christopher Ré
Stanford University and †Numbers Station
{avanika, lorr1, chrismre, simarora}@cs.stanford.edu, ines.chami@numbersstation.ai

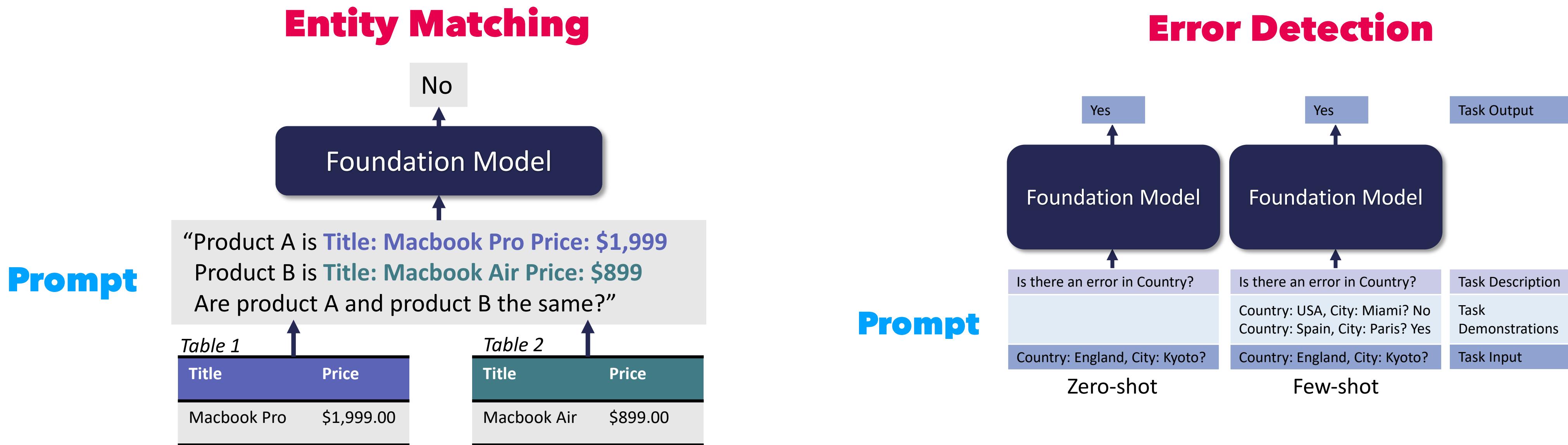


Figure 1: A large FM can address an entity matching task using prompting. Rows are serialized into text and passed to the FM with the question “Are products A and B the same?”. The FM then generates a string “Yes” or “No” as the answer.

Figure 2: Different ways to use FMs with “in-context” learning [20] on an error detection task. For zero-shot (left), the prompt is the task description and the example to complete. For few-shot (right), the prompt adds demonstrations of how to complete the task.

Can Foundation Models Wrangle Your Data?

Avanika Narayan, Ines Chami†, Laurel Orr, Simran Arora, Christopher Ré
 Stanford University and †Numbers Station
 {avanika, lorr1, chrismre, simarora}@cs.stanford.edu, ines.chami@numbersstation.ai

- It might work
- Attributes matter
- Prompt matters
- Few-shot matters



Figure 1: A large FM can address an entity matching task using prompting. Rows are serialized into text and passed to the FM with the question “Are products A and B the same?”. The FM then generates a string “Yes” or “No” as the answer.

Figure 2: Different ways to use FMs with “in-context” learning [20] on an error detection task. For zero-shot (left), the prompt is the task description and the example to complete. For few-shot (right), the prompt adds demonstrations of how to complete the task.

Experimental Results

Dataset	Magellan	Ditto	GPT3-175B ($k=0$)	GPT3-175B ($k=10$)
Fodors-Zagats	100	100	87.2	100
Beer	78.8	94.37	78.6	100
iTunes-Amazon	91.2	97.06	65.9	98.2
Walmart-Amazon	71.9	86.76	60.6	87.0
DBLP-ACM	98.4	98.99	93.5	96.6
DBLP-Google	92.3	95.60	64.6	83.8
Amazon-Google	49.1	75.58	54.3	63.5

Table 1: Entity matching results measured by F1 score where k is the number of task demonstrations.

Code Generation by Examples

Prompt

Generated code

Explanation

NT Can you write a Python program that transforms strings like Nan Tang to N. Tang, and Ju Fan to J. Fan

Certainly! Here's a Python program that transforms strings in the format "First Name Last Name" to the format "Initial. Last Name":

```
python

def transform_name(name):
    parts = name.split() # Split the name into first and last name
    if len(parts) != 2: # Ensure the name has two parts
        return name
    first_name, last_name = parts

    # Extract the initial of the first name
    first_initial = first_name[0] if first_name else ""

    # Construct the transformed name
    transformed_name = f"{first_initial}. {last_name}"
    return transformed_name

# Example usage
name1 = "Nan Tang"
name2 = "Ju Fan"

transformed_name1 = transform_name(name1)
transformed_name2 = transform_name(name2)

print(transformed_name1) # Output: N. Tang
print(transformed_name2) # Output: J. Fan
```

This program defines a function `transform_name()` that takes a name as input, splits it into first and last names, and constructs the transformed name by taking the initial of the first name followed by a dot and the last name. If the input name does not have two parts (i.e., first and last name), it returns the original name as is. Finally, it demonstrates the usage of the function with example names.

PART I: Foundation Models: In-context Learning for Data Preparation Single Tasks

- Early successes: entity matching, error detection, string transformation
- **Limitations**
 - Prompt-sensitive
- Research Opportunities

Prompt Engineering

- Zero-shot
- Prompt matters

Name	City	Profession
Michael Jordan	Berkeley	???

NT name: Michael Jordan, city: Berkeley, profession?

 Michael Jordan is a retired professional basketball player. He is widely regarded as one of the greatest basketball players of all time. There is no information that Michael Jordan has any association with the city of Berkeley, CA.

Prompt with a tuple-like template

NT what is the profession of Michael Jordan living at Berkeley

 As of my training data cut-off of 2021, Michael Jordan does not live in Berkeley. Additionally, Michael Jordan is a retired professional basketball player. He played for the Chicago Bulls in the National Basketball Association (NBA) for most of his career.

Prompt with a natural language template

Prompt Engineering

- Zero-shot
- Prompt matters

Name	City	Profession
Michael Jordan	Berkeley	???

NT name: Michael Jordan, city: Berkeley, profession?

Michael Jordan is a retired professional basketball player. He is widely regarded as one of the greatest basketball players of all time. There is no information that Michael Jordan has any association with the city of Berkeley, CA.

Prompt with a tuple-like template

NT what is the profession of Michael Jordan living at Berkeley

As of my training data cut-off of 2021, Michael Jordan does not live in Berkeley. Additionally, Michael Jordan is a retired professional basketball player. He played for the Chicago Bulls in the National Basketball Association (NBA) for most of his career.

Prompt with a natural language template

NT what is the research field of Michael Jordan working at UC Berkeley

Michael Jordan, a professor at UC Berkeley, works in the field of Artificial Intelligence and machine learning, particularly in statistical machine learning, probabilistic graphical models and large-scale optimization.

NT where does he live

As a language model AI, I do not have information on the personal life of individuals, including their current residence. However, it is known that Michael Jordan is a professor at the University of California, Berkeley.

NT what is the city of University of California, Berkeley

The University of California, Berkeley is located in Berkeley, California, USA.

Poor reasoning: ChatGPT does have the knowledge

Prompt Engineering

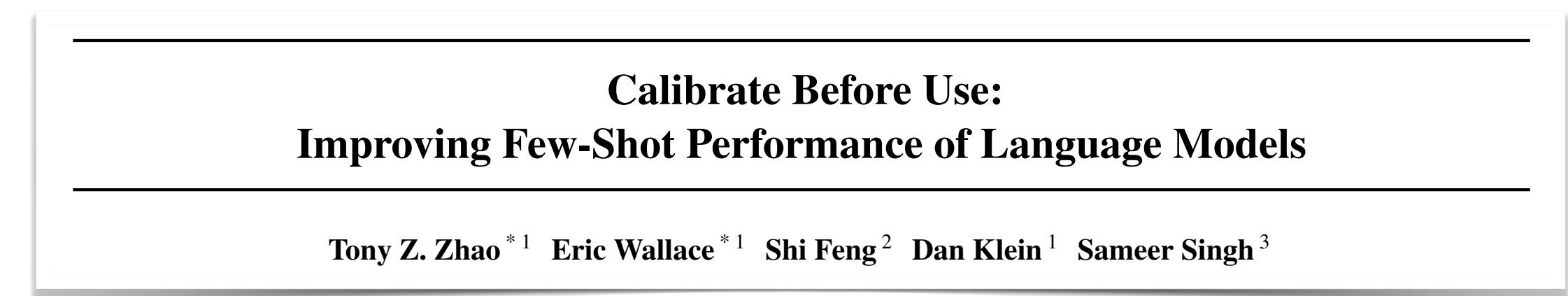
- Zero-shot
 - Prompt matters

Prompt Engineering

- Zero-shot
- Prompt matters
- Few-shot
 - Prompt training examples
 - Training example permutation

Prompt Engineering

- Zero-shot
- Prompt matters
- Few-shot
- Prompt training examples
- Training example permutation



Tony Z. Zhao ^{*1} Eric Wallace ^{*1} Shi Feng ² Dan Klein ¹ Sameer Singh ³

ICML 2021 (UC Berkeley, U Maryland, UC Irvine)

Review: the whole thing's fairly lame, making it par for the course for disney sequels.
Answer: Negative

Review: this quiet, introspective and entertaining independent is worth seeking.
Answer: Positive

Review: this quiet, introspective and entertaining independent is worth seeking.
Answer: Positive

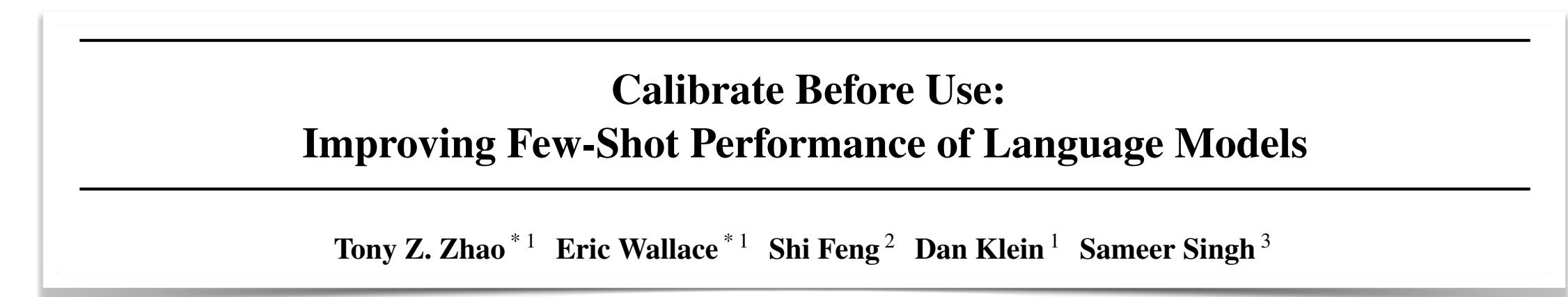
Review: the whole thing's fairly lame, making it par for the course for disney sequels .
Answer: Negative

88.5%

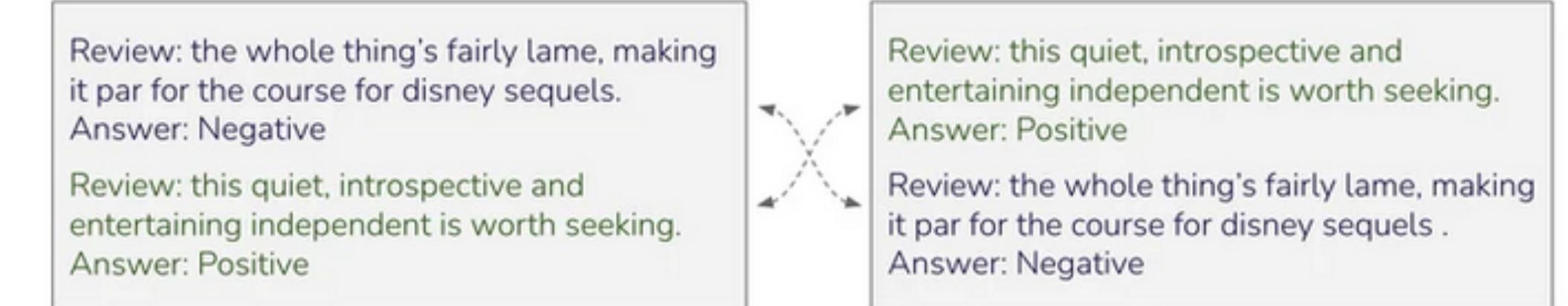
51.3%

Prompt Engineering

- Zero-shot
- Prompt matters
- Few-shot
- Prompt training examples
- Training example permutation



ICML 2021 (UC Berkeley, U Maryland, UC Irvine)



88.5%

51.3%

- Remove the bias:

Context-free :"N\A" 50%

PART I: Foundation Models: In-context Learning for Data Preparation Single Tasks

- Early successes: entity matching, error detection, string transformation
- **Limitations**
 - Prompt-sensitive
 - Dataset-sensitive
- Research Opportunities

Generic Entity Resolution Models

Jiawei Tang*
American School of Doha
23jtang@asd.edu.qa

Yifei Zuo
University of Science and Technology of China
dune@mail.ustc.edu.cn

Lei Cao
MIT CSAIL/Univeristy of Arizona
lcao@csail.mit.edu

Samuel Madden
MIT CSAIL
madden@csail.mit.edu

Abstract

Entity resolution (ER) – which decides whether two data records refer to the same real-world object – is a long-standing data integration problem. The state-of-the-art results on ER are achieved by deep learning based methods, which typically convert each pair of records into a distributed representation, followed by using a binary classifier to decide whether these two records are a match or a non-match. However, these methods are dataset specific; that is, one deep learning based model needs to be trained or fine-tuned for each new dataset, which is not generalizable and thus we call them *specific ER models*. In this paper, we investigate *generic ER models*, which use a single model to serve multiple ER datasets over different datasets from various domains. In particular, we study two types of generic ER models: Employs foundation models (*e.g.*, GPT-3) or trains a generic ER model. Our results show that although GPT-3 can perform ER with zero-shot or few-shot learning, the performance is worse than specific ER models. Our trained generic ER model can achieve comparable performance with specific ER models, but with much less train data and much smaller storage overhead.

Table Representation Learning Workshop
@ NeurIPS 2022

Generic Entity Resolution Models

Jiawei Tang*
American School of Doha
23jtang@asd.edu.qa

Yifei Zuo
University of Science and Technology of China
dune@mail.ustc.edu.cn

Lei Cao
MIT CSAIL/University of Arizona
lcao@csail.mit.edu

Samuel Madden
MIT CSAIL
madden@csail.mit.edu

Abstract

Entity resolution (ER) – which decides whether two data records refer to the same real-world object – is a long-standing data integration problem. The state-of-the-art results on ER are achieved by deep learning based methods, which typically convert each pair of records into a distributed representation, followed by using a binary classifier to decide whether these two records are a match or a non-match. However, these methods are dataset specific; that is, one deep learning based model needs to be trained or fine-tuned for each new dataset, which is not generalizable and thus we call them *specific ER models*. In this paper, we investigate *generic ER models*, which use a single model to serve multiple ER datasets over different datasets from various domains. In particular, we study two types of generic ER models: Employs foundation models (*e.g.*, GPT-3) or trains a generic ER model. Our results show that although GPT-3 can perform ER with zero-shot or few-shot learning, the performance is worse than specific ER models. Our trained generic ER model can achieve comparable performance with specific ER models, but with much less train data and much smaller storage overhead.



"GPT-3 can be easily biased with **few-shot** learning, which is **dataset relevant**."

Generic Entity Resolution Models

Jiawei Tang*
American School of Doha
23jtang@asd.edu.qa

Yifei Zuo
University of Science and Technology of China
dune@mail.ustc.edu.cn

Lei Cao
MIT CSAIL/University of Arizona
lcao@csail.mit.edu

Samuel Madden
MIT CSAIL
madden@csail.mit.edu

Abstract

Entity resolution (ER) – which decides whether two data records refer to the same real-world object – is a long-standing data integration problem. The state-of-the-art results on ER are achieved by deep learning based methods, which typically convert each pair of records into a distributed representation, followed by using a binary classifier to decide whether these two records are a match or a non-match. However, these methods are dataset specific; that is, one deep learning based model needs to be trained or fine-tuned for each new dataset, which is not generalizable and thus we call them *specific ER models*. In this paper, we investigate *generic ER models*, which use a single model to serve multiple ER datasets over different datasets from various domains. In particular, we study two types of generic ER models: Employs foundation models (*e.g.*, GPT-3) or trains a generic ER model. Our results show that although GPT-3 can perform ER with zero-shot or few-shot learning, the performance is worse than specific ER models. Our trained generic ER model can achieve comparable performance with specific ER models, but with much less train data and much smaller storage overhead.



GPT-3

"GPT-3 can be easily biased with **few-shot** learning, which is **dataset relevant**."

Table 4: Sample few-shot results of Restaurant1. Table 5: Sample few-shot results of Movies1.

GPT-3 Result	True Label
0	1
0	0
0	0
0	0
0	0
0	1
0	1
0	0
0	1
0	0

GPT-3 Result	True Label
1	1
1	1
1	1
1	0
1	0
1	1
1	1
1	0
1	1
1	0
1	0

Generic Entity Resolution Models

Jiawei Tang*
American School of Doha
23jtang@asd.edu.qa

Yifei Zuo
University of Science and Technology of China
dune@mail.ustc.edu.cn

Lei Cao
MIT CSAIL/University of Arizona
lcao@csail.mit.edu

Samuel Madden
MIT CSAIL
madden@csail.mit.edu

Abstract

Entity resolution (ER) – which decides whether two data records refer to the same real-world object – is a long-standing data integration problem. The state-of-the-art results on ER are achieved by deep learning based methods, which typically convert each pair of records into a distributed representation, followed by using a binary classifier to decide whether these two records are a match or a non-match. However, these methods are dataset specific; that is, one deep learning based model needs to be trained or fine-tuned for each new dataset, which is not generalizable and thus we call them *specific ER models*. In this paper, we investigate *generic ER models*, which use a single model to serve multiple ER datasets over different datasets from various domains. In particular, we study two types of generic ER models: Employs foundation models (*e.g.*, GPT-3) or trains a generic ER model. Our results show that although GPT-3 can perform ER with zero-shot or few-shot learning, the performance is worse than specific ER models. Our trained generic ER model can achieve comparable performance with specific ER models, but with much less train data and much smaller storage overhead.

Table Representation Learning Workshop
@ NeurIPS 2022



GPT-3

"GPT-3 can be easily biased with **few-shot** learning, which is **dataset relevant**."

Table 4: Sample few-shot results of Restaurant1. Table 5: Sample few-shot results of Movies1.

GPT-3 Result	True Label
0	1
0	0
0	0
0	0
0	0
0	1
0	1
0	0
0	1
0	0

GPT-3 Result	True Label
1	1
1	1
1	1
1	0
1	0
1	1
1	1
1	0
1	1
1	0
1	0



BERT

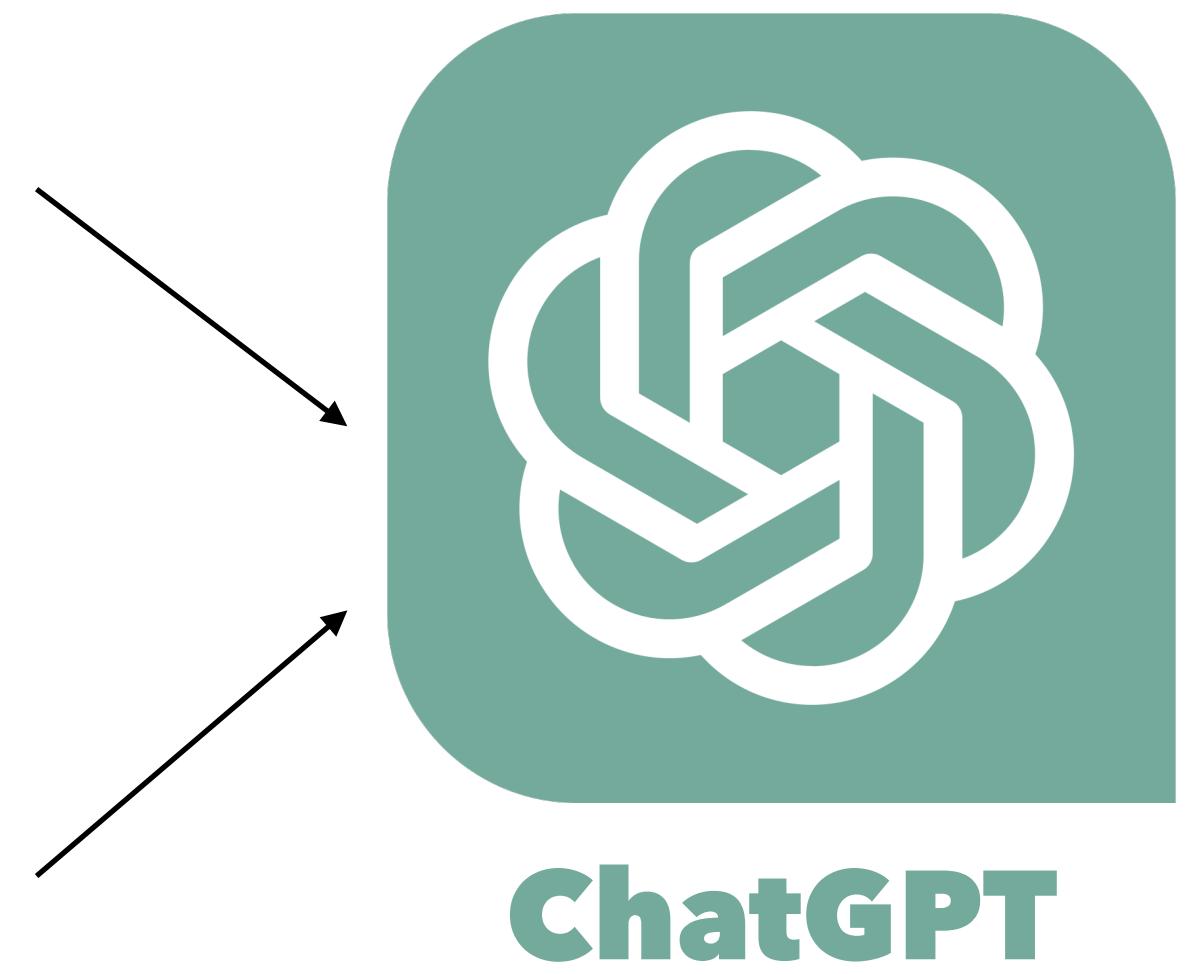
Entity Matching
Adaptation > In-Context Learning

PART I: Foundation Models: In-context Learning for Data Preparation Single Tasks

- Early successes: entity matching, error detection, string transformation
- **Limitations**
 - Prompt-sensitive
 - Dataset-sensitive
 - Not reliable
- Research Opportunities

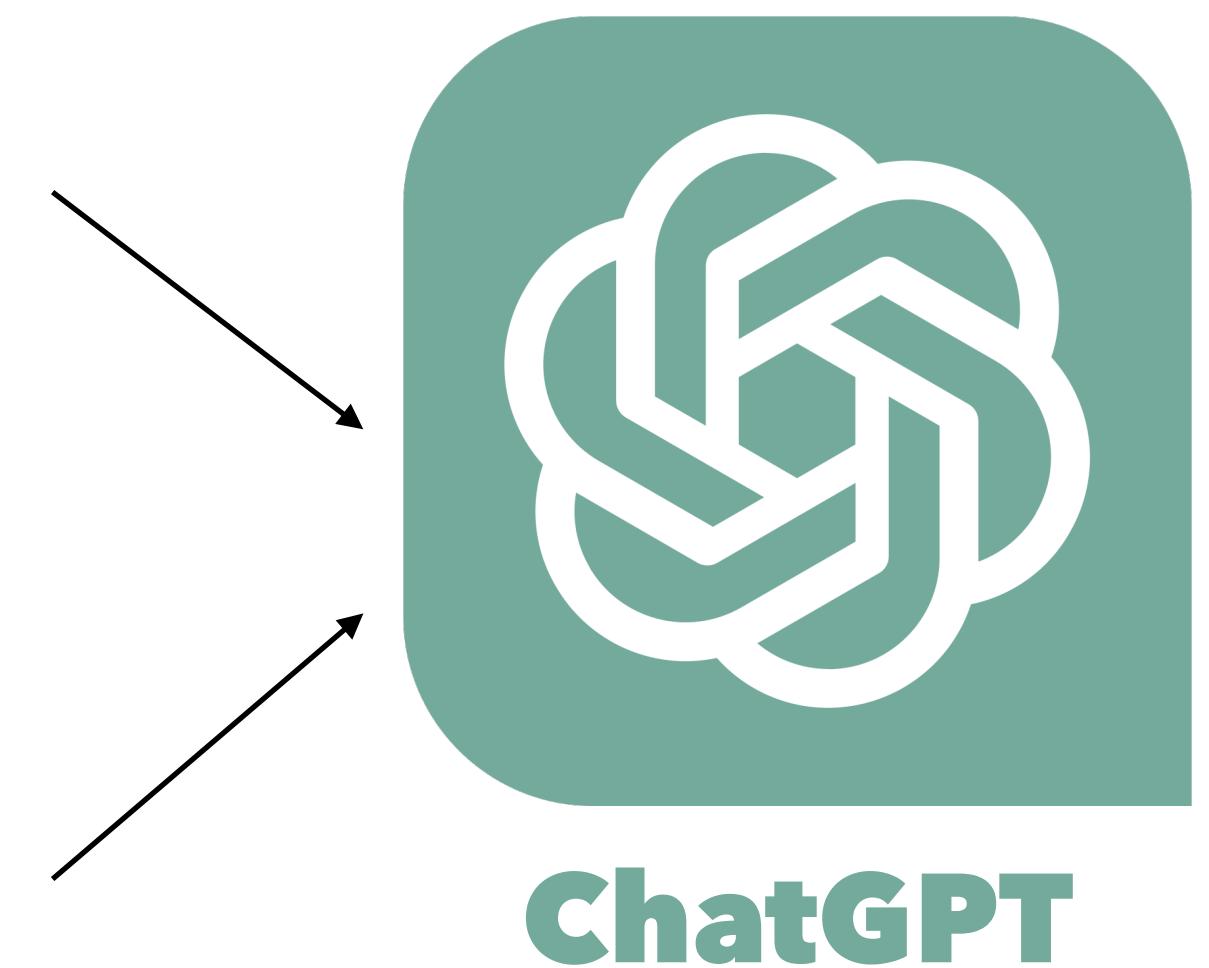
Example from Faisal Farooq

- WHO says COVID-19 is no longer a global health emergency
- Who says COVID-19 is no longer a global health emergency



Example from Faisal Farooq

- WHO says COVID-19 is no longer a global health emergency
- Who says COVID-19 is no longer a global health emergency



→ Exact same answer

PART I: Foundation Models: In-context Learning for Data Preparation Single Tasks

- Early successes: entity matching, error detection, string transformation
- **Limitations**
 - Prompt-sensitive
 - Dataset-sensitive
 - Not reliable
 - Research Opportunities
 - Not up-to-date
 - Not complete
 - Poor reasoning (tabular data)
 - Not explainable

PART I: Foundation Models: In-context Learning for Data Preparation Single Tasks

- Early successes: entity matching, error detection, string transformation
- Limitations
- **Research Opportunities**
 - Table learning
 - Routing-based methods with modular architectures
 - Retrieval-based models

Table Learning

Table Representation Learning workshop

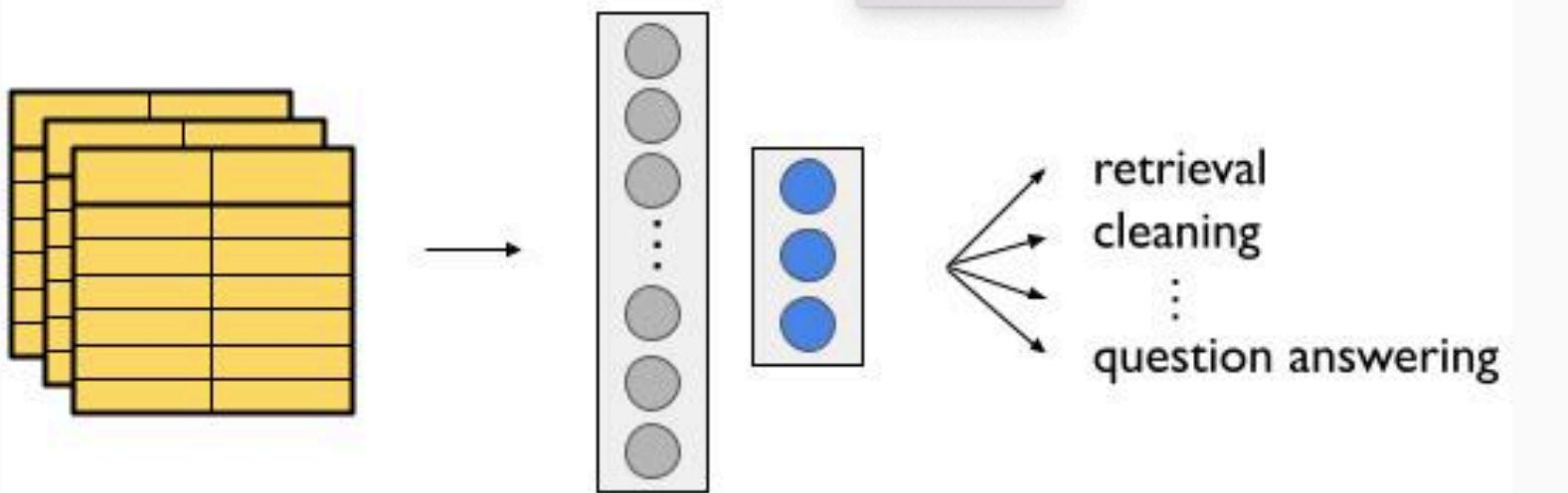
NeurIPS 2022, Room 398, Friday 2 December, New Orleans, USA

FAQ

table-representation-learning-workshop@googlegroups.com

 @TrlWorkshop

TrlWorkshop



About

We develop large models to “understand” images, videos and natural language that fuel many intelligent applications from text completion to self-driving cars. But tabular data has long been overlooked despite its dominant presence in data-intensive systems. By learning latent representations from (semi-)structured tabular data, pretrained table models have shown preliminary but impressive performance for semantic parsing, question answering, table understanding, and data preparation. Considering that such tasks share fundamental properties inherent to tables, representation learning for tabular data is an important direction to explore further. These works also surfaced many open challenges such as finding effective data encodings, pretraining objectives and downstream tasks.

The Table Representation Learning workshop is the first workshop in this emerging research area and has the following main goals: 1) motivating tabular data as a first-class modality for representation learning and further shaping this area, 2) show-casing impactful applications of pretrained table models and discussing future opportunities thereof, and 3) facilitating discussion and collaboration across the machine learning, natural language processing, and data management communities.

Table Learning

Table Representation Learning workshop

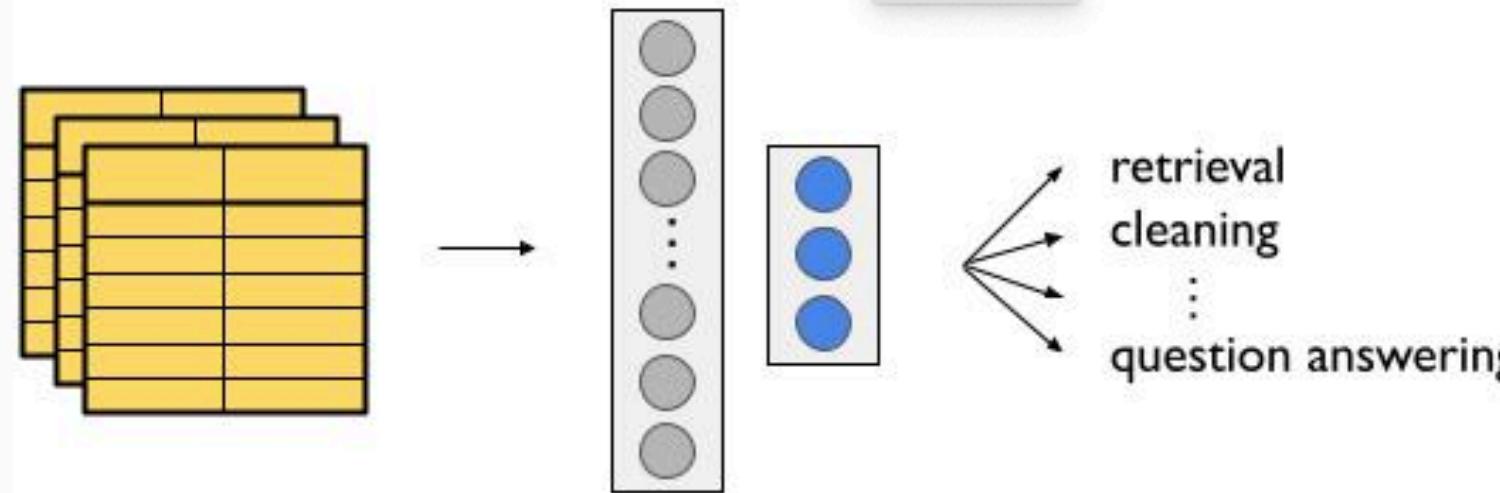
NeurIPS 2022, Room 398, Friday 2 December, New Orleans, USA

FAQ

table-representation-learning-workshop@googlegroups.com

@TrlWorkshop

TrlWorkshop



About

We develop large models to “understand” images, videos and natural language that fuel many intelligent applications from text completion to self-driving cars. But tabular data has long been overlooked despite its dominant presence in data-intensive systems. By learning latent representations from (semi-)structured tabular data, pretrained table models have shown preliminary but impressive performance for semantic parsing, question answering, table understanding, and data preparation. Considering that such tasks share fundamental properties inherent to tables, representation learning for tabular data is an important direction to explore further. These works also surfaced many open challenges such as finding effective data encodings, pretraining objectives and downstream tasks.

The Table Representation Learning workshop is the first workshop in this emerging research area and has the following main goals: 1) motivating tabular data as a first-class modality for representation learning and further shaping this area, 2) show-casing impactful applications of pretrained table models and discussing future opportunities thereof, and 3) facilitating discussion and collaboration across the machine learning, natural language processing, and data management communities.

RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation

Nan Tang
QCRI, HBKU, Qatar
ntang@hbku.edu.qa

Ju Fan
Renmin University, China
fanj@ruc.edu.cn

Fangyi Li
Renmin University, China
fangyili@ruc.edu.cn

Jianhong Tu
Renmin University, China
tujh@ruc.edu.cn

Xiaoyong Du
Renmin University, China
duyong@ruc.edu.cn

Guoliang Li
Tsinghua University, China
liguoliang@tsinghua.edu.cn

Sam Madden
CSAIL, MIT, USA
madden@csail.mit.edu

Mourad Ouzzani
QCRI, HBKU, Qatar
mouzzani@hbku.edu.qa

VLDB 2021

TURL: Table Understanding through Representation Learning

Xiang Deng*
The Ohio State University
Columbus, Ohio
deng.595@buckeyemail.osu.edu

Huan Sun*
The Ohio State University
Columbus, Ohio
sun.397@osu.edu

Alyssa Lees
Google Research
New York, NY
alyssalees@google.com

You Wu
Google Research
New York, NY
wuyou@google.com

Cong Yu
Google Research
New York, NY
congyu@google.com

VLDB 2021

Table Learning

Table Representation Learning workshop

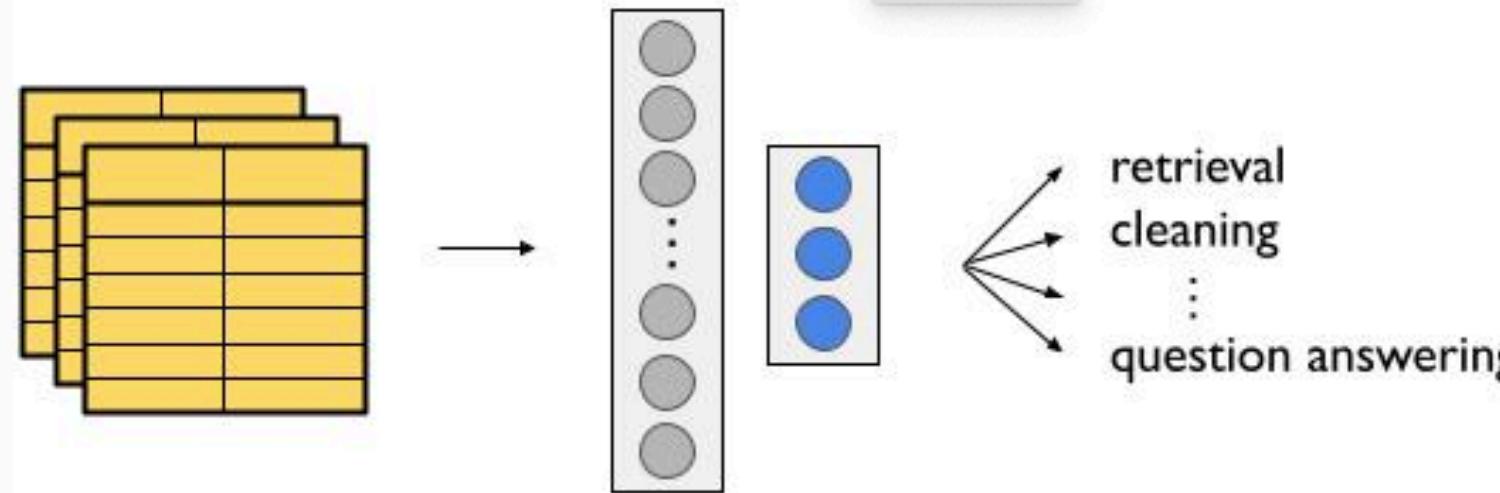
NeurIPS 2022, Room 398, Friday 2 December, New Orleans, USA

FAQ

table-representation-learning-workshop@googlegroups.com

@TrlWorkshop

TrlWorkshop



About

We develop large models to "understand" images, videos and natural language that fuel many intelligent applications from text completion to self-driving cars. But tabular data has long been overlooked despite its dominant presence in data-intensive systems. By learning latent representations from (semi-)structured tabular data, pretrained table models have shown preliminary but impressive performance for semantic parsing, question answering, table understanding, and data preparation. Considering that such tasks share fundamental properties inherent to tables, representation learning for tabular data is an important direction to explore further. These works also surfaced many open challenges such as finding effective data encodings, pretraining objectives and downstream tasks.

The Table Representation Learning workshop is the first workshop in this emerging research area and has the following main goals: 1) motivating tabular data as a first-class modality for representation learning and further shaping this area, 2) show-casing impactful applications of pretrained table models and discussing future opportunities thereof, and 3) facilitating discussion and collaboration across the machine learning, natural language processing, and data management communities.

RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation

Nan Tang
QCRI, HBKU, Qatar
ntang@hbku.edu.qa

Xiaoyong Du
Renmin University, China
duyong@ruc.edu.cn

Ju Fan
Renmin University, China
fanj@ruc.edu.cn

Guoliang Li
Tsinghua University, China
liguoliang@tsinghua.edu.cn

Fangyi Li
Renmin University, China
fangyili@ruc.edu.cn

Sam Madden
CSAIL, MIT, USA
madden@csail.mit.edu

Jianhong Tu
Renmin University, China
tujh@ruc.edu.cn

Mourad Ouzzani
QCRI, HBKU, Qatar
mouzzani@hbku.edu.qa

VLDB 2021

TURL: Table Understanding through Representation Learning

Xiang Deng*
The Ohio State University
Columbus, Ohio
deng.595@buckeyemail.osu.edu

Huan Sun*
The Ohio State University
Columbus, Ohio
sun.397@osu.edu

Alyssa Lees
Google Research
New York, NY
alyssalees@google.com

You Wu
Google Research
New York, NY
wuyou@google.com

Cong Yu
Google Research
New York, NY
congyu@google.com

VLDB 2021

Transformers for Tabular Data Representation: A Tutorial on Models and Applications

Gilbert Badaro
EURECOM
Biot, France
gilbert.badaro@eurecom.fr

Paolo Papotti
EURECOM
Biot, France
paolo.papotti@eurecom.fr

VLDB 2022 Tutorial

Table Learning (e.g. RPT)

Name	Expertise	City
Michael Jordan	Machine Learning	Berkeley

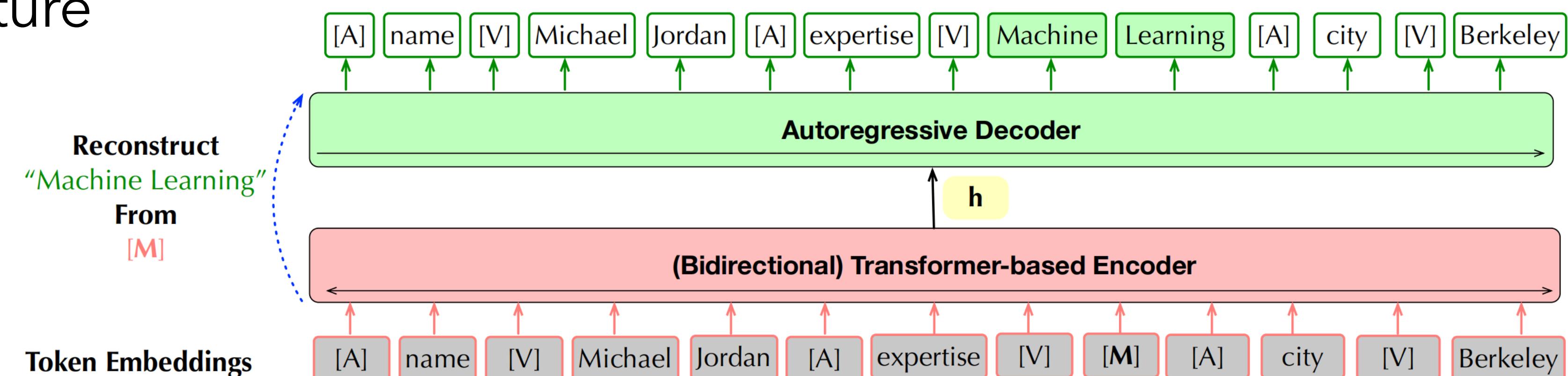
- Tuple Tokenization

```
name Michael Jordan expertise Machine Learning city Berkeley
```

- Token Embedding

```
[A] name [V] Michael Jordan [A] expertise [V] Machine Learning  
[A] city [V] Berkeley
```

- RPT Architecture

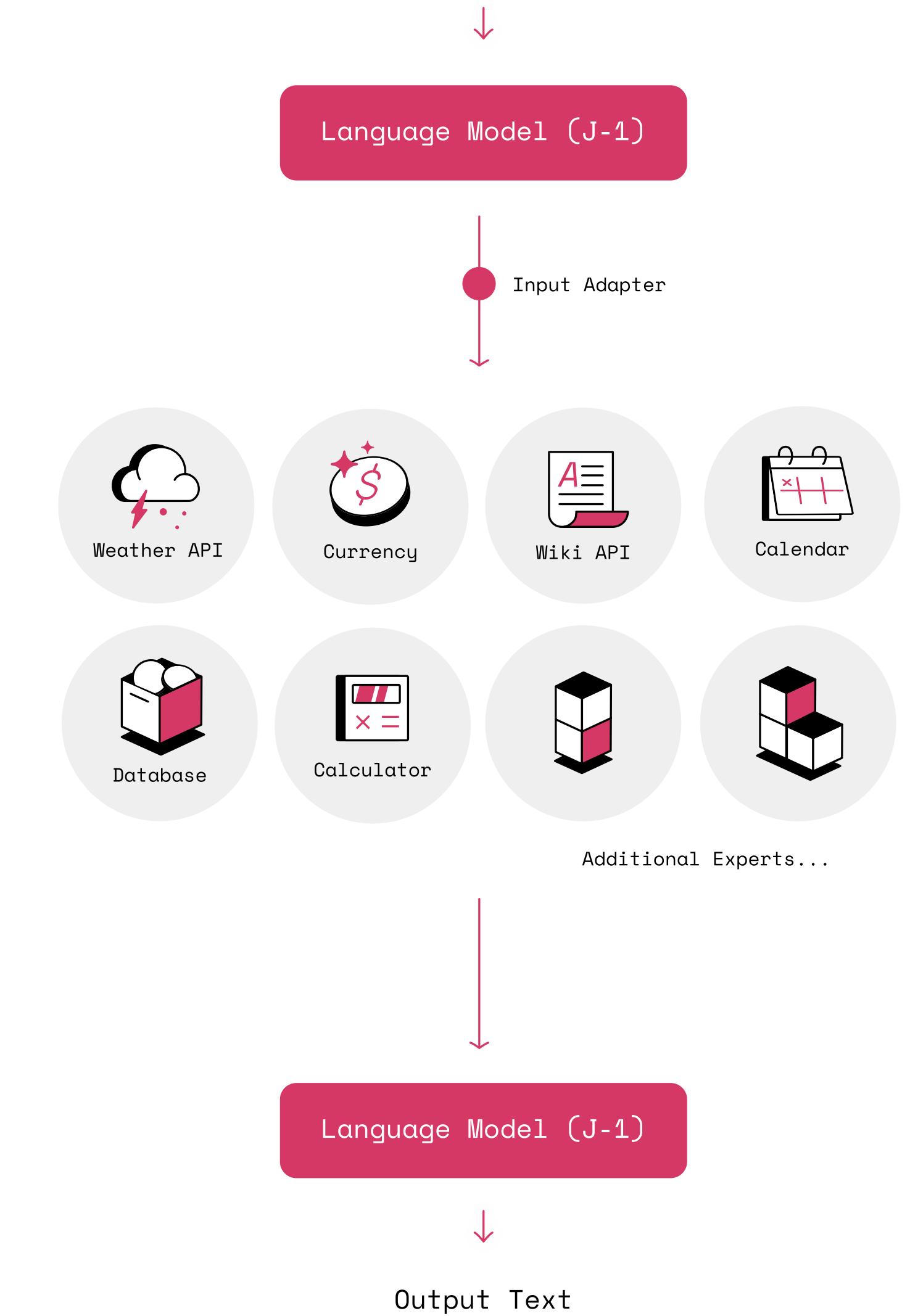
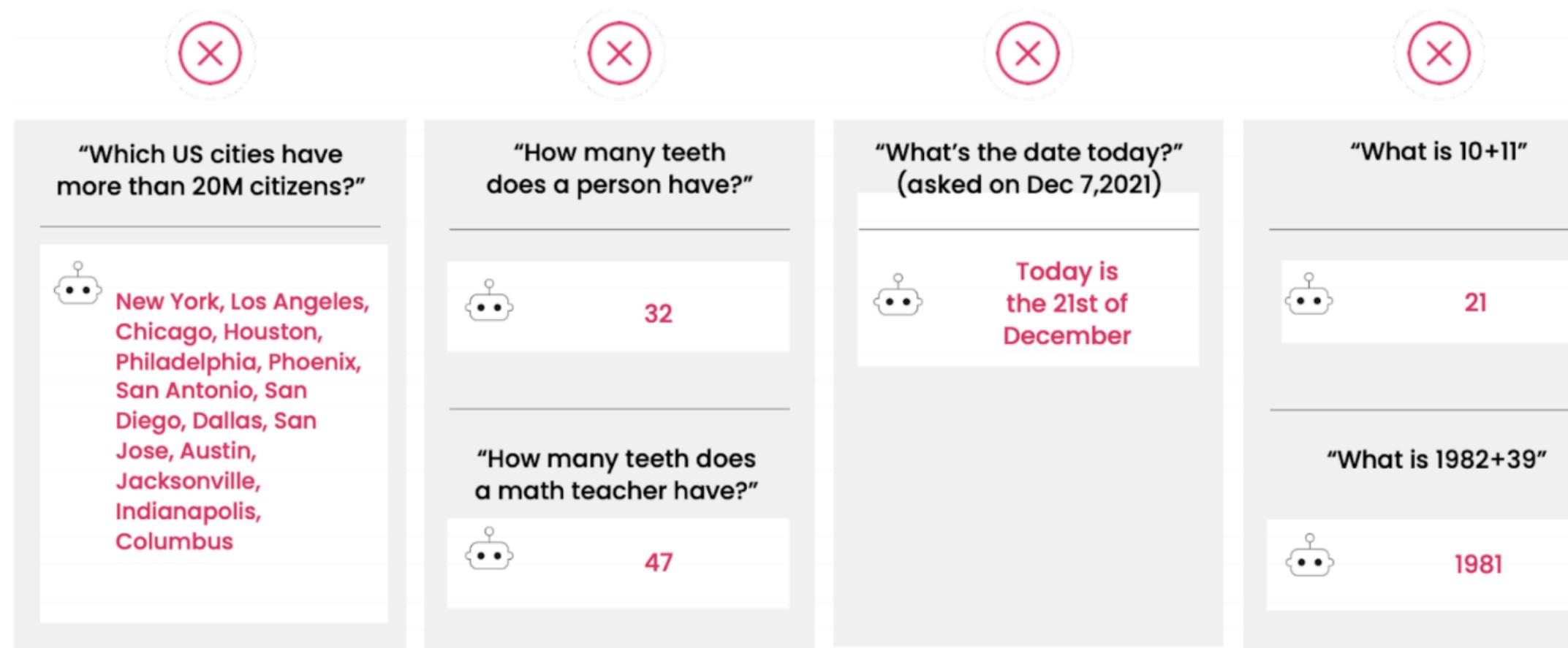
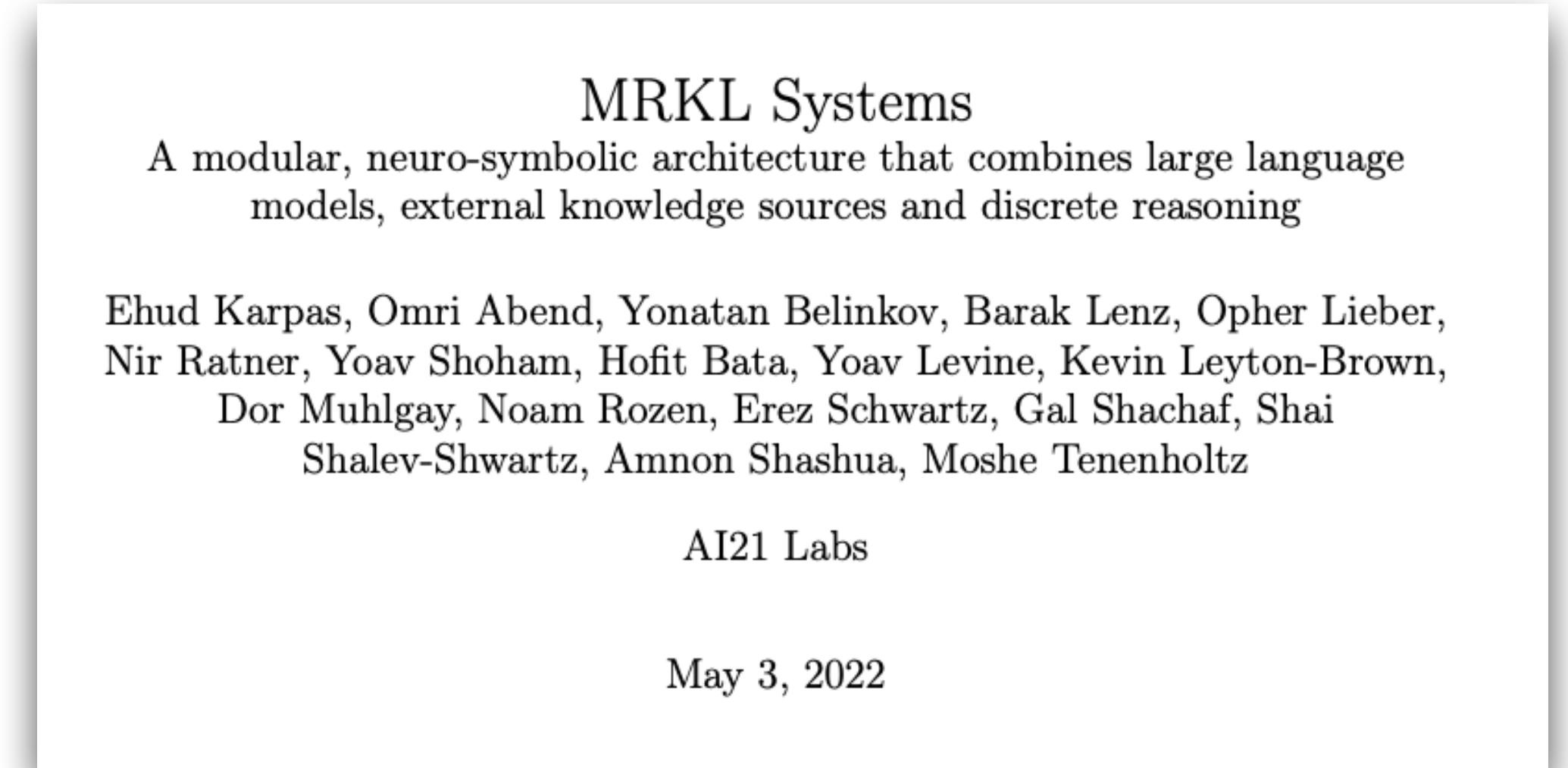


PART I: Foundation Models: In-context Learning for Data Preparation Single Tasks

- Early successes: entity matching, error detection, string transformation
- Limitations
- **Research Opportunities**
 - Table learning
 - Routing-based methods with modular architectures
 - Retrieval-based models

A Modular, Neuro-Symbolic Architecture

MRKL (Modular Reasoning, Knowledge and Language)



SYMPHONY: Towards Natural Language Query Answering over Multi-modal Data Lakes

Zui Chen*

Tsinghua University
China

chenzui19@mails.tsinghua.edu.cn

Zihui Gu*

Renmin University of China
China

guzh@ruc.edu.cn

Lei Cao

MIT CSAIL/University of Arizona
USA

lcao@csail.mit.edu

Ju Fan

Renmin University of China
China

fanj@ruc.edu.cn

Sam Madden

MIT CSAIL
USA

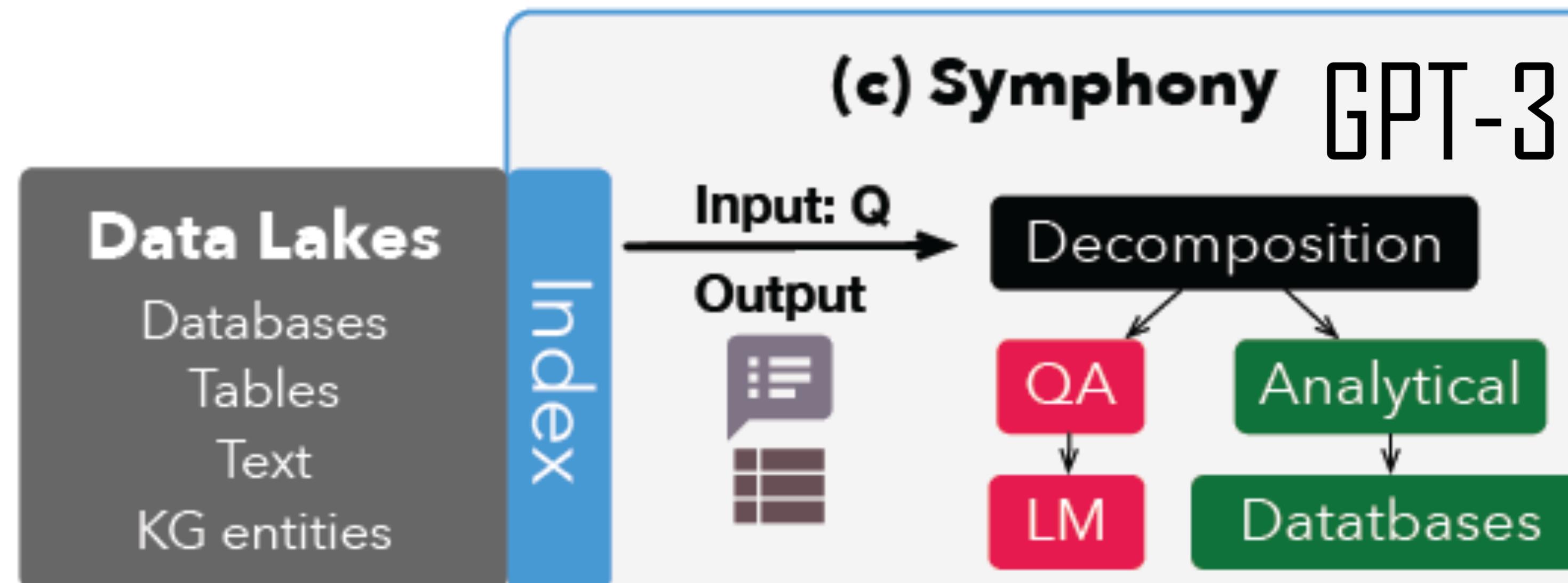
madden@csail.mit.edu

Nan Tang

QCRI, HBKU
Qatar

ntang@hbku.edu.qa

CIDR 2023



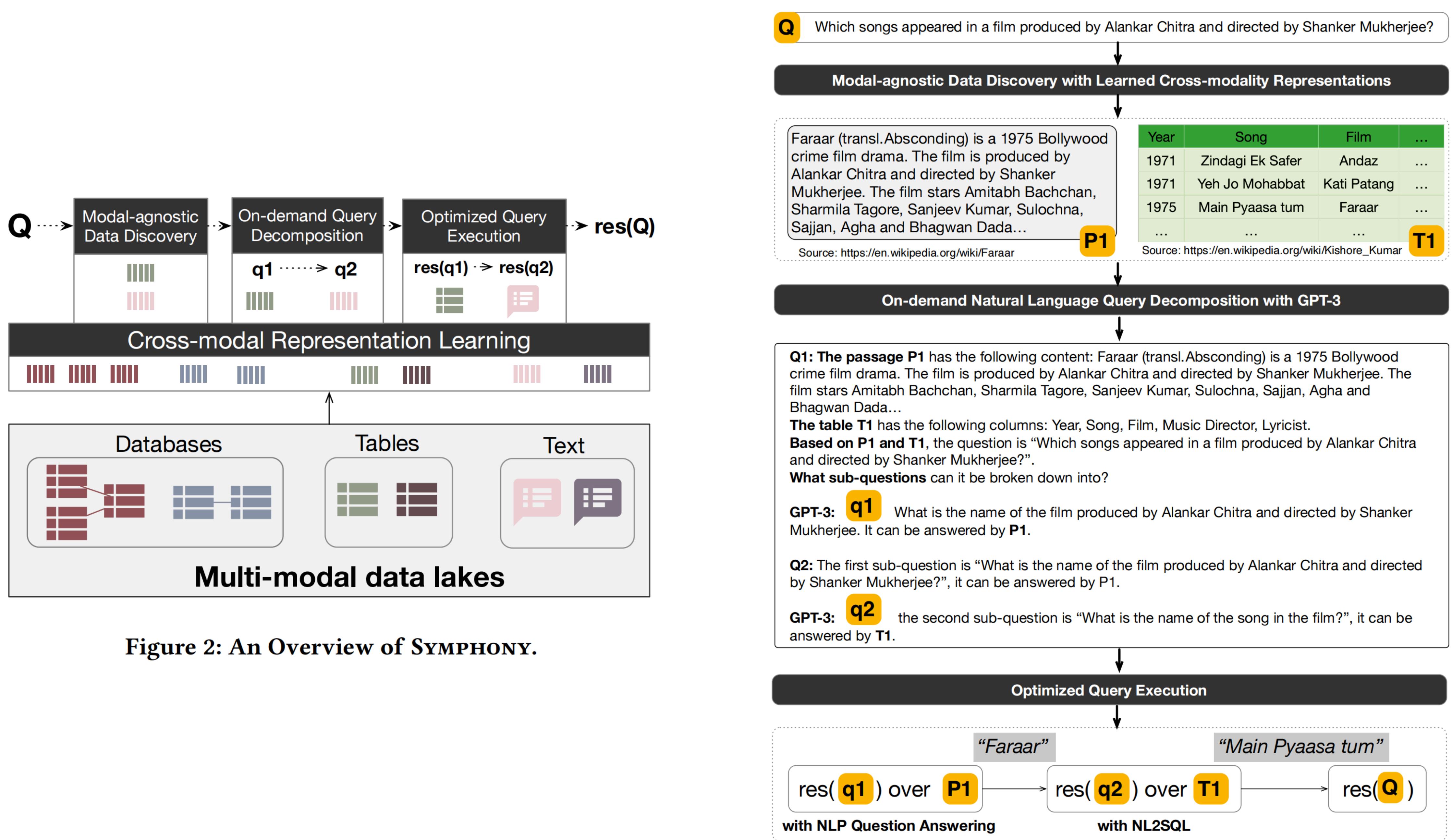


Figure 3: A Running Example of SYMPHONY.

PART I: Foundation Models: In-context Learning for Data Preparation Single Tasks

- Early successes: entity matching, error detection, string transformation
- Limitations
- **Research Opportunities**
 - Table learning
 - Routing-based methods with modular architectures
 - Retrieval-based models

Retrieval-based Models

Improving language models by retrieving
from trillions of tokens

Sebastian Borgeaud[†], Arthur Mensch[†], Jordan Hoffmann[†], Trevor Cai, Eliza Rutherford, Katie Millican,
George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas,
Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones,
Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero,
Karen Simonyan, Jack W. Rae[‡], Erich Elsen[‡] and Laurent Sifre^{†,‡}

All authors from DeepMind, [†]Equal contributions, [‡]Equal senior authorship

Retrieval-based Models



Improving language models by retrieving from trillions of tokens

Sebastian Borgeaud[†], Arthur Mensch[†], Jordan Hoffmann[†], Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae[‡], Erich Elsen[‡] and Laurent Sifre^{†,‡}

All authors from DeepMind, [†]Equal contributions, [‡]Equal senior authorship

INPUT The Dune film was released in _____



LARGE GPT

A Language Information

World Knowledge Information



OUTPUT

2021

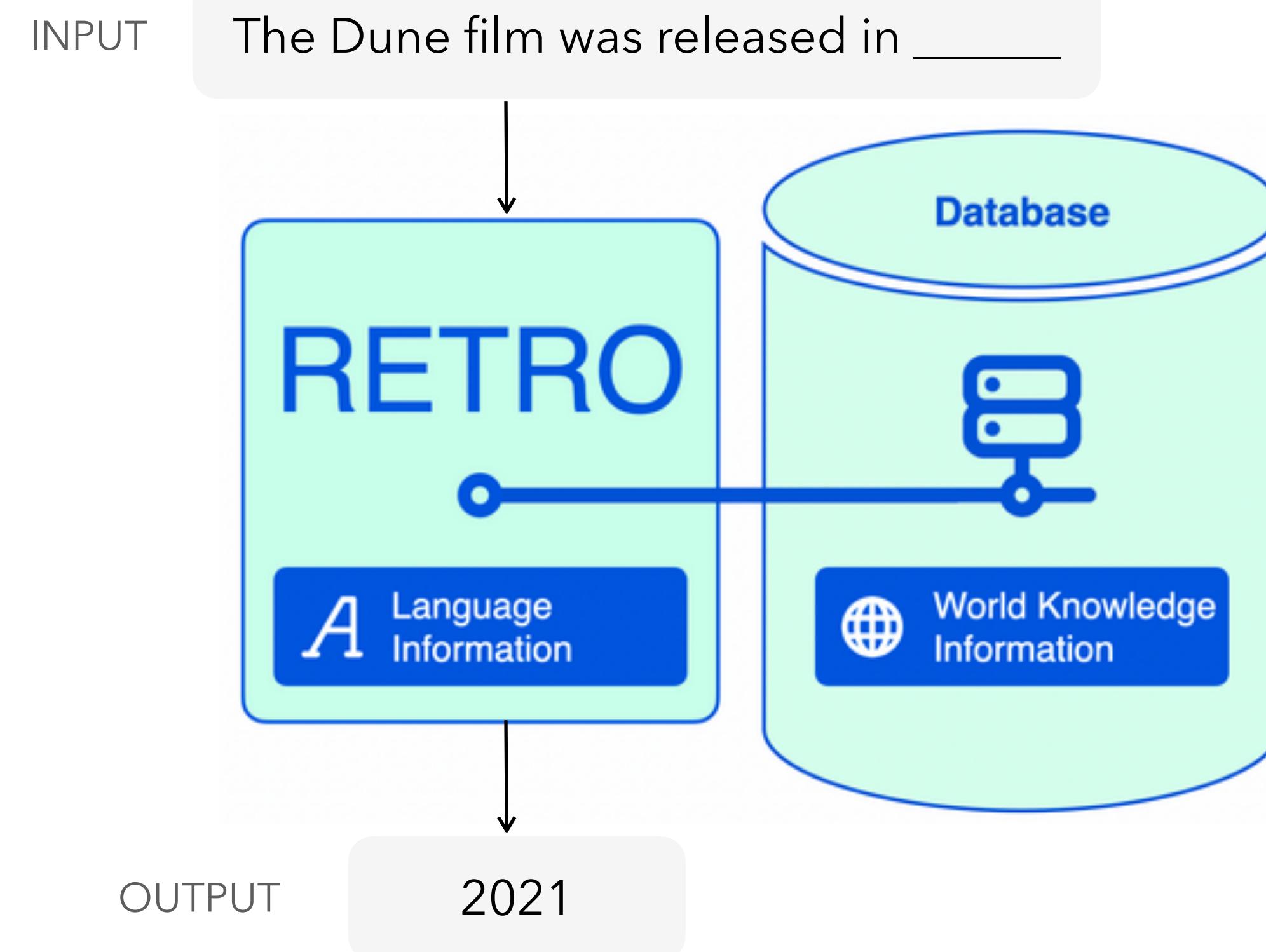
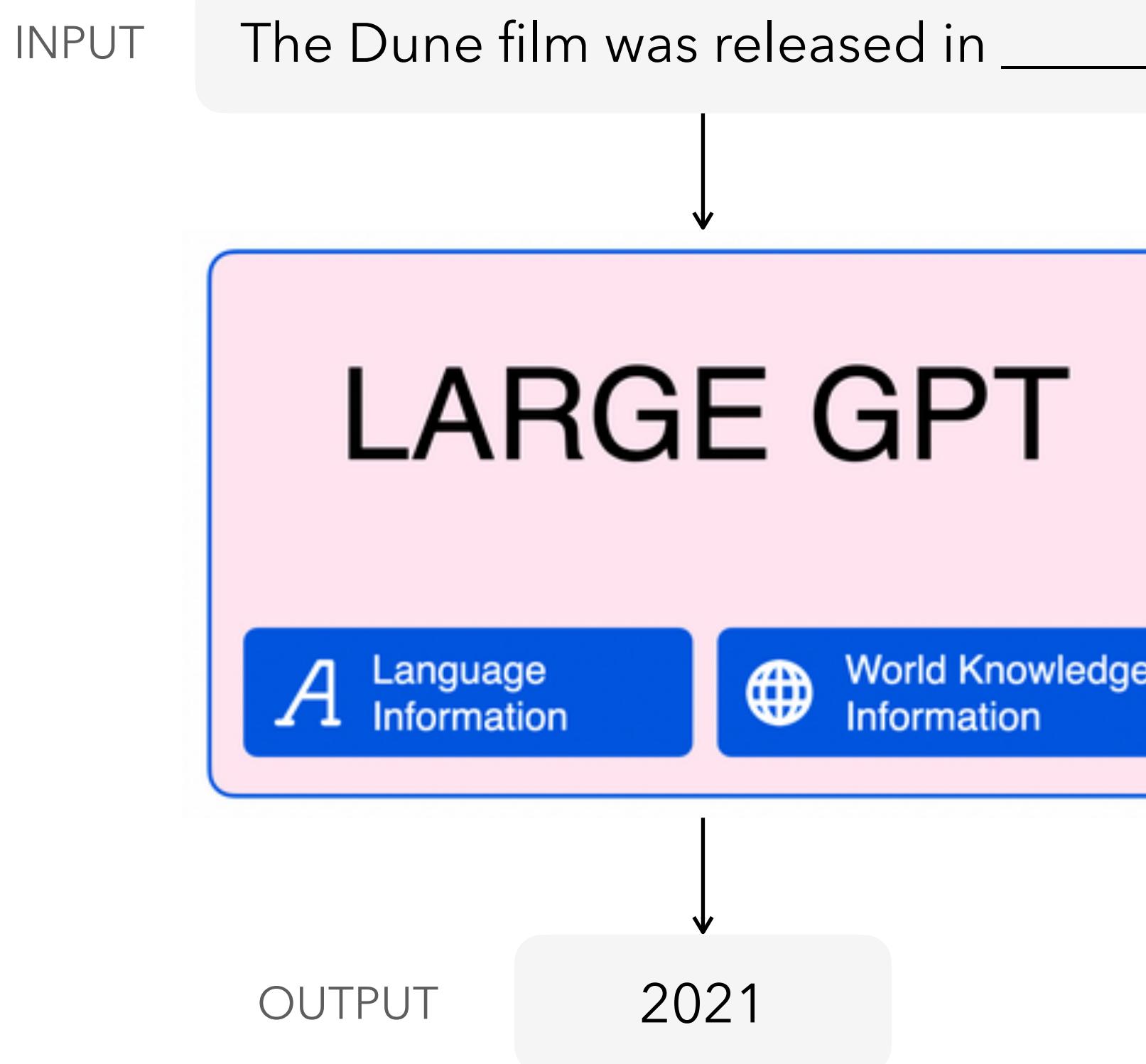
Retrieval-based Models



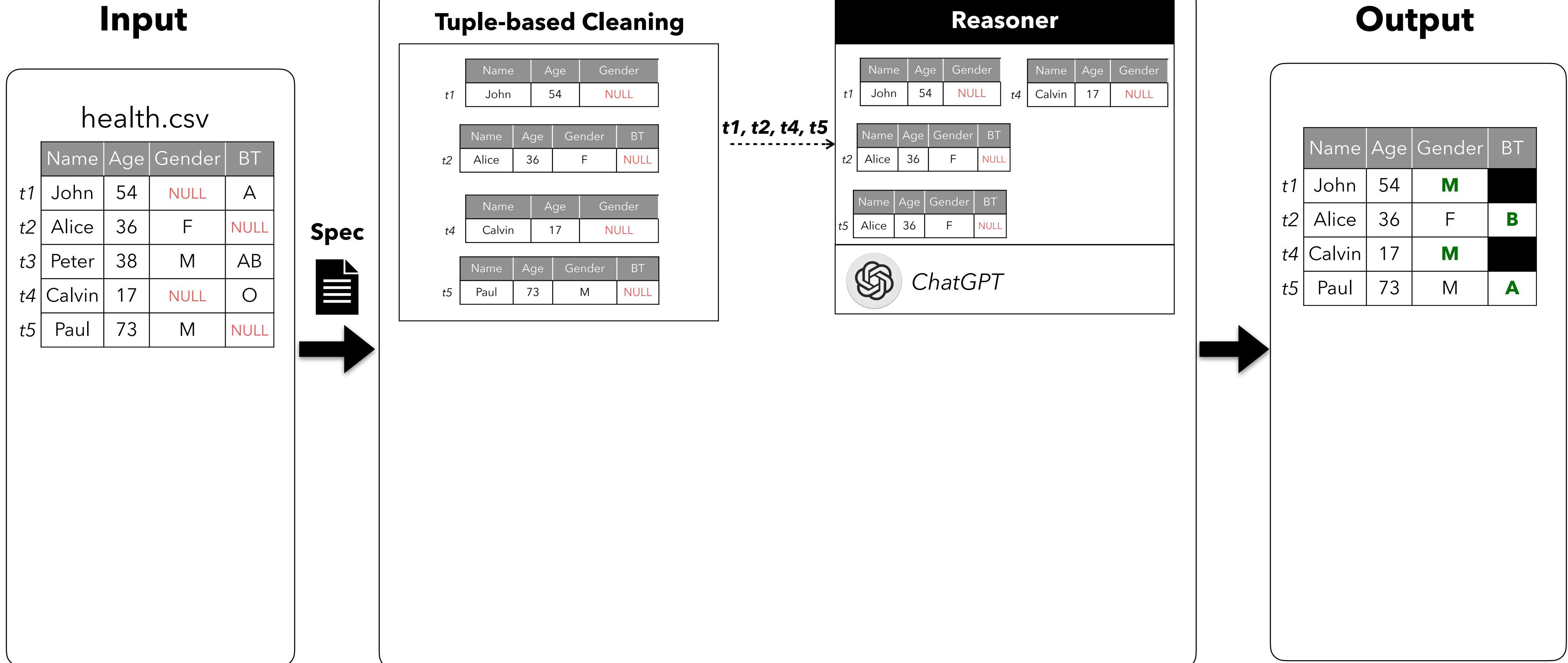
Improving language models by retrieving from trillions of tokens

Sebastian Borgeaud[†], Arthur Mensch[†], Jordan Hoffmann[†], Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae[‡], Erich Elsen[‡] and Laurent Sifre^{†,‡}

All authors from DeepMind, [†]Equal contributions, [‡]Equal senior authorship



Non Retrieval-based Data Cleaning



Non Retrieval-based Data Cleaning

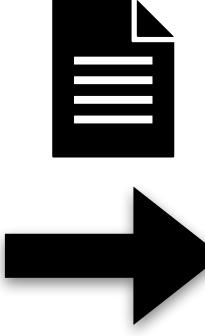
- Not reliable
- Sensitive data

Input

health.csv

	Name	Age	Gender	BT
t1	John	54	NULL	A
t2	Alice	36	F	NULL
t3	Peter	38	M	AB
t4	Calvin	17	NULL	O
t5	Paul	73	M	NULL

Spec



Tuple-based Cleaning

	Name	Age	Gender
t1	John	54	NULL
t2	Alice	36	F
t4	Calvin	17	NULL
t5	Paul	73	M

t1, t2, t4, t5



Reasoner

	Name	Age	Gender	BT
t1	John	54	NULL	
t2	Alice	36	F	NULL
t5	Alice	36	F	NULL



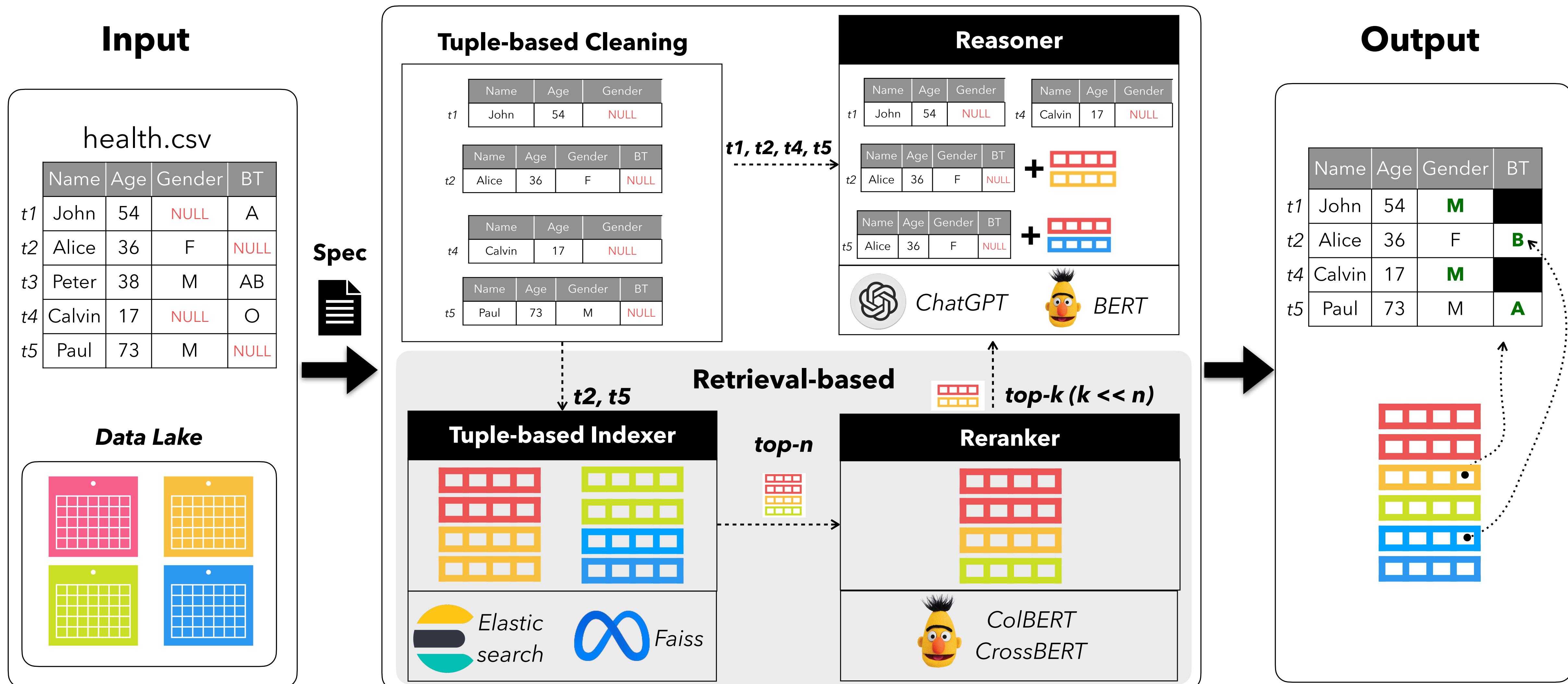
ChatGPT

Output

	Name	Age	Gender	BT
t1	John	54	M	
t2	Alice	36	F	B
t4	Calvin	17	M	
t5	Paul	73	M	A

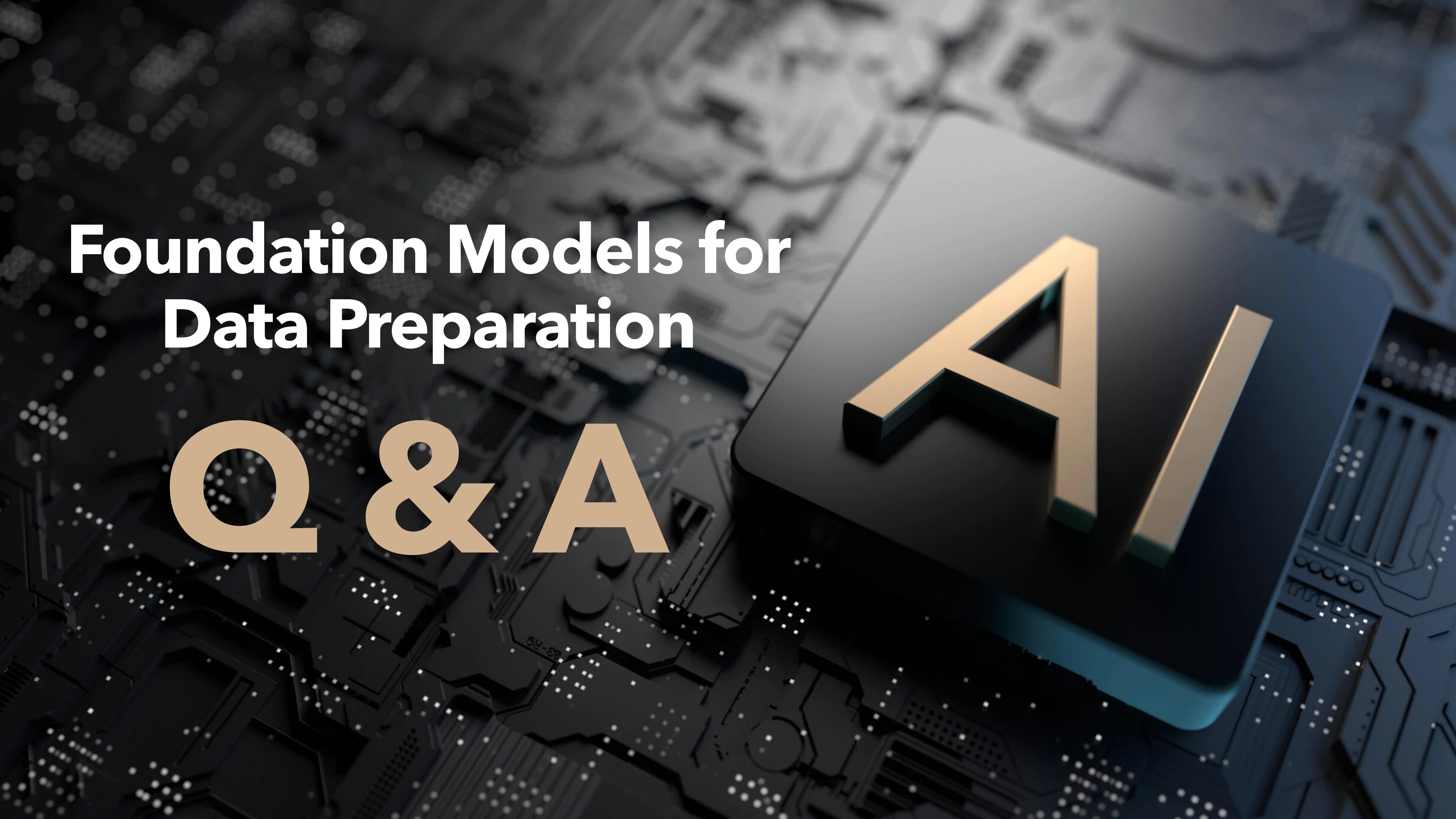
Retrieval-based Data Cleaning

- Not reliable - retrieved data
 - Sensitive data - local models



Conclusions

- **Positioning**
 - Foundation models **cannot** replace humans for data preparation
 - Treat/tune foundation models as your **thought partners**
- **Promising directions**
 - Adapting foundation models for specific tasks (discussed next)
 - Enhancing foundation models with better reasoning (e.g., modularized design)
 - Enhancing foundation models with better explainability (e.g., retrieval-based)



Foundation Models for Data Preparation

Q & A