RelationalAI

Knowledge Graphs and Massive Language Models The future of AI

Vijay Saraswat, Nikolaos Vasiloglou*

Co* Content presented here is the responsibility of the authors and does not necessarily represent the view of their employers. Does not not either…

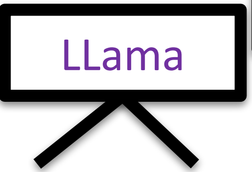
Try all the prompts on your own

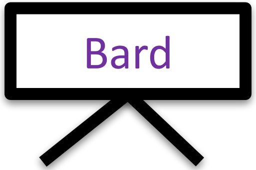
https://bit.ly/raikgc2023

Please Use GPT-4

2

What just happened?





We now have

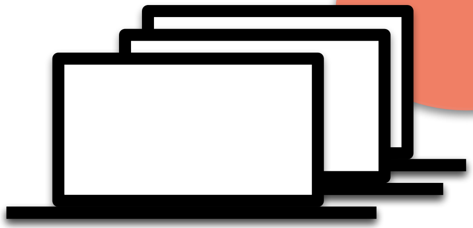
a family of instructable computers

What is the task to be done
What is the form of the output.

that can read all language

(text, tables, documents, data, code, images, audio, pictures, video, …)

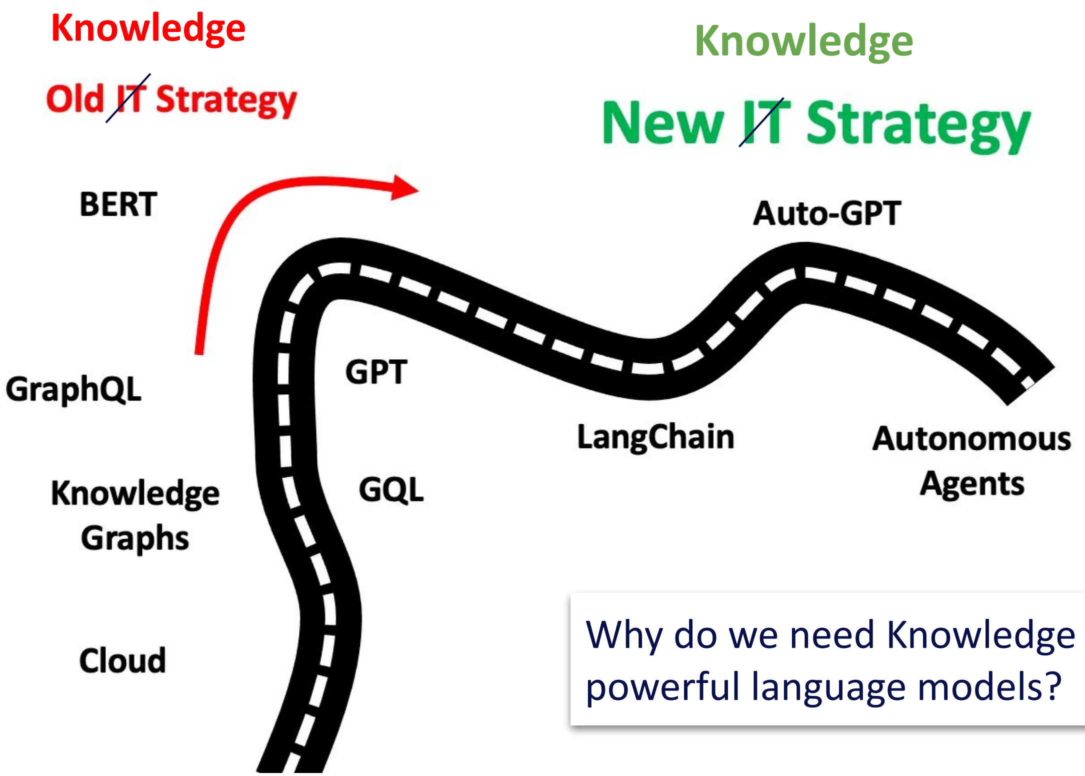
accessible to us,
understand it, and
communicate with us
(and each other)
(somewhat) as humans do.



The First Generation of Computers for Humans

*Language: What humans use to communicate

The "Generative Turn"

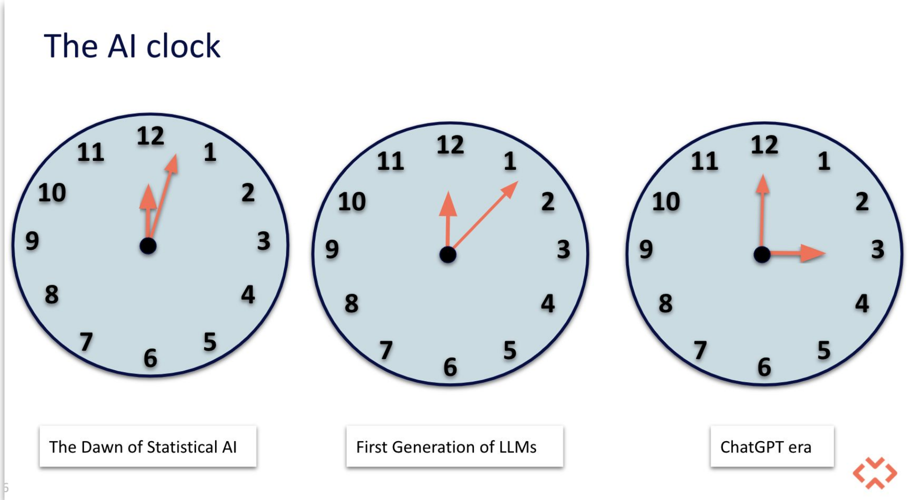


A moment where what we previously understood as how everything from illustration to film directing to publishing works, is all about to change very rapidly.
– Kate Crawford

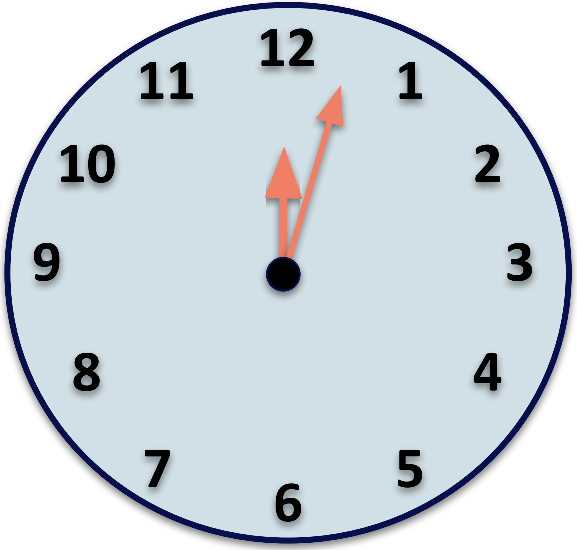
Why do we need Knowledge graphs when we have such powerful language models?

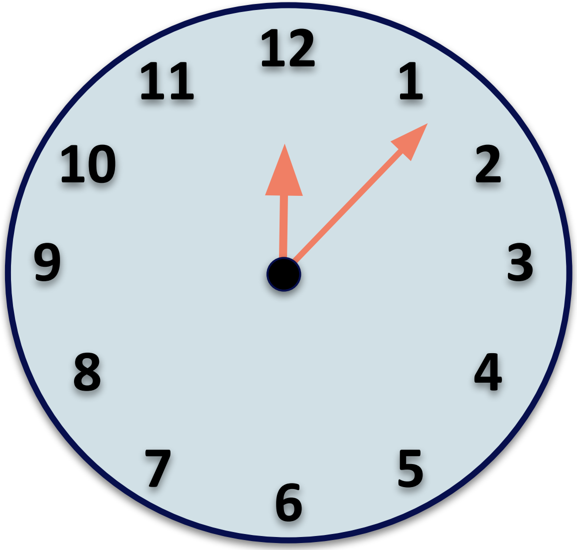
Agenda

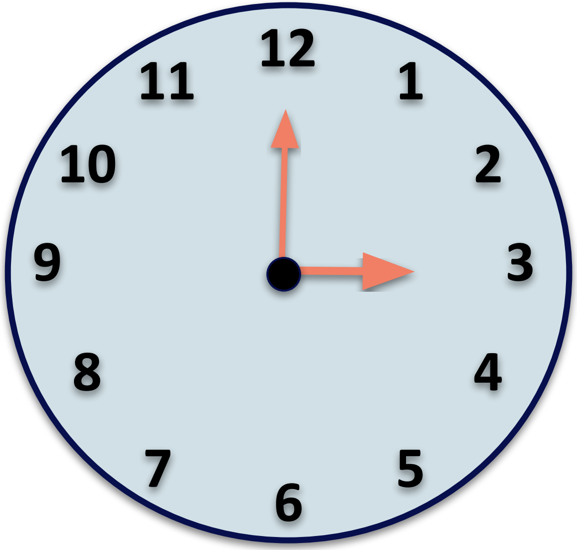
- How did we get here?
- Working with unstructured data
- Working with structured data
- LLM Tradecraft
- Language Machines as the Knowledge Hub
- What is the role of the Knowledge Graph in the new world?



The AI clock







The Dawn of Statistical AI

First Generation of LLMs

ChatGPT era

How we got here

From linguistics to ChatGPT

Computer Science 1.0

Algorithms = Logic + Control
Programs = Algorithms + Data-Structures

Human-generated Code

Programming Models

Assembly

Fortran Lisp

Pascal, C Prolog

C++, Java ML Esterel

Scala X10 Go Python

Rust

Declarative v Prescriptive

Correctness – Type systems, Verification

Transformational v Reactive

Sketching

"When you know how to do it, that's Computer Science, otherwise it's AI"

8

Computer Science 2.0

RelationalAI

Programs = Logic + Data

input

↓

Agent

↓

label

Data is noisy, uncertain , high-dimensional [1]

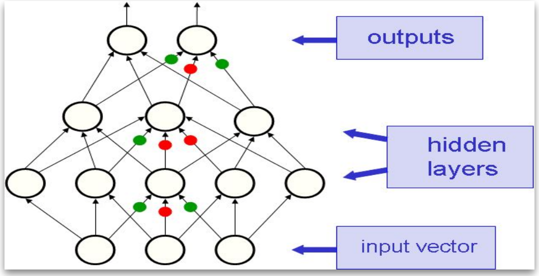
We move to compute with continuous functions.

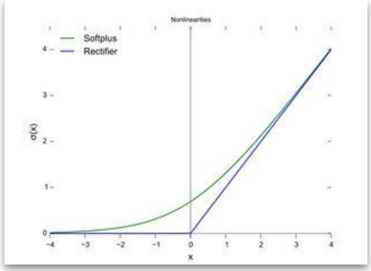
Can we synthesize the code for the agent, simply given (a lot of) examples of input/output pairs?

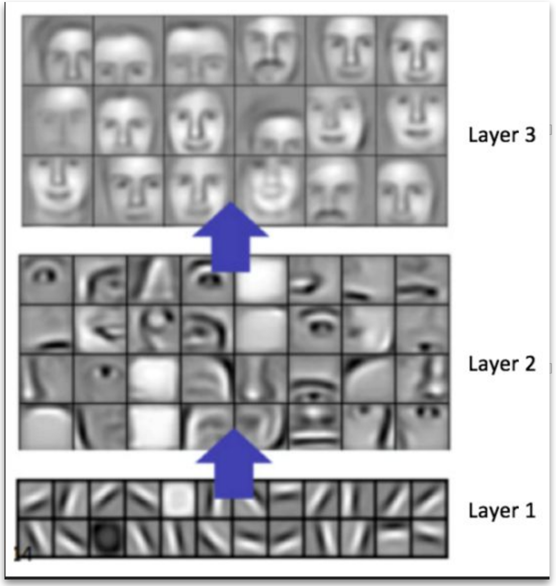
What if we don't know how to write the code?

[1] Sejnowski, et. al., "The unreasonable effectiveness of deep learning in AI", PNAS, 2019

Deep Representations







- Each computing element takes sum of (possibly many) inputs multiplied by learnt weights, and then applies a non-linear function

- 2 layers of neurons (linear weights plus non-linear activation) are adequate to represent any function
- But: Deep layers provide an exponential advantage

Parameterized Function Approximation

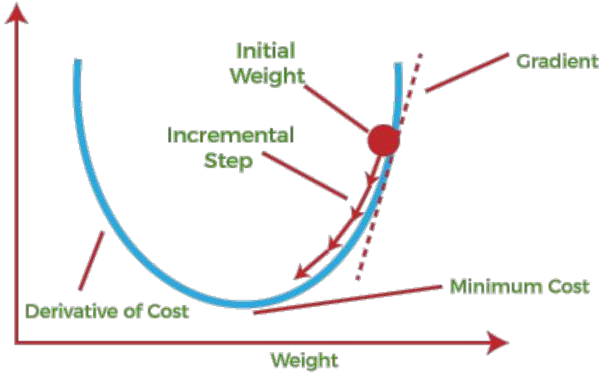
Distributed Representation

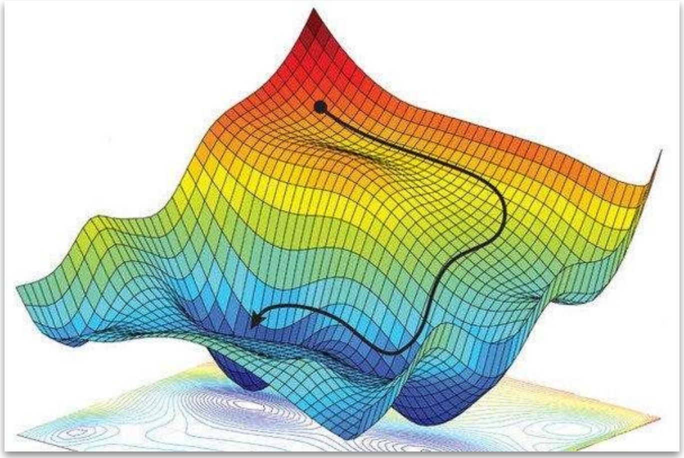
Photo-credit Bengio, Hinton, Le Cun "Deep Learning", NIPS Tutorial, 2015

The intuition of manifold learning

10

Stochastic Gradient Descent





- Find the minimum of a function by moving in the direction of the steepest descent. [1]
- Not guaranteed to work for non-convex problems.
- But: Large, multi-layer networks have many equivalent local minima [2]
- For over-parametrized n/w SGD can find global minima in polynomial time. [3]

[1] Rumelhart, Hinton, Williams “Learning Representations by back-propagating errors”, Nature, 1986
[2] Choromanska et al “The Loss Surfaces of Multilayer Networks, AISTATS 2015
[3] Allen-Zhu et al “A convergencc theory for deep learning via over-parametrization”, ICML 2019

11

RelationalAI

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

In computer chess, the methods that defeated the world champion, Kasparov, in 1997, were based on massive, deep search. At the time, this was looked upon with dismay by the majority of computer-chess researchers who had pursued methods that leveraged human understanding of the special structure of chess. When a simpler, search-based approach with special hardware and software proved vastly more effective, these human-knowledge-based chess researchers were not good losers. They said that ``brute force" search may have won this time, but it was not a general strategy, and anyway it was not how people played chess. These researchers wanted methods based on human input to win and were disappointed when they did not.

A similar pattern of research progress was seen in computer Go, only delayed by a further 20 years. Enormous initial efforts went into avoiding search by taking advantage of human knowledge, or of the special features of the game, but all those efforts proved irrelevant, or worse, once search was applied effectively at scale. Also important was the use of learning by self play to learn a value function (as it was in many other games and even in chess, although learning did not play a big role in the 1997 program that first beat a world champion). Learning by self play, and learning in general, is like search in that it enables massive computation to be brought to bear. Search and learning are the two most important classes of techniques for utilizing massive amounts of computation in AI research. In computer Go, as in computer chess, researchers' initial effort was directed towards utilizing human understanding (so that less search was needed) and only much later was much greater success had by embracing search and learning.

In speech recognition, there was an early competition, sponsored by DARPA, in the 1970s. Entrants included a host of special methods that took advantage of human knowledge---knowledge of words, of phonemes, of the human vocal tract, etc. On the other side were newer methods that were more statistical in nature and did much more computation, based on hidden Markov models (HMMs). Again, the statistical methods won out over the human-knowledge-based methods. This led to a major change in all of natural language processing, gradually over decades, where statistics and computation came to dominate the field. The recent rise of deep learning in speech recognition is the most recent step in this consistent direction. Deep learning methods rely even less on human knowledge, and use even more computation, together with learning on huge training sets, to produce dramatically better speech recognition systems. As in the games, researchers always tried to make systems that worked the way the researchers thought their own minds worked---they tried to put that knowledge in their systems---but it proved ultimately counterproductive, and a colossal waste of researcher's time, when, through Moore's law, massive computation became available and a means was found to put it to good use.

In computer vision, there has been a similar pattern. Early methods conceived of vision as searching for edges, or generalized cylinders, or in terms of SIFT features. But today all this is discarded. Modern deep-learning neural networks use only the notions of convolution and certain kinds of invariances, and perform much better.

This is a big lesson. As a field, we still have not thoroughly learned it, as we are continuing to make the same kind of mistakes. To see this, and to effectively resist it, we have to understand the appeal of these mistakes. We have to learn the bitter lesson that building in how we think we think does not work in the long run. The bitter lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning. The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favored, human-centric approach.

One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great. The two methods that seem to scale arbitrarily in this way are *search* and *learning*.

The second general point to be learned from the bitter lesson is that the actual contents of minds are tremendously, irredeemably complex; we should stop trying to find simple ways to think about the contents of minds, such as simple ways to think about space, objects, multiple agents, or symmetries. All these are part of the arbitrary, intrinsically-complex, outside world. They are not what should be built in, as their complexity is endless; instead we should build in only the meta-methods that can find and capture this arbitrary complexity. Essential to these methods is that they can find good approximations, but the search for them should be by our methods, not by us. We want AI agents that can discover like we can, not which contain what we have discovered. Building in our discoveries only makes it harder to see how the discovering process can be done.

general methods that leverage computation are ultimately the most effective, and by a large margin.

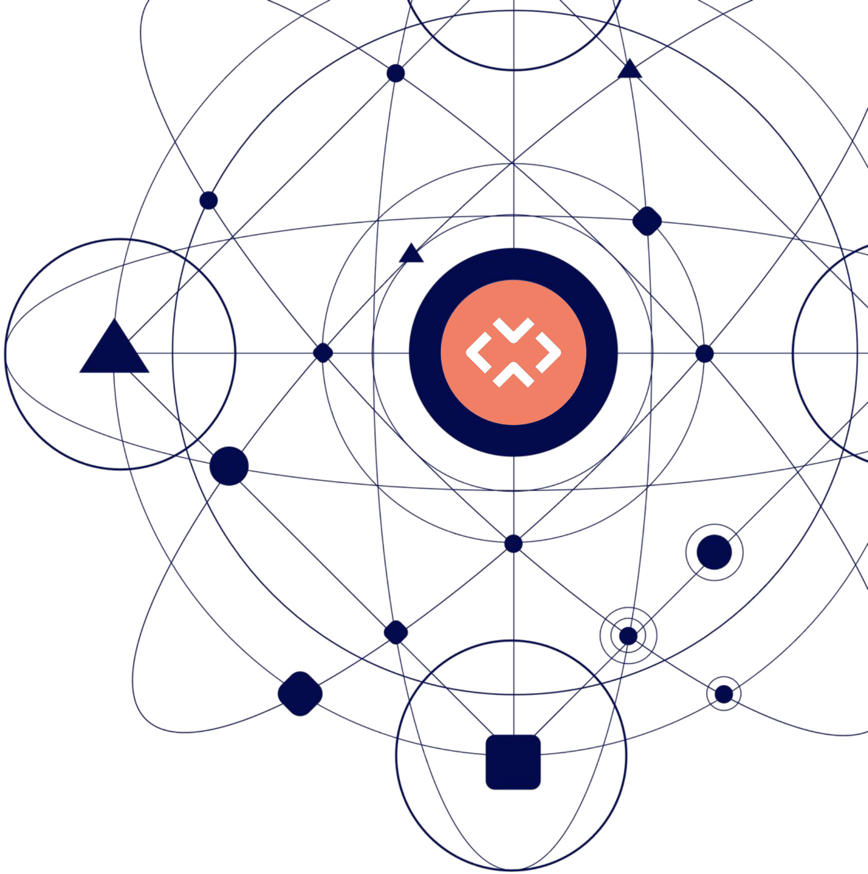
We have to learn the bitter lesson that building in how we think we think does not work in the long run.

… the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great.

The Bitter Lesson

RelationalAI

Language as the wormhole to AI



Traditional Computational Linguistics

Bill NP $(\uparrow \textsc{pred}) = \text{`}\textsc{Bill}\text{'}$
$\uparrow_\sigma = \mathit{Bill}$

kissed V $(\uparrow \textsc{pred}) = \text{`}\textsc{kiss}\text{'}$
$\forall X, Y.\ \mathit{agent}((\uparrow \textsc{pred})_\sigma, X) \otimes \mathit{theme}((\uparrow \textsc{pred})_\sigma, Y) \multimap \uparrow_\sigma = \mathit{kiss}(X, Y)$

Hillary NP $(\uparrow \textsc{pred}) = \text{`}\textsc{Hillary}\text{'}$
$\uparrow_\sigma = \mathit{Hillary}$

Figure 1: Lexical entries for *Bill*, *kissed*, *Hillary*

(1) $!(\forall f, X, Y.\ ((f\ \textsc{subj})_\sigma = X) \otimes ((f\ \textsc{obj})_\sigma = Y) \multimap \mathit{agent}((f\ \textsc{pred})_\sigma, X) \otimes \mathit{theme}((f\ \textsc{pred})_\sigma, Y))$

(2) $!(\forall f, X, Y, Z.\ ((f\ \textsc{subj})_\sigma = X) \otimes ((f\ \textsc{obj})_\sigma = Y) \otimes ((f\ \textsc{obj2})_\sigma = Z) \multimap$
$\mathit{permitter}((f\ \textsc{pred})_\sigma, X) \otimes \mathit{agent}((f\ \textsc{pred})_\sigma, Z) \otimes \mathit{theme}((f\ \textsc{pred})_\sigma, Y))$

Figure 2: Argument mapping principles

bill: $(f_{2\sigma} = \mathit{Bill})$
hillary : $(f_{3\sigma} = \mathit{Hillary})$
kiss : $(\forall X, Y.\ \mathit{agent}(f_{1\sigma}, X) \otimes \mathit{theme}(f_{1\sigma}, Y) \multimap f_{4\sigma} = \mathit{kiss}(X, Y))$
mapping1 : $(\forall X, Y.\ (f_{2\sigma} = X) \otimes (f_{3\sigma} = Y) \multimap \mathit{agent}(f_{1\sigma}, X) \otimes \mathit{theme}(f_{1\sigma}, Y)))$

$(\textbf{bill} \otimes \textbf{hillary} \otimes \textbf{kissed} \otimes \textbf{mapping1})$	(Premises.)
$\multimap \mathit{agent}(f_{1\sigma}, \mathit{Bill}) \otimes \mathit{theme}(f_{1\sigma}, \mathit{Hillary}) \otimes \textbf{kissed}$	(UI, Modus Ponens.)
$\multimap f_{4\sigma} = \mathit{kiss}(\mathit{Bill}, \mathit{Hillary})$	(UI, Modus Ponens.)

Figure 3: Derivation of *Bill kissed Hillary*

LFG Semantics via Constraints

Mary Dalrymple John Lamping Vijay Saraswat
{dalrymple, lamping, saraswat}@parc.xerox.com

Xerox PARC
3333 Coyote Hill Road
Palo Alto, CA 94304 USA

- Build out elaborate grammars for natural language, accounting manually for complex language phenomena (e.g. long-distance relationships)
- Use logic for meaning assembly

Now essentially obsolete for the NL practitioner

5 pillars for the Unreasonable Effectiveness of LLMs

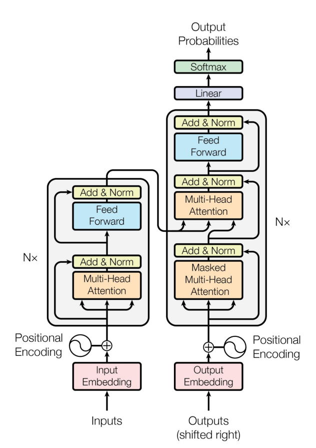
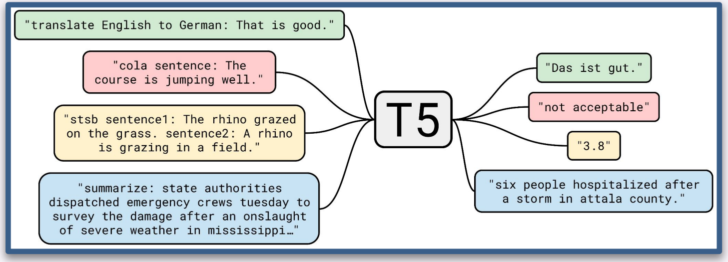


Figure 1: The Transformer - model architecture.

1. Transformer architecture

The cat on the hot tin ____ { roof *8 x 10^-1*
Timbuktu *10^-7*
pot *4 x 10^-5*
Iodine *6 x 10^-6*
bull-session *10^-9*
...

2. Causal Language Model for Self-supervision



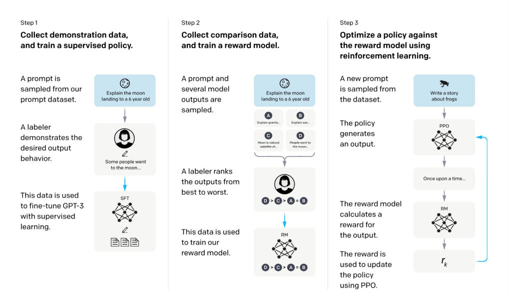
3. Single Task to Rule Them All

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}
GPT-3 Small	125M	12	768	12	64
GPT-3 Medium	350M	24	1024	16	64
GPT-3 Large	760M	24	1536	16	96
GPT-3 XL	1.3B	24	2048	24	128
GPT-3 2.7B	2.7B	32	2560	32	80
GPT-3 6.7B	6.7B	32	4096	32	128
GPT-3 13B	13.0B	40	5140	40	128
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128

Dataset	Quantity (tokens)	Weight in training mix
Common Crawl (filtered)	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

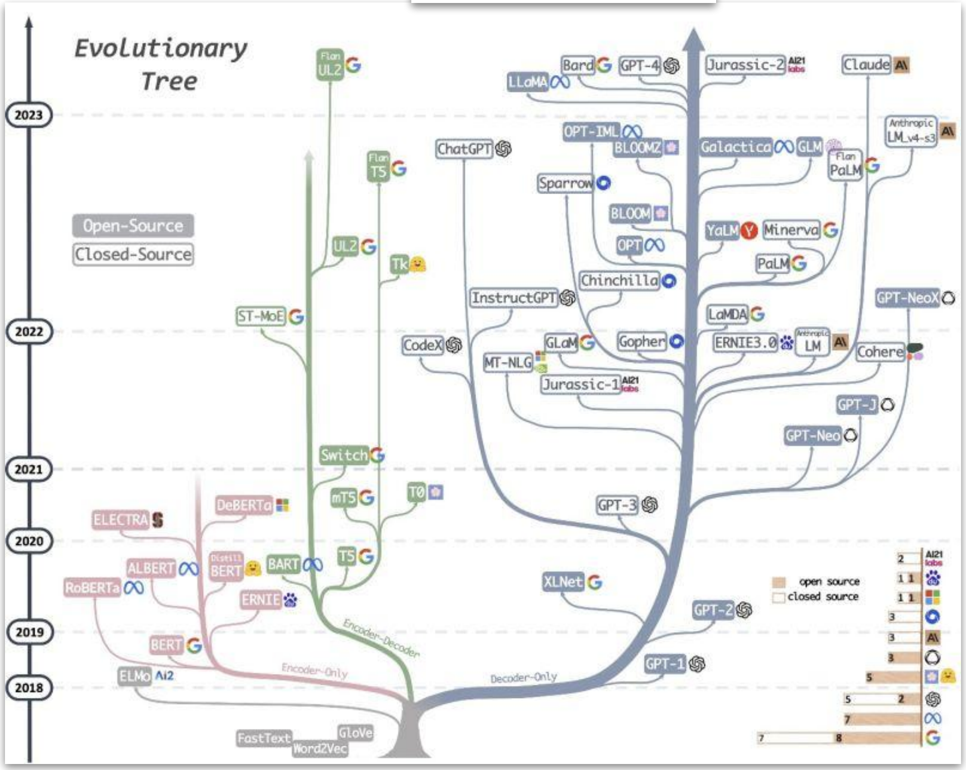
GPT-4: 1T parameters?

4. Scale, baby, scale!



5. Reinforcement Learning w Human Feedback

Explosion of LLMs



Slide from Yan Lecun

16

RelationalAI

GPT-4 Performance on Exams

Exam	GPT-4
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)
LSAT	163 (~88th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)
SAT Math	700 / 800 (~89th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)
USNCO Local Section Exam 2022	36 / 60
Medical Knowledge Self-Assessment Program	75 %
Codeforces Rating	392 (below 5th)
AP Art History	5 (86th - 100th)
AP Biology	5 (85th - 100th)
AP Calculus BC	4 (43rd - 59th)
AP Chemistry	4 (71st - 88th)
AP English Language and Composition	2 (14th - 44th)
AP English Literature and Composition	2 (8th - 22nd)

Exam	GPT-4
AP Environmental Science	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)
AP Microeconomics	5 (82nd - 100th)
AP Physics 2	4 (66th - 84th)
AP Psychology	5 (83rd - 100th)
AP Statistics	5 (85th - 100th)
AP US Government	5 (88th - 100th)
AP US History	5 (89th - 100th)
AP World History	4 (65th - 87th)
AMC 10[3]	30 / 150 (6th - 12th)
AMC 12[3]	60 / 150 (45th - 66th)
Introductory Sommelier (theory knowledge)	92 %
Certified Sommelier (theory knowledge)	86 %
Advanced Sommelier (theory knowledge)	77 %
Leetcode (easy)	31 / 41
Leetcode (medium)	21 / 80
Leetcode (hard)	3 / 45

17

RelationalAI

GPT4 “general purpose” knowledge and capabilities

Knowledge

Knowledge of real world and processes
- News
- Sports, Politics, Culture..
- Entertainment
- Naïve Physics
- …

Capabilities

(Localized) Question Answering
Self-evaluation
Self-correction
Elaboration
Ontology generation, application
Planning for a wide variety of tasks
Summarization
Criticism
Code generation (SQL, Datalog, Python, …)
Identifying trends
Synthetic Data Generation
…

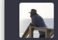
Identifying the range of GPT4 capabilities and the quality of its knowledge is the key intellectual task in front of us. Requires extensive, organized, thoughtful experimentation. (Some initial results follow.)

RelationalAI

Working with unstructured data

GPT4 as an Ontologist’s Assistant

Answering Ontology Questions (I/III)



Consider the following ontology, the direct relationships of which are expressed as a CSV file.

Assume standard meaning for relationships in ontologies, e.g. SubclassOf is transitive. Here is the full CSV file.

1,SubclassOf,organism,entity
2,SubclassOf,plant,organism
3,SubclassOf,hemlock,plant
4,SubclassOf,tree,plant
5,SubclassOf,seaweed,plant
6,SubclassOf,sequoia tree,tree
7,SubclassOf,animal,organism
8,SubclassOf,cat,animal
9,SubclassOf,sea animal,animal
10,SubclassOf,mammal,animal
11,SubclassOf,killer whale,mammal
12,SubclassOf,killer whale,sea animal
13,SubclassOf,dolphin,mammal
14,SubclassOf,dolphin,sea animal
15,SubclassOf,pufferfish,sea animal
16,SubclassOf,dog,mammal
17,SubclassOf,performing animal,animal
18,InstanceOf,Puff,cat
19,InstanceOf,Mase,cat
20,InstanceOf,Springer,killer whale
21,InstanceOf,Luna the Whale,killer whale
22,InstanceOf,Keiko,killer whale
23,InstanceOf,Keiko,performing animal
24,InstanceOf,Rin Tin Tin,dog
25,InstanceOf,Lassie,animal
26,InstanceOf,Rin Tin Tin,dog
27,InstanceOf,Hachiko,dog
28,HasAttribute,dog,four-legged
29,HasAttribute,Duncan,loyal
30,InstanceOf,Duncan,performing animal
31,HasAttribute,poison ivy,poisonous
32,HasAttribute,pufferfish,poisonous
33,HasAttribute,Goforth,accident-prone
34,HasAttribute,sea animal,aquatic
35,HasAttribute,mammal,warm-blooded
36,HasAttribute,animal,gender
37,InstanceOf,Granny (J2),orca
38,HasGender,Granny (J2),female
39,InstanceOf,Humphrey the Humpback Whale,whale
40,InstanceOf,Migaloo,whale
41,InstanceOf,Duncan,beagle
42,SubclassOf,beagle,dog
43,InstaneOf,miasma,gas

Please answer questions about this ontology. Explain your reasoning. Use only the information in the ontology. If the question cannot be answered based on the ontology, say IDK and explain.

A. Is poison ivy poisonous?
B. Is Rin Tin Tin four-legged?
C. Is Humphrey the Humpback Whale four-legged?
D. Is Mase an animal?
E. Is Rin Tin Tin a pet?
F. Is Duncan a dog?
G. Is pufferfish a kind of mammal?
H. Is miasma an entity?

20

Answering Ontology Questions (II/III)

A. Yes, poison ivy is poisonous. The ontology has the following information: (31, HasAttribute, poison ivy, poisonous).

B. It can be inferred that Rin Tin Tin is four-legged. Rin Tin Tin is an instance of a dog (24, InstanceOf, Rin Tin Tin, dog) and dogs have the attribute of being four-legged (28, HasAttribute, dog, four-legged).

C. IDK. The ontology does not provide information on whether Humphrey the Humpback Whale is four-legged or not.

D. Yes, Mase is an animal. Mase is an instance of a cat (19, InstanceOf, Mase, cat), and a cat is a subclass of an animal (8, SubclassOf, cat, animal).

E. IDK. The ontology does not provide information on whether Rin Tin Tin is a pet or not.

F. Yes, Duncan is a dog. Duncan is an instance of a beagle (41, InstanceOf, Duncan, beagle), and a beagle is a subclass of a dog (42, SubclassOf, beagle, dog).

G. No, a pufferfish is not a kind of mammal. The ontology states that pufferfish is a subclass of sea animal (15, SubclassOf, pufferfish, sea animal), and sea animals are a subclass of animals (9, SubclassOf, sea animal, animal). Mammals are also a subclass of animals (10, SubclassOf, mammal, animal), but there is no direct or indirect relationship between pufferfish and mammals in the ontology.

H. IDK. The ontology contains a typo in the line "43, InstaneOf, miasma, gas". Assuming it should be "43, InstanceOf, miasma, gas", we can infer that miasma is a gas, but we don't have any information about the relationship between gas and entity in the ontology.

21

WITH CWA (III/III)

Now make the Closed World Assumption. Answer only those questions whose answers would change.

Under the Closed World Assumption, we assume that any information not explicitly stated in the ontology is false. Based on this assumption, the answers to the following questions would change:

C. Is Humphrey the Humpback Whale four-legged?
Under the Closed World Assumption, since there is no information in the ontology about Humphrey the Humpback Whale being four-legged, we can assume that Humphrey the Humpback Whale is not four-legged.

E. Is Rin Tin Tin a pet?
Under the Closed World Assumption, since there is no information in the ontology about Rin Tin Tin being a pet, we can assume that Rin Tin Tin is not a pet.

H. Is miasma an entity?
Under the Closed World Assumption, since there is no information in the ontology about the relationship between gas (which miasma is an instance of, based on the corrected entry "43, InstanceOf, miasma, gas") and entity, we can assume that miasma is not an entity.

It can do that?!?

22

Constructing a KG from thin air

- Just Ask™!
  - Provide whatever directions are appropriate
- Review, ask for revisions as appropriate
  - Human in the loop mode
  - Semi-automated (ask and revise)
- With each new document processed, remember to ask – should the ontology be revised?

LLM as Collaborator

23

Generation of initial Ontology (I/IV)

Generate an ontology for financial news. It must be rich enough to represent a variety of events of interest in finance, including corporate events (mergers, acquisitions, divestitures), represent information content of earnings reports, 10-K, 10-Q filings, macro-economic information (e.g. interest rate hikes). Design it to be extensible. Emit it as a Datalog schema. -- Datalog Schema

```
; Financial News Ontology in Datalog
; Entities
entity(entity_id, entity_type, entity_name).
; Entity Types
entity_type(type_id, type_name).
; Financial News
news(news_id, headline, content, publication_date, author).
; News-Entity Relationship
news_entity(news_id, entity_id).
; Corporate Events
corporate_event(event_id, event_type, event_date, description).
; Corporate Event Types
event_type(type_id, type_name).
; Earnings Reports
earnings_report(report_id, entity_id, period_end_date, revenue, net_income,
earnings_per_share, publication_date).
; SEC Filings
sec_filing(filing_id, entity_id, filing_type, filing_date, document_url).
; Filing Types
filing_type(type_id, type_name).
; Macro-economic Events
macro_event(event_id, event_type, event_date, description).
; Macro-economic Event Types
macro_event_type(type_id, type_name).
; Interest Rate
interest_rate(rate_id, central_bank, rate, announcement_date).
; Datalog Facts
).
```

```
; Entity Types
entity_type(1, 'company').
entity_type(2, 'central_bank').
entity_type(3, 'government').
entity_type(4, 'currency').
; Event Types
event_type(1, 'merger').
event_type(2, 'acquisition').
event_type(3, 'divestiture').
event_type(4, 'spinoff').
event_type(5, 'bankruptcy').
 event_type(6, 'interest_rate_hike').
event_type(7, 'interest_rate_cut').
event_type(8, 'quantitative_easing').
event_type(9, 'quantitative_tightening'
; Filing Types
filing_type(1, '10-K').
filing_type(2, '10-Q').
; Example Fact
entity(1, 1, 'Acme Corporation').
news(1, 'Acme Corporation Reports Q3 Earnings', 'Acme Corporation reported Q3 earnings...',
'2023-04-24', 'Jane Doe').
news_entity(1, 1). earnings_report(1, 1, '2023-09-30', 1000000, 200000, 0.5, '2023-04-24').
sec_filing(1, 1, 2, '2023-04-24',
'https://www.sec.gov/Archives/edgar/data/000123456/000123456-23-000001-index.htm').
```


24

This uses new fact types! (II/IV)

Extract facts according to this ontology from the following article:
-- News article Coca-Cola (NYSE:KO) reported organic sales rose 12.0% in Q1 to beat the consensus expectation for a gain of 9.6%, driven higher by strong pricing and a positive contribution from mix. Organic revenue was up 23% for the Europe, Middle East & Africa segment and shot up 19% for the Latin America segment. Organic sales were 9% higher for the North America segment. Global unit case volume rose 3% during the quarter, while price and mix both helped to push revenue higher. Volume performance was driven by strength in away-from-home channels and continued investments in the marketplace. Developed markets grew mid single digits, while developing and emerging markets grew low single digits. Comparable operating margin improved to 31.8% of sales from 31.4% a year ago. Comparable operating margin (non-GAAP) expansion was primarily driven by strong topline growth and the impact of refranchising bottling operations, partially offset by an increase in marketing investments and higher operating costs versus the prior year as well as currency headwinds. Cash flow from operations was $160M, a decline of approxir
working capital initiatives an
approximately $520M versus
Of note, Coca-Cola (KO) said
harness the power of ChatGP
to build capabilities through
Shares of Coca-Cola (KO) rose
-- Ontology facts

```
; Entities
entity(1, 1, 'Coca-Cola').
entity(2, 1, 'OpenAI').
entity(3, 1, 'Bain & Company').
; Financial News
news(1, 'Coca-Cola Reports Q1 Organic Sales Growth of 12%', 'Coca-Cola
(NYSE:KO) reported organic sales rose 12.0% in Q1...', '2023-04-24', 'Jane Doe')
; News-Entity Relationship
news_entity(1, 1).
news_entity(1, 2).
news_entity(1, 3).
; Earnings Report Information
earnings_report(1, 1, '2023-03-31', null, null, null, '2023-04-24').
; Quarterly Report Information
; Organic Sales Growth
organic_sales_growth(1, 12.0).
; Segment Revenue segment_revenue(1, 'Europe, Middle East & Africa', 23.0).
segment_revenue(1, 'Latin America', 19.0). segment_revenue(1, 'North America',
9.0).
; Global Unit Case Volume global_unit_case_volume(1, 3.0).
```

```
; Segment Revenue
segment_revenue(1, 'Europe, Middle East & Africa', 23.0).
segment_revenue(1, 'Latin America', 19.0).
segment_revenue(1, 'North America', 9.0).
; Global Unit Case Volume
global_unit_case_volume(1, 3.0).
; Comparable Operating Margin
comparable_operating_margin(1, 31.8).
; Cash Flow from Operations
cash_flow_from_operations(1, 160000000).
; Free Cash Flow
 free_cash_flow(1, -120000000).
; Stock Information
stock_information(1, 'Coca-Cola', 'KO', 1.47, 64.99, 'premarket').
; Partnership Information
partnership(1, 'Coca-Cola', 'OpenAI', 'ChatGPT and DALL-E integration
for marketing and operations').
partnership(1, 'Coca-Cola', 'Bain & Company', 'ChatGPT and DALL-E
integration for marketing and operations').
```


25

It automatically revises the ontology … (III/IV)

How do various facts -- organic_sales_growth, segment_revenue, global_unit_case_volume, comparable_operating_margin etc -- fit into your ontology?	So please give me the new, revised ontology
In the given ontology, these facts can be represented as separate predicates to store specific information related to the financial performance of a company. Here's how these facts can be incorporated into the existing ontology: ; Financial Performance Facts organic_sales_growth(report_id, percentage_growth). segment_revenue(report_id, segment_name, percentage_growth). global_unit_case_volume(report_id, percentage_growth). comparable_operating_margin(report_id, margin_percentage). cash_flow_from_operations(report_id, cash_flow_amount). free_cash_flow(report_id, cash_flow_amount). partnership(partnership_id, entity1_id, entity2_id, partnership_description). stock_information(news_id, entity_name, ticker_symbol, percentage_change, price, market_status). The predicates added to the ontology store the specific financial performance metrics associated with a company's earnings report or financial news. Each predicate connects to the existing earnings report or news facts by using the "report_id" or "news_id" as a foreign key. For example, the "organic_sales_growth" predicate connects to the "earnings_report" by having the same "report_id" as a primary key in the "earnings_report" predicate. This connection allows the ontology to represent facts about a company's organic sales growth in relation to its earnings report. Similarly, other predicates like "segment_revenue", "global_unit_case_volume", and "comparable_operating_margin" store facts related to the financial performance of a company in specific segments or aspects and connect to the "earnings_report" or "news" facts using the corresponding "report_id" or "news_id".	; Financial News Ontology in Datalog (Revised) ; Entities entity(entity_id, entity_type, entity_name). ; Entity Types entity_type(type_id, type_name). ; Financial News news(news_id, headline, content, publication_date, author). ; News-Entity Relationship news_entity(news_id, entity_id). ; Corporate Events corporate_event(event_id, event_type, event_date, description). ; Corporate Event Types event_type(type_id, type_name). ; Earnings Reports earnings_report(report_id, entity_id, period_end_date, revenue, net_income, earnings_per_share, publication_date). ; SEC Filings sec_filing(filing_id, entity_id, filing_type, filing_date, document_url). ; Filing Types filing_type(type_id, type_name). ; Macro-economic Events macro_event(event_id, event_type, event_date, description). ; Macro-economic Event Types macro_event_type(type_id, type_name). ; Interest Rate interest_rate(rate_id, central_bank, rate, announcement_date).

26

Extracts facts according to new ontology (IV/IV)

OK, now extract facts from the Coca-Cola news report according to this revised ontology. The predicate for every fact you extract must already be in this revised ontology.

```
; Financial Performance Facts
organic_sales_growth(report_id, percentage_growth).
segment_revenue(report_id, segment_name, percentage_growth).
global_unit_case_volume(report_id, percentage_growth).
comparable_operating_margin(report_id, margin_percentage).
cash_flow_from_operations(report_id, cash_flow_amount).
free_cash_flow(report_id, cash_flow_amount).
partnership(partnership_id, entity1_id, entity2_id, partnership_description).
stock_information(news_id, entity_name, ticker_symbol, percentage_change, price,
market_status).

; Datalog Facts
; Entity Types
entity_type(1, 'company').
entity_type(2, 'central_bank').
entity_type(3, 'government').
entity_type(4, 'currency').
; Event Types
event_type(1, 'merger').
event_type(2, 'acquisition').
event_type(3, 'divestiture').
event_type(4, 'spinoff').
event_type(5, 'bankruptcy').
event_type(6, 'interest_rate_hike').
event_type(7, 'interest_rate_cut').
event_type(8, 'quantitative_easing').
event_type(9, 'quantitative_tightening').
; Filing Types
….
```

---- Bunch of Datalog rules follow ---

```
; Entities
entity(1, 1, 'Coca-Cola').
entity(2, 1, 'OpenAI').
entity(3, 1, 'Bain & Company').
; Financial News
news(1, 'Coca-Cola Reports Q1 Organic Sales Growth of 12%', 'Coca-Cola (NYSE:KO)
reported organic sales rose 12.0% in Q1...', '2023-04-24', 'Jane Doe').
; News-Entity Relationship
news_entity(1, 1). news_entity(1, 2). news_entity(1, 3).
; Earnings Report Information
earnings_report(1, 1, '2023-03-31', null, null, null, '2023-04-24').
; Organic Sales Growth
organic_sales_growth(1, 12.0).
; Segment Revenue
segment_revenue(1, 'Europe, Middle East & Africa', 23.0).
segment_revenue(1, 'Latin America', 19.0).
segment_revenue(1, 'North America', 9.0).
; Global Unit Case Volume
 global_unit_case_volume(1, 3.0).
; Comparable Operating Margin
comparable_operating_margin(1, 31.8).
; Cash Flow from Operations
cash_flow_from_operations(1, 160000000).
; Free Cash Flow
free_cash_flow(1, -120000000).
; Partnership Information
partnership(1, 1, 2, 'Integration of ChatGPT and DALL-E for marketing and operations').
…..
```

---- Additional explanation follows ---

Very useful approach to bootstrap practical ontologies and extraction systems

Generative Knowledge Tasks

I am a financial analyst tasked with generating economically plausible scenarios that can result from a "shock" to the economic system. A scenario should provide a coherent account of key economic macro-variables, such as changes in rate of inflation, interest rates in the US, Europe and Russia , exchange rate between USD/EUR, USD/Russian rouble, GDP growth, the price of oil.

Consider the ongoing war in Ukraine. Suppose Russia uses a tactical nuclear weapon on the battlefield against Ukraine. Generate a plausible set of scenarios. For each describe why they are plausible and provide your estimation of the changes in key economic variables.

Scenario 1: Escalation of conflict and global economic downturn
Plausibility: High

In this scenario, the use of a tactical nuclear weapon by Russia against Ukraine leads to a significant escalation of the conflict, with NATO and other countries increasing their support for Ukraine. This results in a prolonged and more intense war, leading to a global economic downturn.

Key Economic Variables:

1. Inflation: A surge in global inflation due to supply chain disruptions and increasing energy prices.
2. Interest Rates: Central banks in the US and Europe may initially lower interest rates to support economic growth but may eventually raise them to combat inflation.
3. USD/EUR: The US dollar could appreciate against the Euro, as investors seek safe-haven assets.
4. USD/Russian Rouble: The Russian Rouble would likely depreciate significantly against the US dollar due to economic sanctions and capital outflows.
5. GDP Growth: A slowdown in global GDP growth due to the impact of the war on trade, investment, and consumer confidence.
6. Price of Oil: A sharp increase in oil prices due to supply disruptions and geopolitical tensions.

Scenario 2: Swift resolution and economic recovery
Plausibility: Low

In this scenario, the use of a tactical nuclear weapon by Russia results in swift international intervention, leading to a resolution of the conflict. This allows the global economy to recover relatively quickly from the initial shock.

Key Economic Variables:

1. Inflation: A temporary spike in inflation due to supply chain disruptions, followed by a return to normal levels as the conflict resolves.
2. Interest Rates: Central banks in the US and Europe may lower interest rates temporarily but will normalize them as the economy recovers.
3. USD/EUR: The US dollar could appreciate against the Euro initially, but eventually return to pre-crisis levels.
4. USD/Russian Rouble: The Russian Rouble would initially depreciate significantly but could recover as the conflict is resolved and sanctions are lifted.
5. GDP Growth: A temporary dip in global GDP growth, followed by a swift recovery as the conflict resolves and economic activity resumes.
6. Price of Oil: A temporary increase in oil prices due to supply disruptions, followed by a return to normal levels as the conflict resolves and production resumes.

nged stalemate and economic stagnation
rate

he use of a tactical nuclear weapon by Russia leads to a prolonged
en the warring parties. This uncertainty creates a drag on the global
w growth and heightened risk aversion.

riables:

rate inflation, as supply chain disruptions and energy prices remain elevated but do not spike dramatically.
2. Interest Rates: Central banks in the US and Europe may maintain low interest rates to support economic growth.
3. USD/EUR: The US dollar could appreciate against the Euro, as investors seek safe-haven assets.
4. USD/Russian Rouble: The Russian Rouble would likely depreciate against the US dollar due to ongoing economic sanctions and capital outflows.
5. GDP Growth: Stagnant global GDP growth due to the ongoing conflict and its impact on trade, investment, and consumer confidence.
6. Price of Oil: Persistently elevated oil prices due to ongoing geopolitical tensions and supply disruptions.

Reasonable, though bland   Generable quickly, at scale

28

RelationalAI

Working with structured data

Using your KG with an LLM

Simple Relational Code-gen

Write a Datalog program for the mami(X,Y) relation:
X is the mami of Y if X is the spouse of Z, Z is a brother of W and W is the mother of Y.

You are given the relations parent(X,Y), male(X), female(X).
All other predicates used need to be defined.
Do not provide any commentary.
The program must not be recursive.
Do not define the predicate parent/2, male/1, female/1.

```
spouse(X, Y) :- parent(X, Z), parent(Y, Z), male(X), female(Y).
spouse(X, Y) :- parent(X, Z), parent(Y, Z), male(Y), female(X).

brother(X, Y) :- parent(Z, X), parent(Z, Y), male(X), X != Y.

mother(X, Y) :- parent(X, Y), female(X).

mami(X, Y) :- spouse(X, Z), brother(Z, W), mother(W, Y).
```

GPT-4 already understands how to produce (when asked) output in Python,JSON, Prolog, Javascript, …

30

Planning

I am in New York. I want to spend a day from 9:00am to 6pm. I want to start at an art museum, visit two more, have lunch in an Indian restaurant and visit two history museums, ending at one. I want to minimize the amount of time I spend walking.

Please help me by creating a Python program that I can run in Jupyter. The program should use the Google Map API, and pulp, the integer programming library. I will supply the Google MAP API key in the variable API_KEY.

The program should look like this:
1. It should have a list of 3 art museums in NYC.
2. It should have a list of 2 history museums in NYC.
3. It should have a list of 1 site that is an Indian restaurants in NYC.
4. It should compute the address for each site x, using the Google Map API.
6. It should use the Google MAP API to compute the distance[x,y] between each pair of distinct sites x and y.
7. It should use pulp to solve a traveling salesman problem. This problems should ensure that each of the selected sites (3 art museums, 2 history sites, 1 Indian restaurant) is visited exactly once. The length of the trip should be minimized.

Please give me the program. Make sure it is using pulp APIs correctly.

```
sites = art_museums + history_museums + indian_restaurant
site_coordinates = {}
for site in sites:
    geocode_result = gmaps.geocode(site)
    location = geocode_result[0]['geometry']['location']
    site_coordinates[site] = (location['lat'], location['lng'])
def compute_distance(site1, site2):
    result = gmaps.distance_matrix(site_coordinates[site1], site_coordinates[site2])
    distance = result['rows'][0]['elements'][0]['distance']['value']
    return distance
distances = {}
for site1, site2 in combinations(sites, 2):
    distances[frozenset([site1, site2])] = compute_distance(site1, site2)
problem = LpProblem("Traveling_Salesman", LpMinimize)
x = LpVariable.dicts("X", ((i, j) for i in sites for j in sites if i != j), cat='Binary')
problem += lpSum(distances[frozenset([i, j])] * x[(i, j)] for i in sites for j in sites if i != j)
for site in sites:
    problem += lpSum(x[(i, j)] for j in sites if i != j) == 1
    problem += lpSum(x[(j, i)] for j in sites if i != j) == 1
u = LpVariable.dicts("u", sites, lowBound=0, upBound=len(sites)-1, cat='Integer')
for i in sites:
n, site)].varValue == 1][0]
```

```
import googlemaps
from pulp import LpMinimize, LpProblem, LpStatus, lpSum, LpVariable
from itertools import combinations

API_KEY = 'YOUR_GOOGLE_MAP_API_KEY_HERE'

gmaps = googlemaps.Client(key=API_KEY)

art_museums = ['The Metropolitan Museum of Art, New York, NY',
            'Museum of Modern Art, New York, NY',
            'Solomon R. Guggenheim Museum, New York, NY']

history_museums = ['American Museum of Natural History, New York, NY',
            'New-York Historical Society, New York, NY']

indian_restaurant = ['Junoon, New York, NY']
```

Almost correct! Constraints need a few tweaks.

Very useful assistant for knowledgeable programmer.

31

RelationalAI

Planning

I am in New York. I want to spend a day from 9:00am to 6pm. I want to visit 3 sites related to art (museums, galleries, etc). I want to visit 2 sites related to history and 1 related to science. I want to also have lunch in an Indian restaurant. I want to leave the order in which I visit the sites open. But I dont want to travel more than 5 miles between sites.

In order to solve the problem you will need to write a python program that uses an integer programming library. Choose one that allows two symbolic variables to be multiplied, e.g. the pulp library. The program does the following
1. Create a python list with 10 sites related to art in NY
2. Create a python list with 10 sites related to history in NY
3. Create a list related to 10 Indian restaurants in NY
4. For all the sites and restaurants use the Google Map API to fetch their addresses
5. Each site (art or history) and restaurant is a node on a graph
6. Use Google Maps to compute the distance between each node
7. Use an integer programming library to solve an optimization problem. Your optimization variables are binary and each of them represents a node in the graph. If a node is selected as part of the itinerary then the variable gets the value 1 if not it gets the value 0. Your optimization problem should have the following constraints:
7a. The sum of the variables representing the art nodes should be 3.
7b. The sum of the variables representing the history nodes should be 2.
7c. The sum of the variables representing the restaurants should be 1.
7c. The maximum distance between any two chosen nodes should be less than 5 miles

Show me the Python code only.

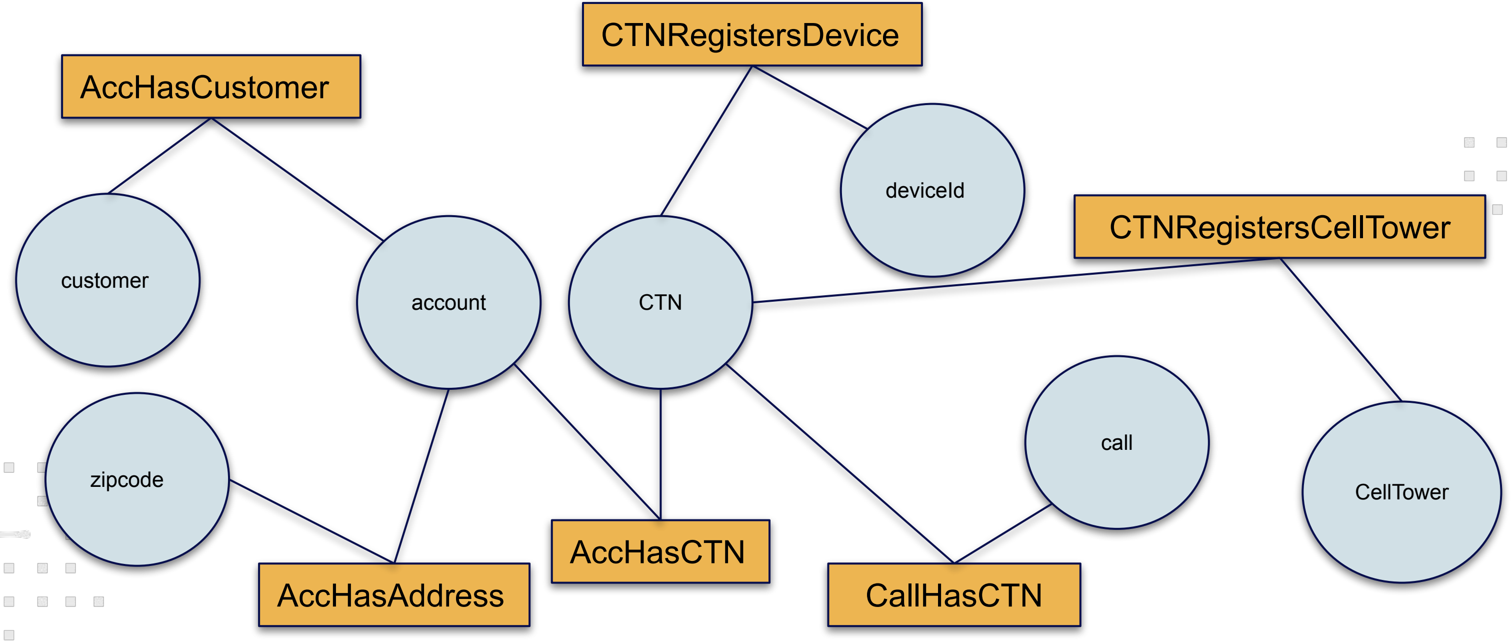
Providing a detailed outline to increase chance of success.

Note: you do not have to worry about remembering specific API calls.

32

Semantic Search

- Find the cell tower who receives most of the calls from customers who live in a specific address?



RelationalAI

ChatGPT 3.5 doesn't work for datalog

NV

Given this ontology in datalog:
AccHasCustomer(acc, cust)
CTNRegistersDeviceID(ctn, dev)
AccHasAddress(acc, zip)
AccHasCTN(acc, ctn)
CallHasCTN(call, ctn)
CTNRegistersCellTower(ctn, cell)
Write a query that answers that question:
Find the cell tower who receives most of the calls from customers who
live in a specific address?

To find the cell tower that receives the most calls from customers who live in a specific address, we need to first identify the customers who live at that address, and then find the calls associated with those customers, and finally determine the cell tower that received the most calls.

Here's the datalog query to achieve this:

css Copy code

```
maxcount(cell, count) :-
  AccHasAddress(acc, zip),
  AccHasCustomer(acc, cust),
  CTNRegistersDeviceID(ctn, dev),
  AccHasCTN(acc, ctn),
  CallHasCTN(call, ctn),
  CTNRegistersCellTower(ctn, cell),
  AccHasAddress(acc, desired_zip),
  desired_zip == zip,
  count = #count {call : CTNRegistersCellTower(ctn, cell)}.
```

In this query, we first specify the desired zip code in the variable `desired_zip`. We then find all the accounts that have this zip code using `AccHasAddress(acc, desired_zip)` and `AccHasCustomer(acc, cust)`. Next, we find all the CTNs associated with those accounts using `AccHasCTN(acc, ctn)`, and then all the calls associated with those CTNs using `CallHasCTN(call, ctn)`.

Finally, we count the number of calls associated with each cell tower using `count = #count {call : CTNRegistersCellTower(ctn, cell)}` and determine the cell tower with the highest count using the `maxcount` predicate.

To use this query for a specific address, replace `desired_zip` with the desired zip code. For example, if the desired zip code is 12345, the query would look like this:

css Copy code

```
maxcount(cell, count) :-
  AccHasAddress(acc, 12345),
  AccHasCustomer(acc, cust),
  CTNRegistersDeviceID(ctn, dev),
  AccHasCTN(acc, ctn),
  CallHasCTN(call, ctn),
  CTNRegistersCellTower(ctn, cell),
  count = #count {call : CTNRegistersCellTower(ctn, cell)}.
```


34

ChatGPT 4 Works

NV

Given this ontology in SQL tables:
AccHasCustomer(account, customer)
CTNRegistersDeviceID(ctn, deviceID)
AccHasAddress(account, zip)
AccHasCTN(account, ctn)
CallHasCTN(call, ctn)
CTNRegistersCellTower(ctn, celltower)
Write a SQL query that answers that question:
Find the cell tower who receives most of the calls from customers who live in a specific address?

To answer this question, we can use the following SQL query:

sql Copy code

```sql
SELECT celltower, COUNT(call) AS call_count
FROM CallHasCTN
JOIN CTNRegistersCellTower ON CallHasCTN.ctn = CTNRegistersCellTower.ctn
JOIN AccHasCTN ON CallHasCTN.ctn = AccHasCTN.ctn
JOIN AccHasAddress ON AccHasCTN.account = AccHasAddress.account
WHERE AccHasAddress.zip = 'your_specific_zip_code'
GROUP BY celltower
ORDER BY call_count DESC
LIMIT 1;
```


35

A different approach

NV

Given this ontology in the form (relation, entity, entity)
(AccHasCustomer, account, customer)
(CTNRegistersDeviceID, ctn, deviceID)
(AccHasAddress, account, zip)
(AccHasCTN, account, ctn)
(CallHasCTN, call, ctn)
(CTNRegistersCellTower, ctn, celltower)
Identify the entities involved in the following question:
Find the cellphone tower who receives most of the calls from people who live in a specific address?

Based on the given ontology and the question, the involved entities are:

1. zip (address)
2. celltower (cellphone tower)

In the given question, the following entities are involved:

1. Cellphone tower: This is the target entity to be identified in the question and is represented by 'celltower' in the ontology.
2. Calls: These are used to determine the cellphone tower that receives the most calls, represented by 'call' in the ontology.
3. Address: This is the specific address mentioned in the question, represented by 'zip' in the ontology.

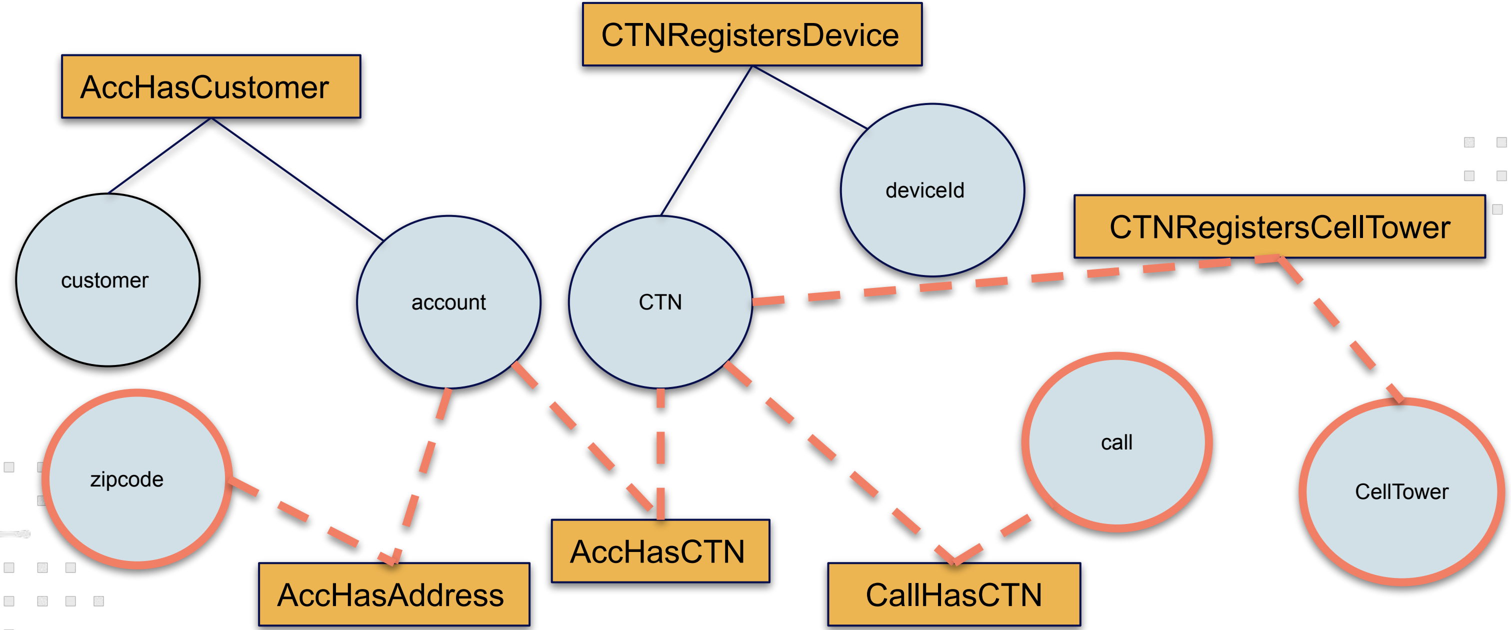
The entities involved in the question are:

1. zip (specific address)
2. call (calls from people)
3. ctn (cellular telephone number)
4. celltower (cellphone tower)

36

Traverse the Ontology

LLMs can help us anchor on the ontology. Then we can deterministically traverse it to form a query. As ontologies become larger this is the most viable solution.



RelationalAI

LLM Tradecraft

Recursion, subgoaling, literate agents

Prompting

- Early stages of LLMs:
  rispa => paris
  utefuac => faucet
  gboinr =>

- It was more about showing examples to implicitly describe the task
- LLMs proved to be very good analogical few shot learners
- Prompting was very brittle

From Prompting to instructions

- With 15 thousand conversations it was possible to move from brittle prompts to instructions
- ChatGPT (InstructGPT)
  - Treat the LLM as your colleague
  - Give instructions that describe the task.
  - ChatGPT knows enough and does not need an example
  - However it is possible that it might have no prior knowledge
    - Let’s say you want to teach it a new concept (ie a new programming language)

From GPT-3 to GPT-3.5 (ChatGPT)

40

Fine-tuning

- Develop a task-specific data-set (prompt, response)_n
- Fine-tune a pre-trained model.
  - For instance, a T5 model or GPT-3.
  - Works well for a number of tasks, e.g. classification
    - E.g. de-duplication of entity records from different vendors
    - E.g. labeling columns in relational tables with concepts from an ontology.
- Main drawback: Expensive to generate dataset
- But: Use a more powerful model to generate data for a smaller model (distillation), cf [Yejin Choi 2022](#) "Symbolic Knowledge Distillation"

41

RelationalAI

Symbolic knowledge distillation

Symbolic Knowledge Distillation:
from General Language Models to Commonsense Models

Peter West[†‡*] Chandra Bhagavatula[‡] Jack Hessel[‡] Jena D. Hwang[‡]
Liwei Jiang[†‡] Ronan Le Bras[‡] Ximing Lu[†‡] Sean Welleck[†‡] Yejin Choi [†‡*]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence

Abstract

The common practice for training commonsense models has gone *from–human–to–corpus–to–machine*: humans author commonsense knowledge graphs in order to train commonsense models. In this work, we investigate an alternative, *from–machine–to–corpus–to–machine*: general language models author these commonsense knowledge graphs to train commonsense models.

Our study leads to a new framework, **Symbolic Knowledge Distillation**. As with prior art in Knowledge Distillation (Hinton et al., 2015), our approach uses larger models to teach smaller models. A key difference is that we distill knowledge symbolically–as text–in addition to the resulting neural model. We distill only one aspect–the commonsense of a general language model teacher, allowing the student to be a different type of model, a commonsense model. Altogether, we show that careful prompt engineering and a separately trained

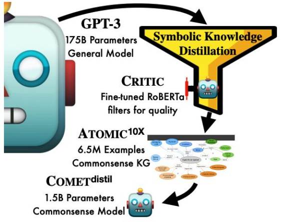


Figure 1: **Symbolic knowledge distillation** extracts the commonsense from the large, general language model GPT-3, into 2 forms: a large commonsense knowledge graph ATOMIC^{10x}, and a compact commonsense model COMET$^{DIS}_{TIL}$. The quality of this knowledge can be controlled and improved by adding a **critic** model, making GPT-3 a stronger teacher.

X starts running	**xEffect** *so, X*	gets in shape	X sings a song	**HinderedBy** *but not if*	X can't remember the lyrics
X and Y engage in an argument	**xWant** *so, X wants*	to avoid Y	X is not well liked	**xReact** *so, X feels*	lonely
X learns to type fast	**xNeed** *X needed*	to have taken typing lessons	X takes care of a monkey	**xAttr** *X is seen as*	kind
X steals his grandfather's sword	**xEffect** *so, X*	is punished by his grandfather	X butts in	**HinderedBy** *but not if*	X is too shy to speak up
X takes up new employment	**xIntent** *because X wants*	to be self sufficient	X waits for the storm to break	**xEffect** *so, X*	is safe from the storm

Figure 2: Example **automatically generated** ATOMIC triples from our ATOMIC^{10x} commonsense knowledge graph. Each example includes a generated **event**, **relation** (with natural language interpretation), and generated **inference**.

<TASK-PROMPT>
<EX$_1$-INP><EX$_1$-OUT>
...
<EX$_{N-1}$-INP><EX$_{N-1}$-OUT>
<EX$_N$-INP>

What needs to be true for this event to take place?

...

Event <i>: X goes jogging
Prerequisites: For this to happen, X needed to wear running shoes

...

Event <N>: X looks at flowers
Prerequisites: For this to happen,

1. Event: X overcomes evil with good
2. Event: X does not learn from Y
...
10. Event: X looks at flowers
11.

42

Embeddings

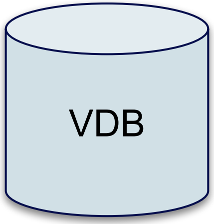
- The cornerstone of deep learning.
- It took many years to figure out how to represent anything in a vector with semantic meaning
- Embedding: Summarize text into a vector
- All modern LLM offer a service for converting text to embeddings
- Example:

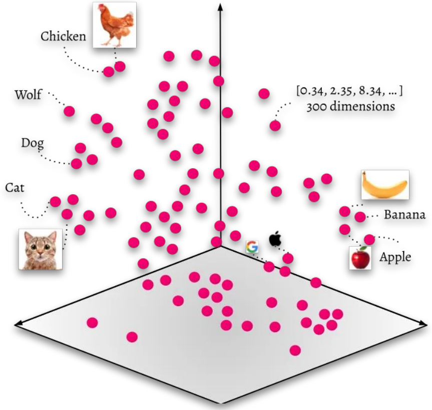
```
The kids enjoyed splashing water => Embedding1
My children love water parks => Embedding2

distance(Embedding1, Embedding2)  ~ small
```


43

Vector search

- A new type of database
- Key - value store where key is a vector
- Retrieval is done in an approximate fashion.
    - Given a vector find other vectors in the database





44

Vector Database Systems

FAISS
Scalable Search With Facebook AI

Pinecone

GSI TECHNOLOGY

haystack
by deepset

Qdrant

Jina

weaviate

vespa

Vald

milvus

Dmitry Kan
Oct 2, 2021 · 7 min read · Member-only · Listen

Not All Vector Databases Are Made Equal

A detailed comparison of Milvus, Pinecone, Vespa, Weaviate, Vald, GSI and Qdrant

While working on this blog post I had a privilege of interacting with all search engine key developers / leadership: Bob van Luijt and Etienne Dilocker (Weaviate), Greg Kogan (Pinecone), Pat Lasserre, George Williams (GSI Technologies Inc), Filip Haltmayer (Milvus), Jo Kristian Bergum (Vespa), Kiichiro Yukawa (Vald) and Andre Zayarni (Qdrant)

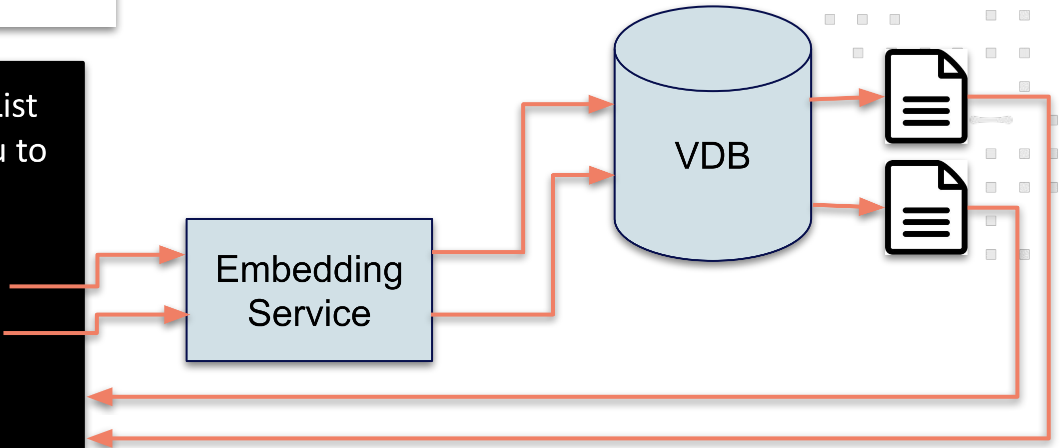
This blog has been discussed on HN: https://news.ycombinator.com/item?id=28727816

45

A vector database as external memory

- Split documents in passages
- Embed each doc with LLM
- Store (Embedding, passage) inside a VDB
- When the LLM has a question it can embed it and get relevant passages from the VDB

```
user> Is the debt of GE higher than GM? List
documents that would be relevant for you to
answer this question
gpt>
  -   Balance Sheet for General Motors
  -   Balance Sheet for General Electric
>
```



Iteratively augment prompt

- Most of the times you will need to work with prompting
- You can solve many problems by iteratively
  - Ask LLM a question
  - Retrieve information from an external VDB
  - Ask LLM to refine given the new information again
- Pros
  - You can dynamically change your Knowledge Base by adding and deleting records on your vector database
  - You have full control of your workflow
- Cons
  - Multiple calls to the LLM for reasoning queries
    - Higher costs
    - Higher latency

47

Multi-hop reasoning: Ask for residuals

```
[17]: print(f"Question: {hotpot[0]['question']}")
      a = solve_chat(hotpot[0],force=True)
      print(f"Answer: {hotpot[0]['answer']}")
      print(f"Result: {a}")
```

```
Question: Which magazine was started first Arthur's Magazine or First for Women?
----- 0 ------
[0 Search results] [7, 4].
[User]:
-- (Additional) Context Passages
First for Women: First for Women is a woman's magazine published by Bauer Media Group in the USA.  The magazine was started in
1989.  It is based in Englewood Cliffs, New Jersey.  In 2011 the circulation of the magazine was 1,310,696 copies.
First Arthur County Courthouse and Jail: The First Arthur County Courthouse and Jail, was perhaps the smallest court house in
the United States, and serves now as a museum.
--Question
Which magazine was started first Arthur's Magazine or First for Women?

[Assistant]:{"residual": "When was Arthur's Magazine started?"}

----- 1 ------
[1 Search results] [7, 4, 5, 9].
[User]:
-- (Additional) Context Passages
First for Women: First for Women is a woman's magazine published by Bauer Media Group in the USA.  The magazine was started in
1989.  It is based in Englewood Cliffs, New Jersey.  In 2011 the circulation of the magazine was 1,310,696 copies.
First Arthur County Courthouse and Jail: The First Arthur County Courthouse and Jail, was perhaps the smallest court house in
the United States, and serves now as a museum.Arthur's Magazine: Arthur's Magazine (1844–1846) was an American literary period
ical published in Philadelphia in the 19th century.  Edited by T.S. Arthur, it featured work by Edgar A. Poe, J.H. Ingraham, S
arah Josepha Hale, Thomas G. Spear, and others.  In May 1846 it was merged into "Godey's Lady's Book".
William Rast: William Rast is an American clothing line founded by Justin Timberlake and Trace Ayala.  It is most known for th
eir premium jeans.  On October 17, 2006, Justin Timberlake and Trace Ayala put on their first fashion show to launch their new
William Rast clothing line.  The label also produces other clothing items such as jackets and tops.  The company started first
as a denim line, later evolving into a men's and women's clothing line.
--Question
Which magazine was started first Arthur's Magazine or First for Women?

[Assistant]:{"ans": "Arthur's Magazine", "quote": "Arthur's Magazine (1844–1846)"}
Answer: Arthur's Magazine
Result: ({'ans': "Arthur's Magazine", 'quote': "Arthur's Magazine (1844–1846)"}, 1)
```

Setting: "Multi-hop" question answering – question Q can only be found by identifying multiple passages, which together contain info to answer Q. (Hotpot QA)

Use semantic embeddings to find passages close to question.

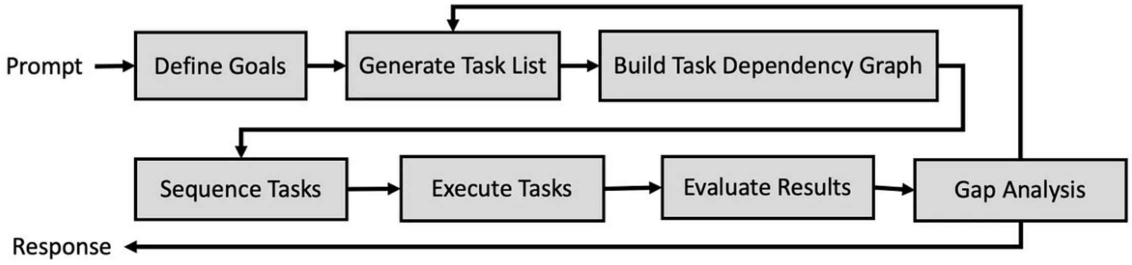
Idea: You can ask LM to answer Q, and if it cannot, **generate a question Q'**, which when answered will help answer Q.

Find more passages related to Q'.

Iterate

48

Language Machines



Use language for communication and representation.

- Organize problem-solving as
  - Interaction between a dynamic collection of agents
  - Using language for input, output and internal representation
  - Using language analyzers (LLMs) to understand input, generate updates to internal state, generate output, specify actions in the world
- Agents may
  - Operate autonomously (standing agents)
  - Generate and execute code
  - Spawn new agents

Powerful and natural extension of Language Models

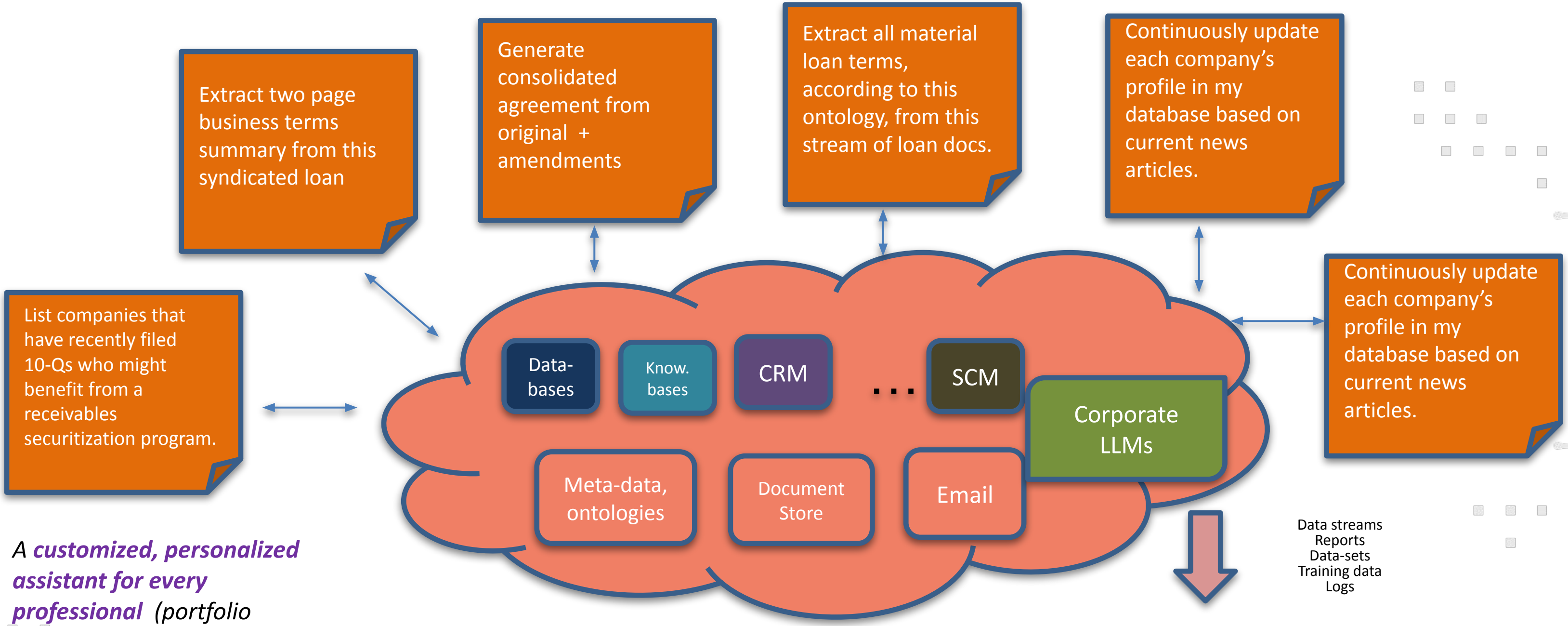
49

RelationalAI

Language Machines as the Knowledge Hub

Rethinking Enterprise Architecture

First Generation of Enterprise Computers for humans



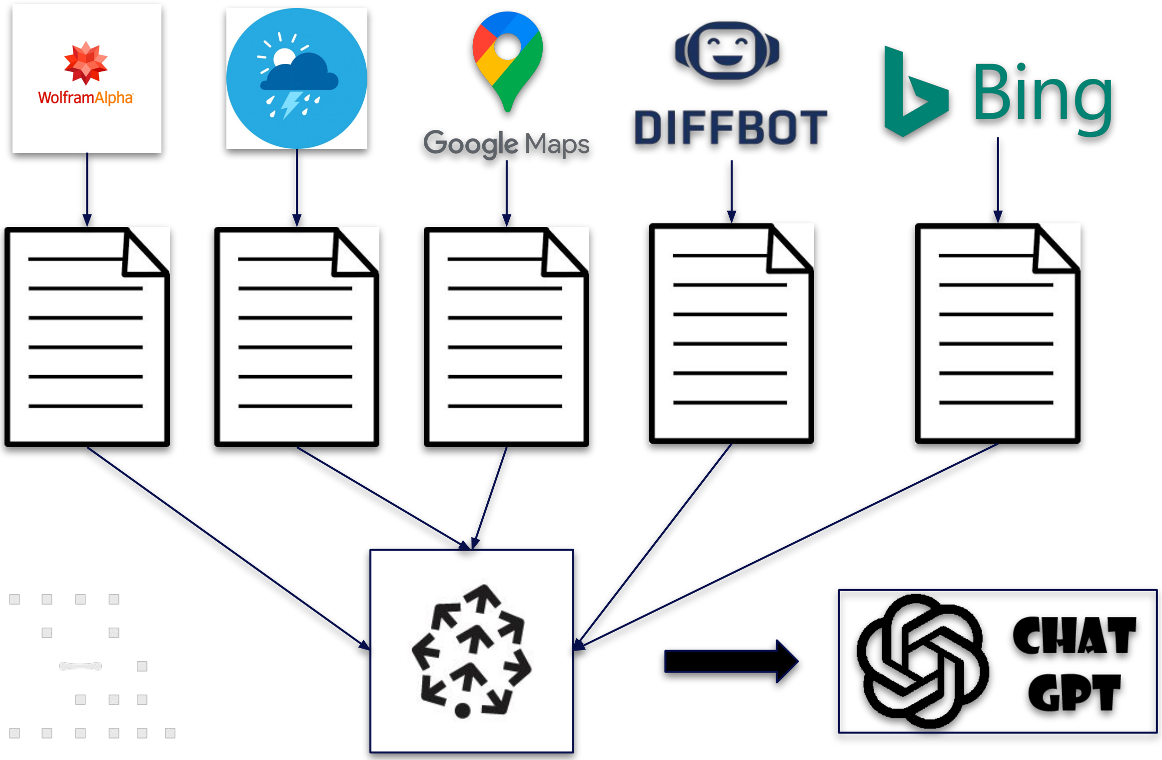
*A **customized, personalized assistant for every professional** (portfolio managers, traders, sales, risk, controllers, engineers, …)*

Language Machine Hub

Use data logs to retrain models with RLHF

Foundation of Institutional Memory

Implementing Language Machine Hub



- Describe APIs and entry-points using text.
- Register with LM Hub

RelationalAI

Example interaction

```
./kgchat -v 2
chat> :load-plugin googlemaps
module 'googlemaps' has no attribute 'plugin'
chat> :load-plugin maps
maps:directions 687 tokens
maps:distancematrix 629 tokens
maps:geolocate 588 tokens
chat> :load-plugin worldtime
worldtime:* 392 tokens
chat> What is the time right now in Faridabad, Haryana?
---- context segments (0.43s)
0.784568548 maps:distancematrix (629 tokens)
0.770304561 worldtime:* (392 tokens)
0.758450000 maps:directions (687 tokens)
0.708505094 maps:geolocate (588 tokens)
2332 tokens
---- context source (24.28s)
from worldtime import get_timezones_by_area_location

area = "Asia"
location = "Kolkata"  # Faridabad, Haryana is in the same timezone as Kolkata
datetime_timezone = get_timezones_by_area_location(area, location)

print(f"Current time in Faridabad, Haryana: {datetime_timezone.datetime}")
---- context (4.51s)
Current time in Faridabad, Haryana: 2023-04-19T23:28:24.720148+05:30
---- context prompt
49 tokens
---- (5.81s)
The current time in Faridabad, Haryana is 2023-04-19T23:28:24.720148+05:30.
----
```

```
chat> Ok that was totally awesome. But are you sure that Faridabad is in the same timezone as Calcutta? Why?
---- context segments (0.76s)
0.767981231 maps:directions (687 tokens)
0.767508447 worldtime:* (392 tokens)
0.743990302 maps:distancematrix (629 tokens)
0.740608811 maps:geolocate (588 tokens)
2332 tokens
---- context source (17.22s)
None
---- (16.05s)
Yes, I am sure that Faridabad and Kolkata (Calcutta) are in the same timezone. Both cities are located in India, and the entire country follows a single timezone, which is Indian Standard Time (IST). IST is 5 hours and 30 minutes ahead of Coordinated Universal Time (UTC+5:30). This uniform timezone helps in maintaining consistency in various aspects like business, communication, and transportation across the country.
----
chat>
```

User can ask questions about intermediate state!

53

RelationalAI

Example Interaction

Workflow

- A user asks a question Q
- Use LLM to generalize Q to Q’
- Embed Q’
- Look up relevant plugins R via VDB
- Use LLM to write the Python code P that uses R to produce data helpful in answering Q
- Use LLM to assemble resources into answer

“How long does it take to go from Acropolis to the parliament on foot? “

NV Generalize the following question by removing any specific information
How long does it take to go from Acropolis to the parliament on foot?

How long does it take to travel between two points of interest on foot?

NV Write the python code that answers the following question by using the google maps API

To use the Google Maps API to answer a question about travel time, you first need to set up a project and enable the Google Maps API. You can follow the instructions in the official documentation: https://developers.google.com/maps/documentation/directions/get-api-key

Once you have your API key, you can use the following Python code to find the travel time between two points of interest:

python Copy code

```python
import requests

def get_travel_time(origin, destination, api_key):
    base_url = "https://maps.googleapis.com/maps/api/directions/json?"

    # Prepare request parameters
    params = {
        "origin": origin,
        "destination": destination,
        "mode": "walking",
        "key": api_key
    }

    # Send request to Google Maps API
    response = requests.get(base_url, params=params)
    data = response.json()

    # Extract travel time information
    if data["status"] == "OK":
        route = data["routes"][0]
        leg = route["legs"][0]
        duration = leg["duration"]["text"]
        return duration
    else:
        return "Error: Could not retrieve travel time."

if __name__ == "__main__":
    origin = "Acropolis, Athens, Greece"
    destination = "Hellenic Parliament, Athens, Greece"
    api_key = "YOUR_API_KEY"

    travel_time = get_travel_time(origin, destination, api_key)
    print(f"Travel time on foot from {origin} to {destination}: {travel_time
```


54

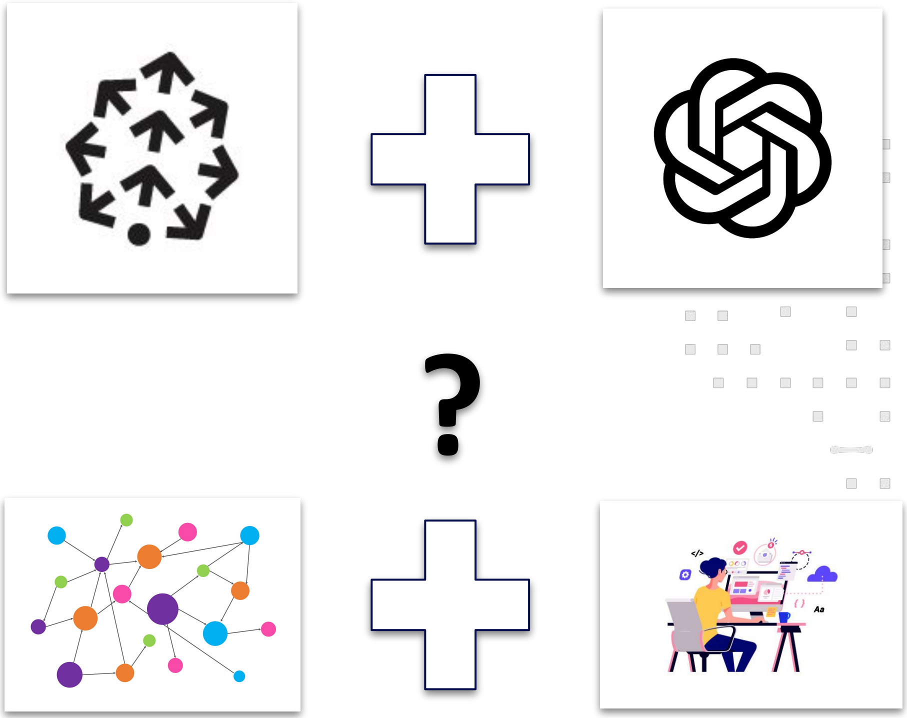
RelationalAI

The Elephant

What is the role of Knowledge Graphs in the new world?

Raw text vs Organized KG

- Should an enterprise convert unstructured data (documents) to a KG?
- Is it better to embed them into a vector database and query them with LLMs?
- Maybe finetune an LLM?



56

Do LLMs represent concepts better than KGs?

Knowledge Graphs abstract things from strings

```
// write query

def original_country = {
    "United States";
    "England";
    "Mexico";
    "France";
    "Iceland";
    "Russia";
    "Turkey"
}

def insert:country_name = {
  ^Country[name], name from name in original_country
}
```

```
// write query

def delete:country_name(e, name) = {
    country_name(e, name) and name = "Turkey"
}

def insert:country_name(e, "Türkiye") = {
    e = ^Country["Turkey"]
}
```

#	Hash	String
1	66962613565377794269208588323099458022	Iceland
2	70351310082327252121206171765925792648	Russia
3	74729024159465200830213900492854100902	Mexico
4	198593132371554076706272655681648002548	France
5	232831030175921651993477065381021933716	United States
6	243556965892236413755237816118635800139	Türkiye
7	313568593215643973623607804124061869773	England

#	Symbol	String	Mixed
1	:name	England	England
2	:name	France	France
3	:name	Iceland	Iceland
4	:name	Mexico	Mexico
5	:name	Russia	Russia
6	:name	Türkiye	Türkiye
7	:name	United States	United States
8	:name_alias	United States	USA
9	:name_alias	United States	United States of America

57

How do LLMs deal with concepts?

What was the population of Turkey in 1990?

The estimated population of Turkey in 1990 was approximately 56.5 million people.

what was the population of Türkiye in 1990?

The estimated population of Turkey in 1990 was approximately 56.5 million people.

GPT-4

The power of synonyms

58

A more difficult example

Who was the president of Rhodesia in 1970?

In 1970, Rhodesia was an unrecognized state, which had unilaterally declared independence from Britain in 1965. The country was led by Prime Minister Ian Smith, who had taken power after the declaration of independence. There was no president of Rhodesia at that time. Smith held executive power as the head of government, while the Queen of England remained the head of state, a position she held until 1979, when Rhodesia became Zimbabwe.

Who was the president of Zimbabwe in 1970?

In 1970, Zimbabwe was known as Rhodesia and it was not yet an independent country. At that time, the Prime Minister of Rhodesia was Ian Smith. He had declared Rhodesia's independence from Britain in 1965 and led the country until it became Zimbabwe in 1980.

The power goes beyond synonyms

GPT-4

59

RelationalAI

Another Example

NV

Who was the CEO of Truist in 2015?

In 2015, Truist Financial Corporation did not yet exist. Truist was formed in December 2019 as a result of the merger between BB&T (Branch Banking and Trust Company) and SunTrust Banks. During 2015, Kelly S. King was the CEO of BB&T, while William H. Rogers Jr. was the CEO of SunTrust Banks.

60

Sometimes things go wrong

NV what was the first animated movie?

The first animated feature film was "Snow White and the Seven Dwarfs," released in 1937. It was produced by Walt Disney Productions and directed by David Hand. The film was based on the fairy tale "Snow White" by the Brothers Grimm and was a critical and commercial success. It set the standard for future animated films and established Disney as a leader in the animation industry.

NV are you sure this is the correct answer?

I apologize for the mistake in my previous response. The first animated feature film was actually "El Apóstol," a silent Argentine film directed by Quirino Cristiani and released in 1917. The film was a political satire and used cutout animation techniques. I hope this clarifies your question.

NV what was the first animated movie?

The first animated feature film was "Snow White and the Seven Dwarfs," released in 1937. It was produced by Walt Disney Productions and directed by David Hand. The film was based on the fairy tale "Snow White" by the Brothers Grimm and was a critical and commercial success. It set the standard for future animated films and established Disney as a leader in the animation industry.

NV are you sure this is the correct answer?

2 / 2

I apologize for the mistake in my previous answer. While "Snow White and the Seven Dwarfs" was the first animated feature-length film produced in color and with sound, there were actually several animated short films that predate it. The first animated short film was "Fantasmagorie," directed by French animator Émile Cohl, which was released in 1908. Other early animated shorts include "Gertie the Dinosaur" (1914) by Winsor McCay and "Felix the Cat" (1919) by Pat Sullivan and Otto Messmer.

Was this response better or worse? Better Worse Same

People also ask

Who made the first animated film?

The first film-based animation was made by J. Stuart Blackton, whose Humorous Phases of Funny Faces in 1906 launched a successful series of animated films for New York's pioneering Vitagraph Company.

61

Converting text to KG

Pros

- Facts are reliable, since they are reviewed by humans
- KGs are faster and cheaper than LLMs
- Heavy lifting is done offline, applications are lighter
- KGs have more powerful reasoning capabilities

Cons

- You need to know a priori what entities/relations to extract
- In case of a new query type, you need to re-extract info from raw docs

62

Working directly with text and LLMs

Pros

- Documents can be added/removed dynamically
- Very powerful reasoning capabilities
- Natural Language for Querying
- Easy to do multihop querying

Cons

- LLMs have high latency
- Multiple LLM calls might be required
- LLMs have high cost
- Need an extra Vector DB
- Embedding docs has an associated cost

63

RelationalAI

Knowing what you don't know

- Knowledge Graphs know their boundaries really well
- We can't claim the same for LLMS

What is the distance between Anchorage and New York?

The distance between Anchorage, Alaska, and New York City, New York, is approximately 4,060 kilometers (2,520 miles) if you were to travel in a straight line between the two cities. However, actual travel distance between the two cities can vary depending on the mode of transportation and the route taken.

Anchorage, Alaska

New York

71 hr (4,357.6 mi) via Alaska Hwy

Directions

People also ask

How far is Alaska from New York by plane?

How far is Anchorage from New York by plane?

3363 miles

A direct flight to Anchorage will take on average 10h 08m to travel the 3363 miles flying distance from New York.

64

RelationalAI

LLMs controlling KGs

S

How far is it from Chicago to Tokyo?

How far is it from Tokyo to Chicago?

Used **Wolfram**

The distance from Chicago to Tokyo is approximately 6,313 miles (10,160 kilometers). If you were to travel by aircraft at a speed of 550 mph, it would take about 11 hours and 30 minutes to cover this distance.

Toolformer: Language Models Can Teach Themselves to Use Tools

Timo Schick **Jane Dwivedi-Yu** **Roberto Dessì**[†] **Roberta Raileanu**
Maria Lomeli **Luke Zettlemoyer** **Nicola Cancedda** **Thomas Scialom**
Meta AI Research [†]Universitat Pompeu Fabra

Abstract

Language models (LMs) exhibit remarkable abilities to solve new tasks from just a few examples or textual instructions, especially at scale. They also, paradoxically, struggle with basic functionality, such as arithmetic or factual lookup, where much simpler and smaller models excel. In this paper, we show that LMs can teach themselves to *use external tools* via simple APIs and achieve the best of both worlds. We introduce *Toolformer*, a model trained to decide which APIs to call, when to call them, what arguments to pass, and how to best incorporate the results into future token prediction. This is done in a self-supervised way, requiring nothing more than a handful of demonstrations for each API. We incorporate a range of tools, including a calculator, a Q&A

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

LLMs still struggle to know when to use KGs

LLMs embedded in the KG query plan

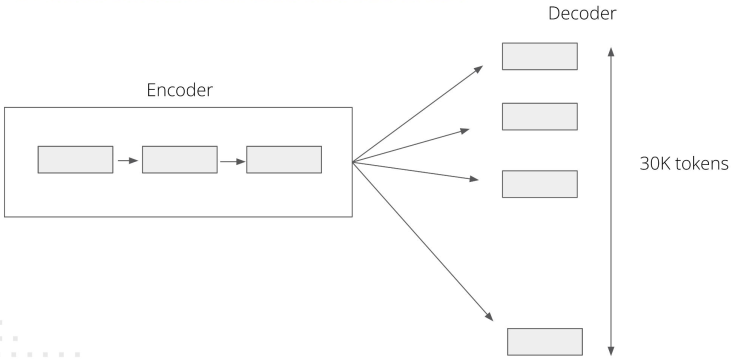
Question: Which startups received investment from funds that were over $1B in 2018?

```
result(investor) =  transformer(query+corpus, answer)
and toInvestor[answer]=investor and fund[investor] > $1B
```

Compute this first

Use these results to reduce the search space of the transformer

Transformers brute forces a lot



At every step the decoder **needs to consider 30K** tokens, but **very few of them are valid** based on the constraints

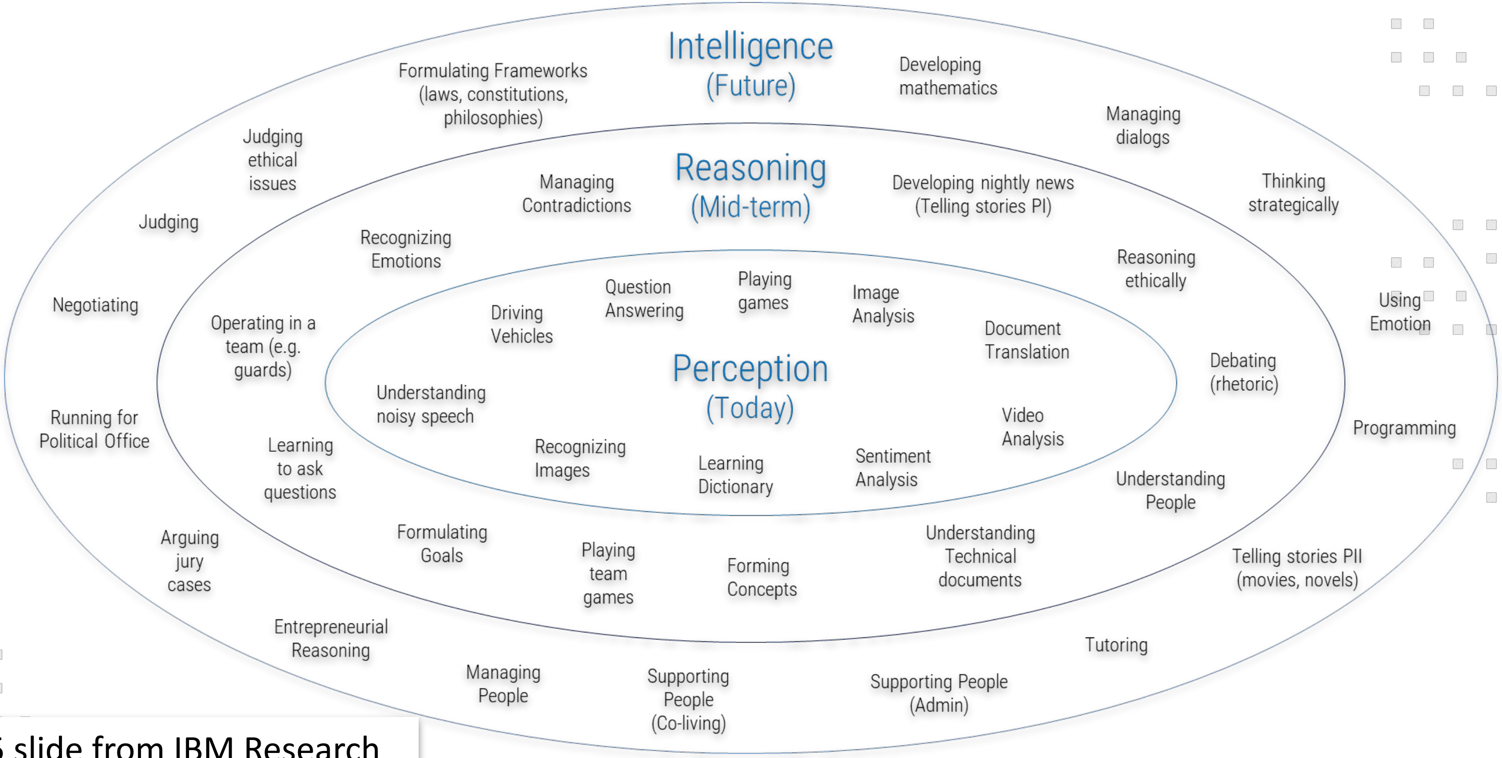
66

Conclusion

Where we go next

Conclusion – Tremendous progress, but a long way to go!

Central Q: Can we build systems that can work w/people and accomplish tasks in the world?



Vijay's 2015 slide from IBM Research

Thank you

69

RelationalAI

Appendix

Financial Applications

PIPG: Extract Formulas from Structured Notes

Consider the security note with language:
Issuer: Denominations:
Initial Valuation Date: Issue Date:
Final Valuation Date:* Maturity Date:* Reference Assets:
Payment at Maturity:
Barclays Bank PLC
Minimum denomination of $1,000, and integral multiples of $1,000 in excess thereof
December 22, 2021
December 28, 2021
December 23, 2024
December 27, 2024
The Russell 2000® Index (the "RTY Index") and the S&P 500® Index (the "SPX Index"), as set forth in the following table:
RTY Index RTY <Index> 2,221.90 1,555.33 1,555.33
SPX Index SPX <Index> 4,696.56 3,287.59 3,287.59
The RTY Index and the SPX Index are each referred to herein as a "Reference Asset" and, collectively, as the "Reference Assets."
If the Notes are not redeemed prior to scheduled maturity, and if you hold the Notes to maturity, you will receive on the Maturity Date a cash payment per $1,000 principal amount Note that you hold (in each case, in addition to any Contingent Coupon that m ay be payable on such date) determined as follows:

- If the Final Value of the Least Performing Reference Asset is greater than or equal to its Barrier Value, you will receive a payment of $1,000 per $1,000 principal amount Note.
- If the Final Value of the Least Performing Reference Asset is less than its Barrier Value, you will receive an amount per $1,000 principal amount Note calculated as follows:

$1,000 + [$1,000 × Reference Asset Return of the Least Performing Reference Asset]
If the Notes are not redeemed prior to scheduled maturity, and if the Final Value of the Least Performing Reference Asset is less than its Barrier Value, your Notes will be fully exposed to the decline of the Least Performing Reference Asset from its Initial Value. You may lose up to 100.00% of the principal amount of your Notes at maturity.
Please provide the formula for payment at maturity date in Prolog.

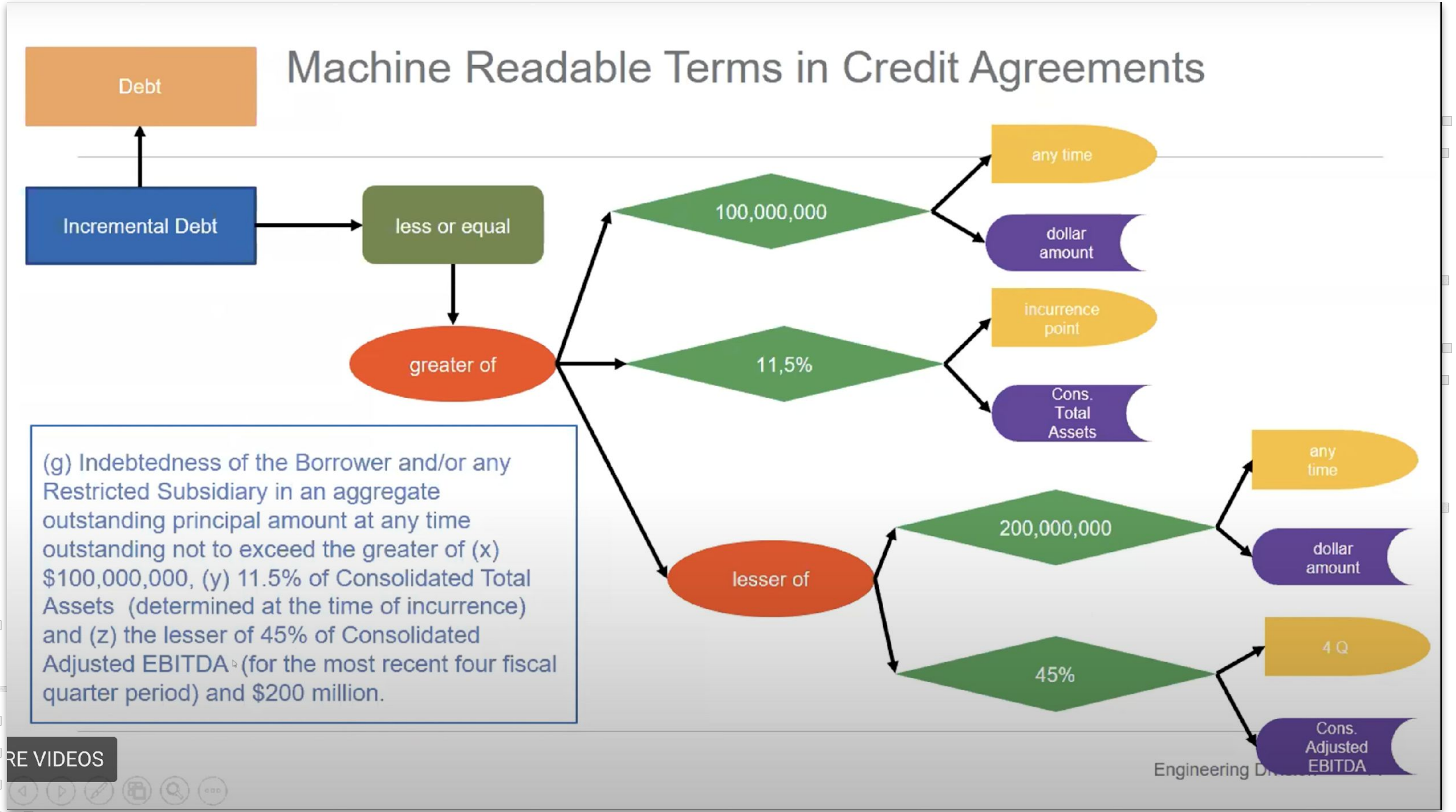
The formula for payment at the maturity date can be represented in Prolog as follows:

```
payment_at_maturity(FinalValue, BarrierValue, InitialValue, Payment) :-
FinalValue >= BarrierValue,
Payment is 1000.
payment_at_maturity(FinalValue, BarrierValue, InitialValue, Payment) :-
FinalValue < BarrierValue,
Payment is 1000 + (1000 * (FinalValue - InitialValue) / InitialValue).
```

This formula first checks whether the Final Value of the Least Performing Reference Asset is greater than or equal to its Barrier Value. If it is, the Payment is set to 1000. If it is not, the Payment is calculated as 1000 plus 1000 times the Reference Asset Return of the Least Performing Reference Asset (which is calculated as the difference between the Final Value and the Initial Value, divided by the Initial Value).

71

Creda: Extract Negative Covenants from Loans



RelationalAI

Nessie: News Analysis for Credit officers I/II

A counter-party credit risk officer is interested in monitoring any news about a company that may affect its credit-rating. Examples of such news:

Exec departure: Senior executive departure.
Positive Earnings: Positive earnings update with comments about outperforming expectations, or comments about increased operating profit, higher production, higher sales.
Negative Earnings: Operating loss announcement.
Changed market outlook: Positive or negative market outlook for the company or sector. Amount and percentage CAGR is discussed.
Acquisition target: Company is a target for acquisition, by another company or by the government.
New business: Getting a new deal or project with definite indication of cash inflow.
Material cash inflow: All events inolving material cash inflow are relevant (e.g. disposal of loss-making assets, inflow of cash through an IPO).
Cash out: e.g. layoffs and pay cuts are significant.
Litigation: Legal disputes and litigation that is material from a credit risk perspective.
Poor company performance: Poor execution at the firm.

Is the following news article relevant to such a risk officer at a bank that has loaned money to BBBY? Justify succinctly. Which of the above categories are relevant?
--- Article
Bed Bath & Beyond drops another 12% as investors eye survivability
BBBY -29.88%Jan. 06, 2023 7:13 AM ET
Bed, Bath & Beyond (NASDAQ:BBBY) swung lower again in premarket action on Friday.

Investors are scrambling following BBBY's going concern warning issued on Thursday and reports that a bankruptcy may be imminent. Even if BBBY survives, more store closings are anticipated. The retailer closed 37 locations in the U.S. across 19 states last year, including shuttering locations in New York, California, Florida, Alabama, Arizona, Georgia, Idaho, Missouri, Mississippi, Montana, Michigan, New Jersey and Minnesota.

In its SEC filing, BBBY said it is considering alternatives including restructuring or refinancing its debt, seeking additional debt or equity capital, reducing or delaying the company's business activities and strategic initiatives.

Shares of Bed, Bath & Beyond (BBBY) fell 11.85% premarket to $1.49 on heavy early volume.

The Seeking Alpha Quant Rating on BBBY has been flashing Sell since November.

This news article is relevant to a counter-party credit risk officer at a bank that has loaned money to BBBY. The article mentions the company's going concern warning and the possibility of bankruptcy, which would likely have a negative impact on BBBY's credit rating. The article also mentions that the company is considering restructuring or refinancing its debt, seeking additional debt or equity capital, and reducing or delaying the company's business activities and strategic initiatives, which suggests that BBBY may be facing financial difficulties and may be at risk of defaulting on its loans.

The relevant categories in the above list are:

- Poor company performance
- Operating loss announcement
- Negative Earnings
- Litigation
- Cash out
- Acquisition target

73

Nessie: News Analysis for credit officers II/II

Excellent. How about:

And:

Pfizer's pneumococcal vaccine for infants gets FDA priority review
PFE -0.94%Jan. 06, 2023 7:13 AM ET1 Comment
The U.S. Food and Drug Administration (FDA) granted priority review to Pfizer's (NYSE:PFE) application seeking approval of its 20-valent pneumococcal conjugate vaccine candidate (20vPnC).
The FDA accepted the supplemental Biologics License Application (sBLA) for 20vPnC to prevent invasive pneumococcal disease (IPD) caused by the 20 Streptococcus pneumoniae (pneumococcus) serotypes contained in the vaccine in infants and children 6 weeks through 17 years of age, and for preventing otitis media caused by seven of the 20 Streptococcus pneumoniae serotypes contained in the vaccine.
The FDA is expected to make a decision on the sBLA by April 2023.
Under priority review, the FDA's goal is to take action within six months, compared to 10 months under standard review.
The sBLA was backed by data from phase 3 and phase 2 trial programs for the pediatric indication for 20vPnC, according to the company.
20vPnC had previously received the FDA's fast track and breakthrough therapy designations for the pediatric indication for IPD.

This news article is not relevant to a counter-party credit risk officer at a bank that has loaned money to Pfizer. The article mentions the FDA granting priority review to Pfizer's application for its 20-valent pneumococcal conjugate vaccine candidate, which is a positive development for the company. The FDA's decision is expected to be made by April 2023. This news is related to the development of Pfizer's new drug and does not affect the company's credit-rating.

74

RelationalAI

Lani: Extract Operational terms from Corporate Loans

2.1 Commitment. Subject to and upon the terms and conditions hereof, and upon satisfaction of the conditions precedent set forth in Section 3.1, Lender shall make a loan to Borrower on the Financial Closing Date in an aggregate principal amount of $20,000,000 (the "Loan"). The Loan will be made in up to three advances relating to the three Projects (each, an "Advance"), which may occur on the same date. Lender shall not be required to make any Advance after December 31, 2011, and upon such date its obligation to make Advances shall expire. Lender shall not have an obligation to make an Advance in respect of a Project unless the NRG Investor and the Google Investor are required by their respective Equity Funding Agreements to make, or have made, equity contributions with respect to such Project and have furnished all letters of credit and other collateral required thereunder.

2.2 Promissory Note for Loan. The Loan made by Lender shall be evidenced by a promissory note executed by Borrower in favor of Lender in an amount equal to the aggregate amount of the Loan as provided in Section 2.1, substantially in the form of Exhibit A attached hereto (the "Note").

2.3 Use of Proceeds. The proceeds of the Loan shall be used by Borrower in such manner as Borrower determines is appropriate, acknowledging that Borrower intends to use the loan to reimburse Sponsor for Project Costs (as defined in the DOE Loan Documents) of the Project for which the Advance is made.

2.4 Loan Advance.

(a) Subject to and upon the terms and conditions set forth in this Agreement, Lender shall make each Advance to Borrower on the Financial Closing Date and Additional Advance Date, as applicable, by deposit of Advance proceeds to one or more accounts of a bank located in New York designated by Borrower. Borrower shall notify Lender in writing prior to the Financial Closing Date and Additional Advance Date, as applicable, of the account(s) (and, if there are multiple accounts, the amounts to be deposited into such accounts) into which the Loan proceeds shall be deposited.

(b) Borrower may request that Lender deposit Advance proceeds into an account of Sponsor by delivering to Lender an Advance Request executed by Borrower and Sponsor pursuant to which (i) Borrower certifies to Lender that such request is made to facilitate Borrower's distribution of such proceeds to Sponsor as authorized pursuant to the resolutions delivered to Lender under Section 3.1(h)(iv)(B), (ii) Borrower agrees to deliver to Lender on the Business Day after Sponsor's receipt of such proceeds a receipt confirming receipt of such proceeds on behalf of Borrower, and (iii) Borrower and Sponsor acknowledge and agree that any such deposit of Advance proceeds into an account of Sponsor constitutes for all purposes Lender's delivery of Advance proceeds to Borrower and a transfer of such proceeds by and from Borrower to Sponsor.

2.5 Interest.

(a) Interest Rate. From the Financial Closing Date until paid in full, the outstanding principal balance of the Loan shall bear interest at a rate of [*] (the "Base Rate") per annum, [*] as provided in the Note;

* Confidential Treatment Requested

22

- Around 120 attributes, including:
  - Underlying facilities (term, revolver, LOC…)
  - Base rates (Eurodollar, …)
  - Margin
  - Tenure
  - Currencies
  - Business Days
  - …
  - Plus capture cross product (not all options are available)

75

Working with Text: The Transformer

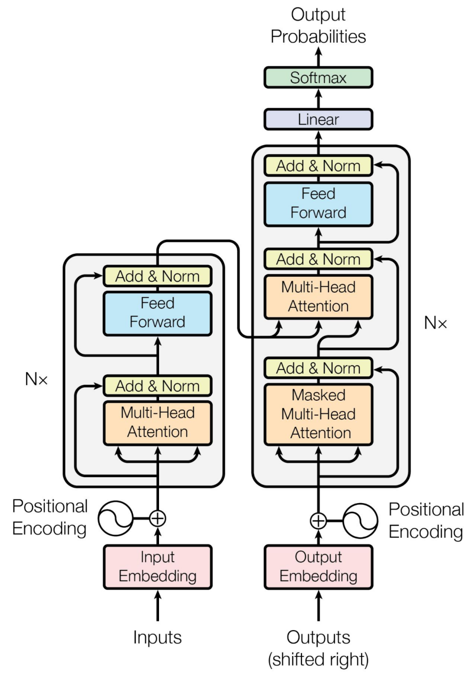


Figure 1: The Transformer - model architecture.

- Set up a path between any pair of tokens in a layer, to learn the strength of that pair – *multi-headed self-attention*
- Build up these patterns recursively, layer on layer.
- Have to explicitly represent position of token now.
- Self-attention supports *permutation-equivariance*
  - Key to compositional generalization? (cf Lake and Baroni, ICML 2018).

Breakthrough in NLU

Vaswani et al "Attention is all you need", 2017

Identifying data invariants and building them into the network is key to "programming"?

RelationalAI

Causal Language Model Self-supervision

The cat on the hot tin _____

roof 0.8
Timbuktu .0000001
pot 0.00004
Iodine 0.000006
bull-session 0.000000001

Generate billions of such examples automatically, train the model to predict.

Simply learn to predict the probability of the next token, conditioned on the context.

(Variants: Masked LM, see also modern "UL" work)

77

RelationalAI

Scale, baby, Scale

GPT-3 paper

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

GPT-4 is said to have 1T parameters

Realizing “GPT-3 performance” with small LLMS

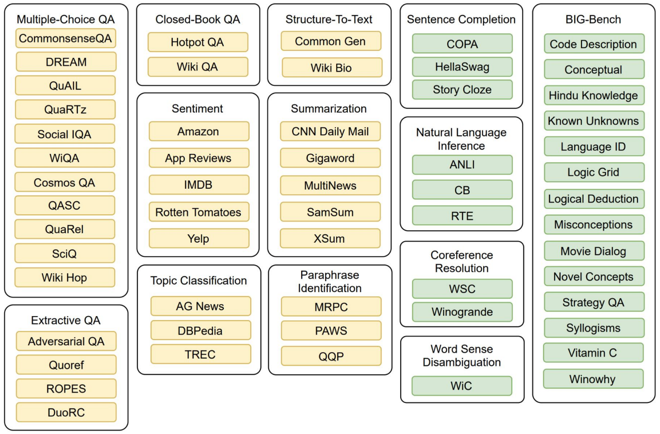


Figure 2: T0 datasets and task taxonomy. (T0+ and T0++ are trained on additional datasets. See Table 5 for the full list.) Color represents the level of supervision. Yellow datasets are in the training mixture. Green datasets are held out and represent tasks that were not seen during training. Hotpot QA is recast as closed-book QA due to long input length.

Published as a conference paper at ICLR 2022

MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh* Hugging Face
Albert Webson* Brown University
Colin Raffel* Hugging Face
Stephen H. Bach* Brown & Snorkel AI

Fine tune on labeled data-sets.

First convert all tasks to text-to-text tasks.

Hold out some tasks.
Train on data sets for other tasks.

Test for in-context learning on held out tasks.

79

RLHF: Training to better follow instructions

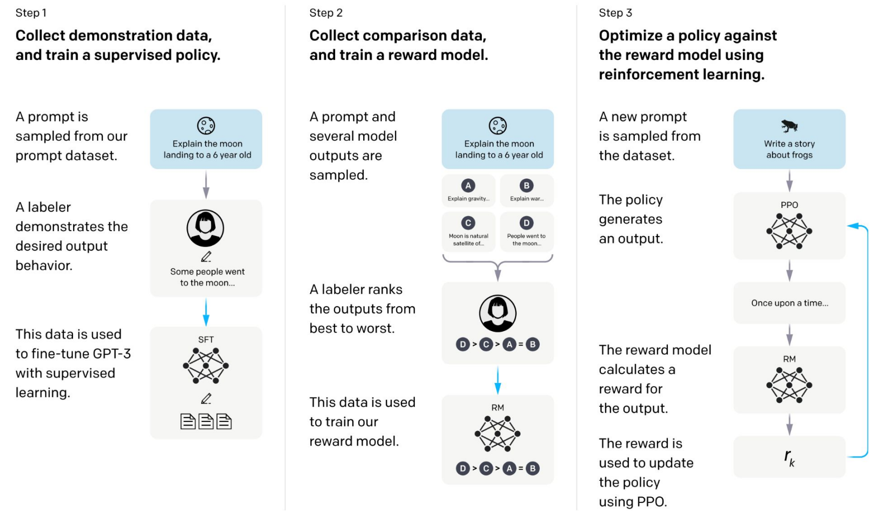


Figure 2: A diagram illustrating the three steps of our method: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. Blue arrows indicate that this data is used to train one of our models. In Step 2, boxes A-D are samples from our models that get ranked by labelers. See Section 3 for more details on our method.

Key to ChatGPT performance

https://openai.com/research/instruction-following

Credit: OpenAI

80