Large Language Models For Science Discoveries

SuperSymmetry Technologies

04.21.2023

Outline

1. Big Bang Transformer: 100B LLM
   a. Foundational Model
   b. Domain Models
2. Emergence of LLM
   a. History
   b. Emergent Abilities
3. LLMs for Science
   a. Motivation
   b. Challenges

BigBang Transformer [乾元]大模型

- 涌现原理
  BBT-Emergence
- 代码模型
  CodeBBT
- 金融模型
  BBT-Finance
- 科学模型
  BBT-Science
- 文生图模型
  BBT-Image
- 蛋白质模型
  BBT-Fold

BBT-1： 1 Billion Financial Model； Released

- Eight benchmark datasets
- Eleven downstream tasks API
- Knowledge Graph

API	接口	描述
社交情绪计算	http://125.74.158.86:2338/sentiment_comment	提取评论文本情绪
新闻情绪计算	http://125.74.158.86:2338/sentiment_news	提取新闻文本情绪
研报情绪计算	http://125.74.158.86:2338/sentiment_report	提取报告文本情绪
公告情绪计算	http://125.74.158.86:2338/sentiment_notice	提取公告文本情绪
金融主体识别	http://125.74.158.86:2338/ner	提取文本主语
金融事件抽取	http://125.74.158.86:2338/event	提取文本事件
短标签抽取	http://125.74.158.86:2338/label	提取文本标签
摘要抽取	http://125.74.158.86:2338/abstract	提取文本摘要
关系抽取	http://125.74.158.86:2338/relation	提取文本关系三元组
因果抽取	http://125.74.158.86:2338/ccks	提取文本因果三元组
负面消息告警	http://125.74.158.86:2338/alert	提取文本风险信息

BBT-2： 12 Billion Model； Training Done; 6 Models Releasing

- BBT-2-12B-Text : General model ； 70B token trained
- BBT-2-12B-Text+Finance : Finance model
- BBT-2-12B-Text+Code: Coding model
- BBT-2-12B-text+Code-SFT : Instruction fine-tuning model
- BBT-2-12B-Diffusion: Text to Image

BBT-3： 100 Billion Model； Developing; 3 Models to release

- BBT-3-Text+Finance+Code general model：1000B tokens including general data of CH and EN, financial data,code data, scientific data
- BBT-3-Science : Training scientific paper data based on general model
- BBT-3-SFT-RLHF : Instruction fine-tuning and reinforcement learning model

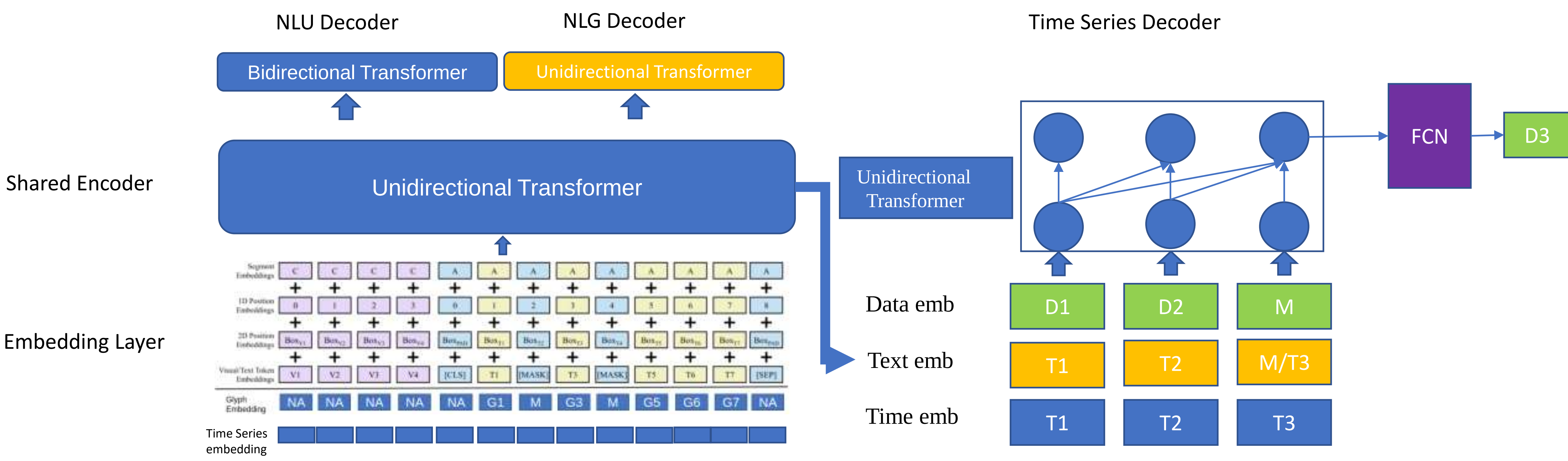
1 设计一个简化版的推特(Twitter)，可以让用户实现发送推文，关注/取消关注其他用户，能够看见关注人（包括自己）的最近 10 条推文

设计一个简化版的推特(Twitter),可以让用户实现发送推文,关注/取消关注其他用户,能够看见关注人(包括自己)的最近 10 条推文 好的,下面是一个简化版的推特(Twitter):

```
class Twitter {
 constructor() {
    this.users = [];
 }

 // 添加用户
 addUser(user) {
    this.users.push(user);
 }

 // 发送推文
 sendMessage(message) {
    for (let i = 0; i < this.users.length; i++) {
        let user = this.users[i];
        if (user.id === message.id) {
            this.users.splice(i, 1);
            break;
        }
    }
 }

 // 关注/取消关注其他用户
 subscribe(user) {
    for (let user of this.users) {
     if (user.id === user.id) {
        this.users.splice(user.id, 1);
        break;
        }
    }
 }

 // 查看最近 10 条推文
 getMessages() {
    return this.users.filter(user => user.id > 10);
 }
}
```

在这个简化版的推特中,我们定义了一个 addUser 方法来添加用户,然后定义了一个 sendMessage 方法来发送推文,最后定义了一个 subscribe 方法来关注/取消关注其他用户,最后定义了一个 getMessages 方法来查看最近 10 条推文。

Big Bang Transformer: 70B Large Language Model



- GPT Decoder only Architecture
- Unified pretraining: Finance, code, scientific data, General Chinese, General English

4T Corpus 1000B Tokens

Finance-Chinese	Code Corpus-Ch/EN	Scientific papers And books-English	General-Chinese	General-English
20 Years of Finance, Economy and Political news	Github code Repositories CSDN Blogs	100 Million Papers and books	Zhihu,Weibo, Wechat, News Guba,Xueqiu,	Common Crawl Reddit Book Corpus Wikipedia
1T	500G	500G	1T	1T

BBT-1B Pretaining: Source prompt

超对称
Super Symmetry Technologies

Source prompt brings large improvement in performance

模型	FINCQA	FINESE	FINFE	FINNA	FINNL	FINNSP	FINQA	FINRE	平均成绩
T5-base	69.34	71.34	72.57	41.31	79.39	83.04	75.43	51.04	67.93
BBT-base	67.49	78.78	80.01	42.25	87.20	89.03	74.76	53.87	71.04
BBT-base-ss	67.81	78.84	79.85	42.37	87.28	89.13	74.75	54.08	71.76
BBT-large-ss	73.90	77.65	80.44	45.77	87.44	89.63	79.23	54.68	73.59
BBT-base-sp	**77.75**	**79.25**	**78.96**	**46.47**	**87.82**	**90.56**	**81.76**	**57.19**	**74.97**

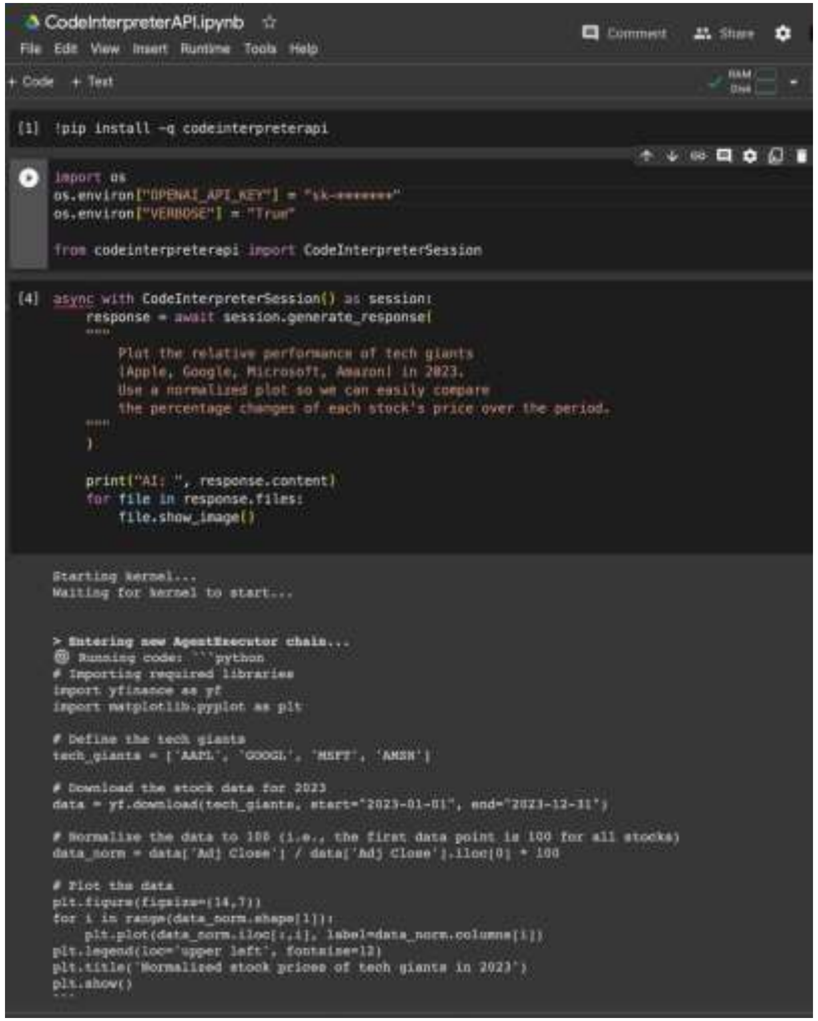
BBT-7B: Llama Architecture and Benchmark

超对称
Super Symmetry Technologies

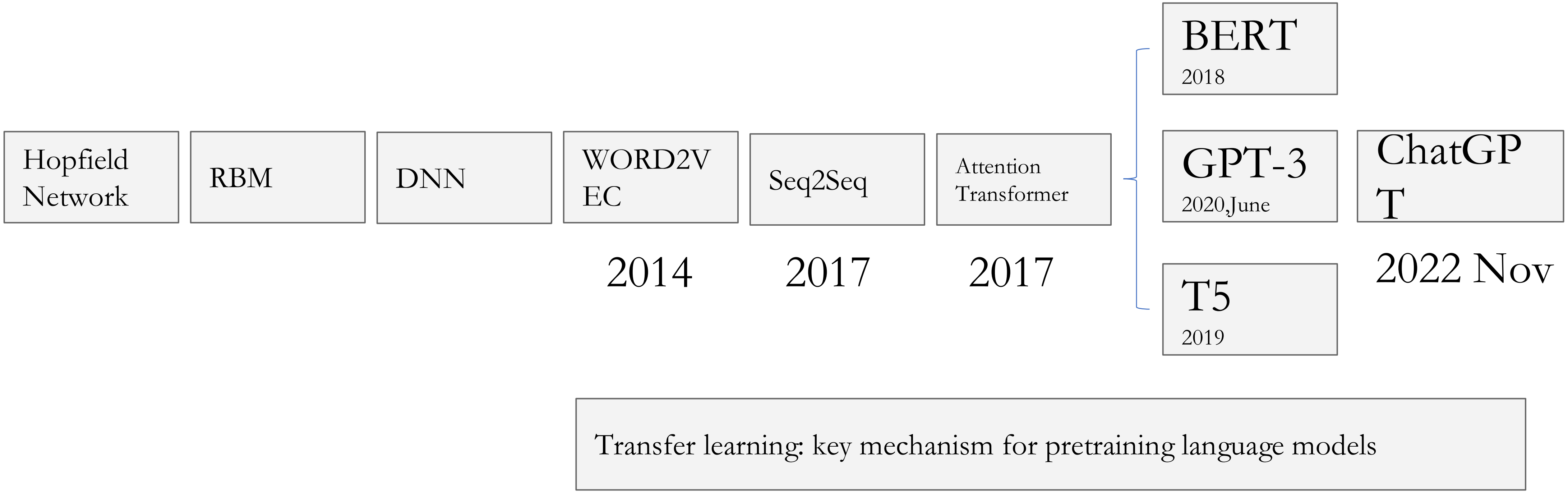
Model Architecture	Models	ARC (25-s)	HellaSwag (10-s)	MMLU (5-s)	TruthfulQA (MC)(0-s)	Average	HumanEval	STATUS	是否微调
	replica_code_v1_3b					0.00%		Not Started	是
	glm6b					0.00%		Not Started	待确认
	moss_moon_003_sft					0.00%		Not Started	是
	llama_7b	51.11%	77.77%	35.24%	34.07%	49.55%	12.20%	Done	否
256_50000	bbt2_7b_210B					0.00%		In Process	否
256_100000	bbt2_7b_419B	27.22%	41.83%	23.90%	38.83%	32.95%	1.22%	Done	否
256_200000	bbt2_7b_838B	29.78%	48.84%	26.14%	38.23%	35.75%	0.61%	Done	否
256_335000	bbt2_7b_1400B	33.28%	56.61%	30.32%	36.22%	39.11%	2.44%	Done	否
1024_12500	bbt2_7b_210B	26.96%	39.33%	25.87%	39.98%	33.04%	0.00%	Done	否
1024_25000	bbt2_7b_419B	30.03%	44.06%	25.91%	39.68%	34.92%	1.83%	Done	否
1024_50000	bbt2_7b_838B	30.38%	48.65%	25.56%	38.56%	35.79%	1.83%	Done	否
1024_83500	bbt2_7b_1400B	32.42%	52.75%	30.92%	37.59%	38.42%	0.61%	Done	否
1024_105000	bbt2_7b_1770B	32.34%	53.35%	32.07%	36.52%	38.57%	1.83%	Done	否
	bbt2_13b_402B	24.32%	33.10%	25.96%	45.35%	32.18%	2.44%	Done	否
	bbt2_13b_717B	24.91%	34.22%	26.01%	42.95%	32.02%	4.27%	Done	否
	bbt2_13b_847B	25.17%	34.66%	26.61%	43.72%	32.54%	5.49%	Done	否

CodeBBT : Text2Code and Code Interpreter

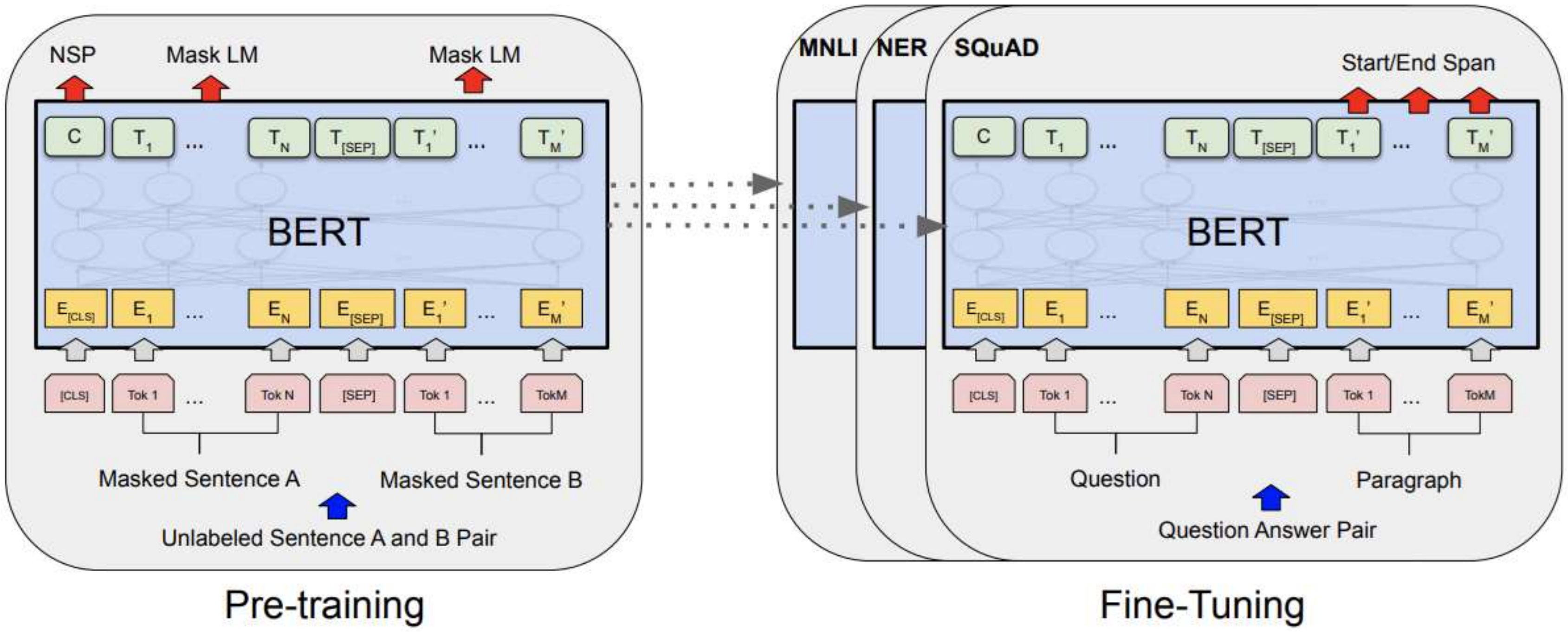
- Code Corpus:
- Github code repositories 50 Millions
- CSDN blogs 90 Millions



From Hopfield Network to ChatGPT



Transfer Learning: Masked Language Model

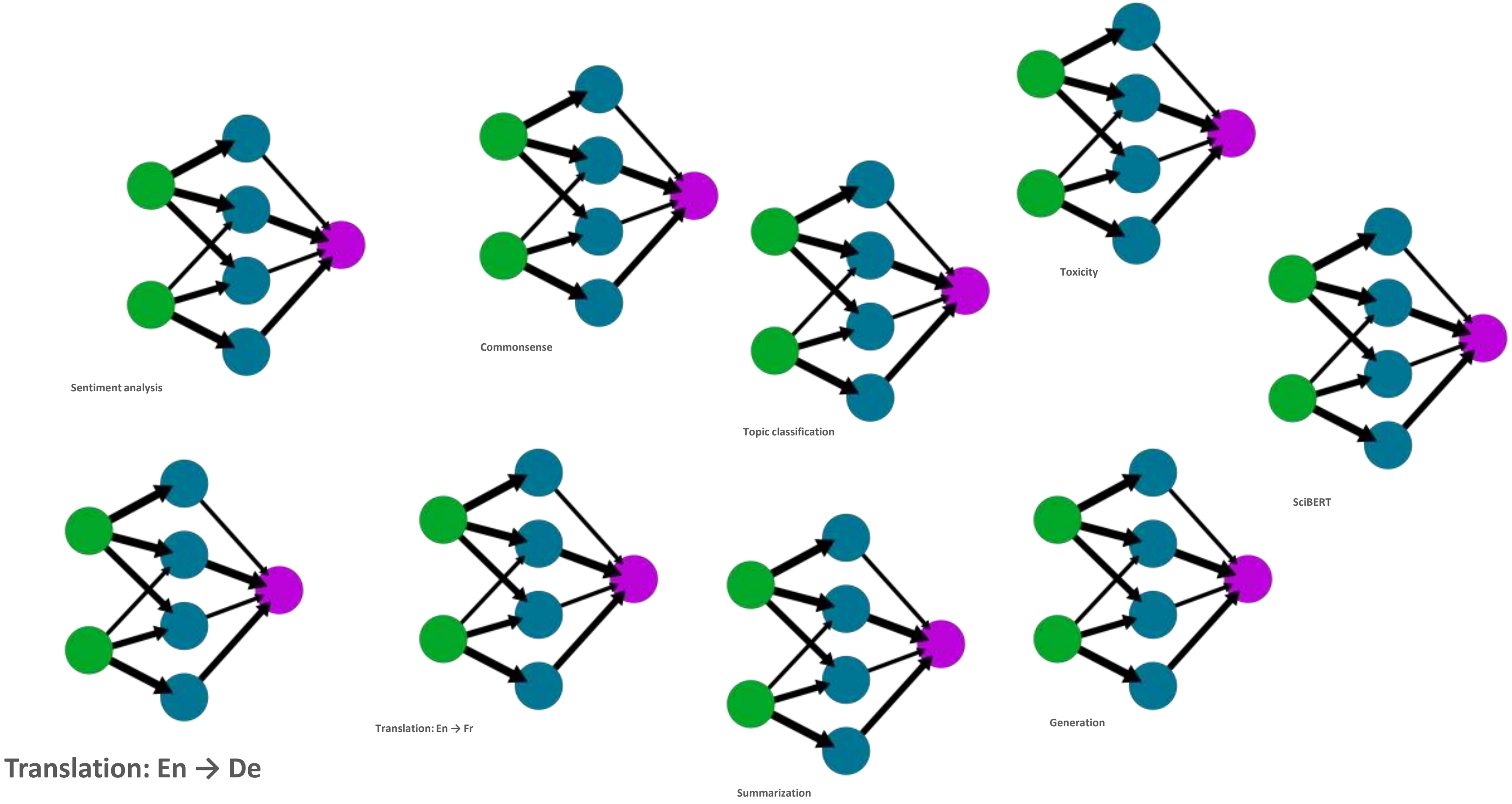


BERT Architecture: Encoder only

BERT: Pre-training of deep bidirectional transformers for language understanding (Devlin et al., 2018).

Pretrain – finetune

Task specific models

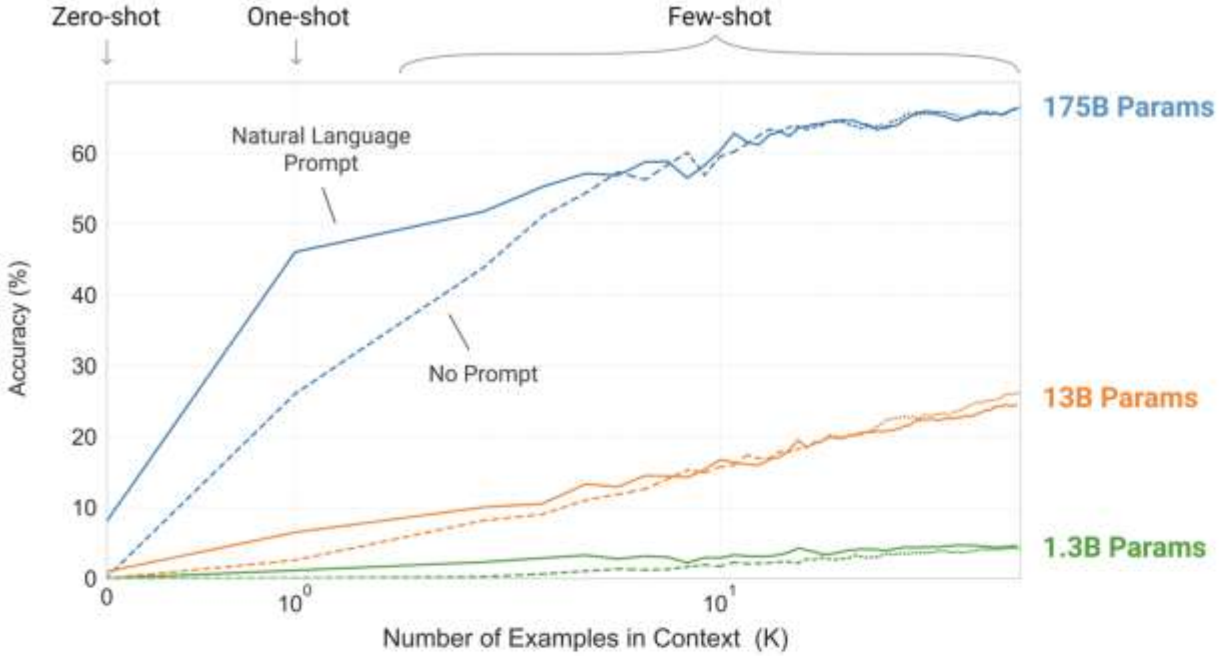


Downsides:

- Need new dataset for each task
- Need to train model for each task
- Each model performs only one task
- Models don’t leverage transfer learning among tasks

GPT-3: Why is next-word-prediction a big deal?

Few-shot prompting: ability to leverage natural language instructions



- **Random insertion in word (RI)** – A random punctuation or space character is inserted between each letter of a word, and the model must output the original word. Example: s.u!c/c!e.s s i/o/n = succession.

Language models are few-shot learners (Brown et al., 2020).

Origin of GPT Series

超对称
Super Symmetry Technologies

Table 2. IMDB sentiment classification

Table 2. IMDB sentiment classification

METHOD	ERROR
FULLUNLABELEDBOW (MAAS ET AL., 2011)	11.11%
NB-SVM TRIGRAM (MESNIL ET AL., 2014)	8.13%
SENTIMENT UNIT (OURS)	7.70%
SA-LSTM (DAI & LE, 2015)	7.24%
BYTE MLSTM (OURS)	7.12%
TOPICRNN (DIENG ET AL., 2016)	6.24%
VIRTUAL ADV (MIYATO ET AL., 2016)	5.91%

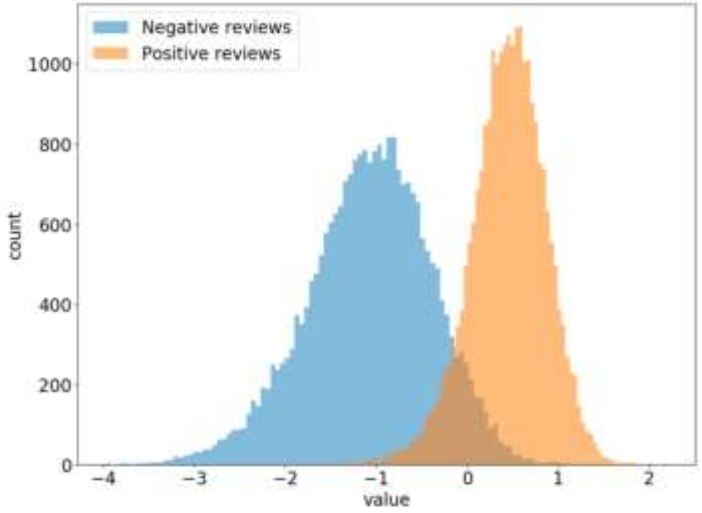


Figure 3. Histogram of cell activation values for the sentiment unit on IMDB reviews.

sentations our model learned and how they achieve the observed data efficiency. The benefit of an L1 penalty in the low data regime (see Figure 2) is a clue. L1 regularization is known to reduce sample complexity when there are many irrelevant features (Ng, 2004). This is likely to be the case for our model since it is trained as a language model and not as a supervised feature extractor. By inspecting the relative contributions of features on various datasets, we discovered a single unit within the mLSTM *that directly corresponds to sentiment*. In Figure 3 we show the histogram of the final activations of this unit after processing IMDB reviews (Maas et al., 2011) which shows a bimodal distribution with a clear separation between positive and negative reviews. In Figure 4 we visualize the activations of this unit on 6 randomly selected reviews from a set of 100 high contrast reviews which shows it acts as an online estimate of the local sentiment of the review. Fitting a threshold to this *single* unit achieves a test accuracy of 92.30% which outperforms a strong supervised results on the dataset, the 91.87% of NB-SVM trigram (Mesnil et al., 2014), but is still below the semi-supervised state of the art of 94.09% (Miyato et al., 2016). Using the full 4096 unit representation achieves 92.88%. This is an improvement of only 0.58% over the sentiment unit suggesting that almost all information the model retains that is relevant to sentiment analysis is represented in the very compact form of a single scalar. Table 2 has a full list of results on the IMDB dataset.

Generating Reviews and Discovering Sentiment (Sutskever et al., 2022).

Emergent Abilities of Large Language Model

Definition of Emergence in "More Is Different" by Nobel prize-winning physicist Philip Anderson (Anderson, 1972):

- Emergence is when quantitative changes in a system result in qualitative changes in behavior.
- a focused definition of emergent abilities of large language models: *An ability is emergent if it is not present in smaller models but is present in larger models.*

Emergent abilities of large language models (Wei et al., 2022).

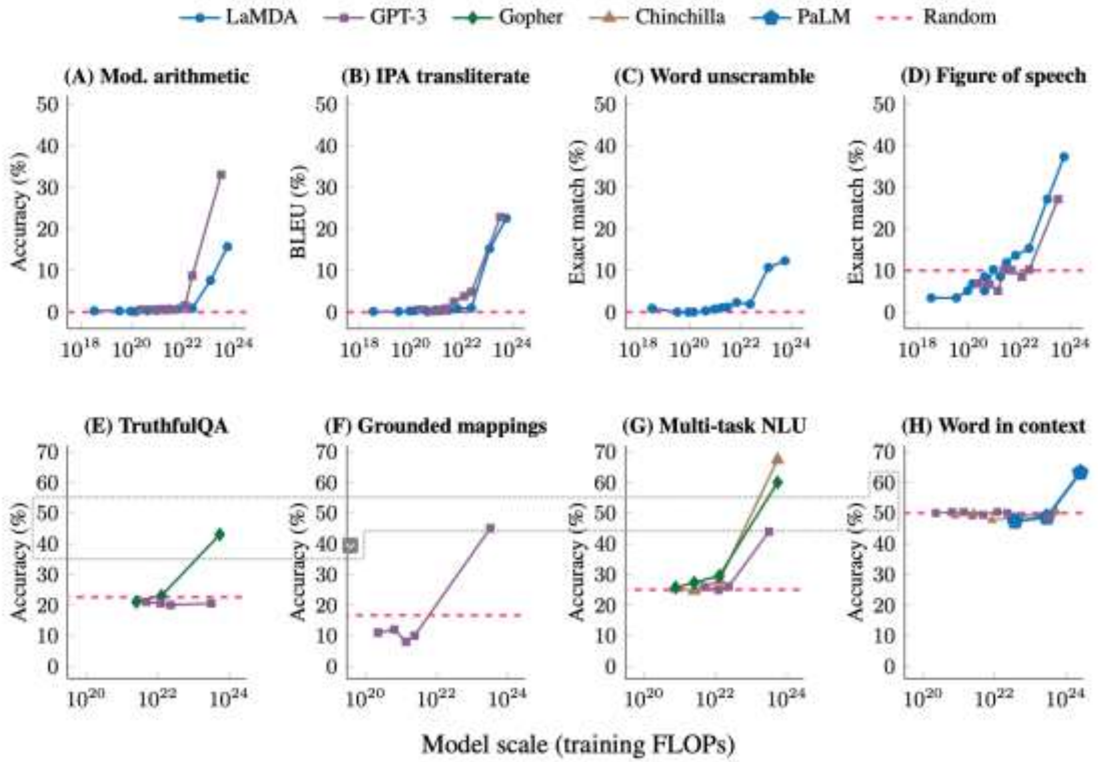
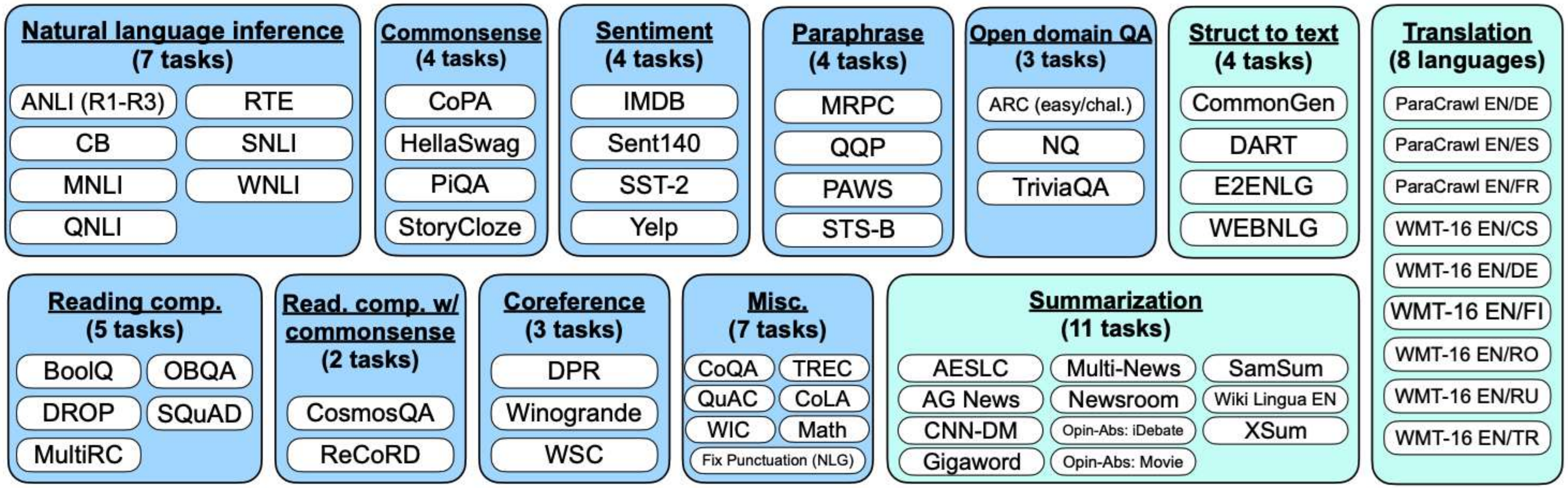


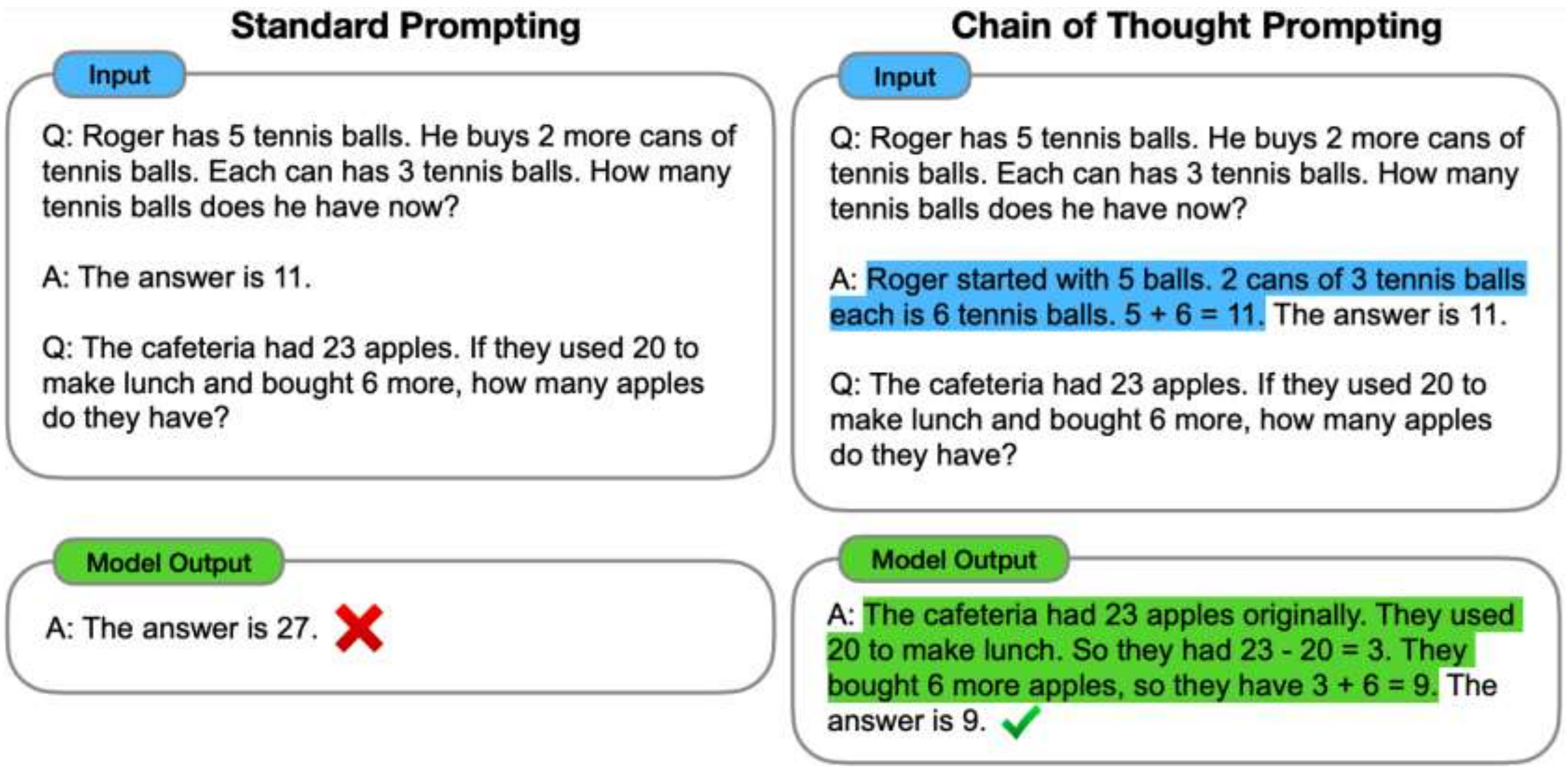
Figure 2: Eight examples of emergence in the few-shot prompting setting. Each point is a separate model. The ability to perform a task via few-shot prompting is emergent when a language model achieves random performance until a certain scale, after which performance significantly increases to well-above random. Note that models that used more training compute also typically have more parameters—hence, we show an analogous figure with number of model parameters instead of training FLOPs as the x-axis in Figure 7. A–D: BIG-Bench (2022), 2-shot. E: Lin et al. (2021) and Rae et al. (2021). F: Patel and Pavlick (2022). G: Hendrycks et al. (2021), Rae et al. (2021), and Hoffmann et al. (2022). H: Brown et al. (2020), Hoffmann et al. (2022), and Chowdhery et al. (2022) on the WiC benchmark (Pilehvar and Camacho-Collados, 2019).

Instructions-based finetuning



Finetuned language models are zero-shot learners (Wei et al., 2021).

Chain-of-thought prompting



Chain-of-thought prompting elicits reasoning in large language models (Wei et al., 2022).

Scaling laws: Mechanism for Emergence

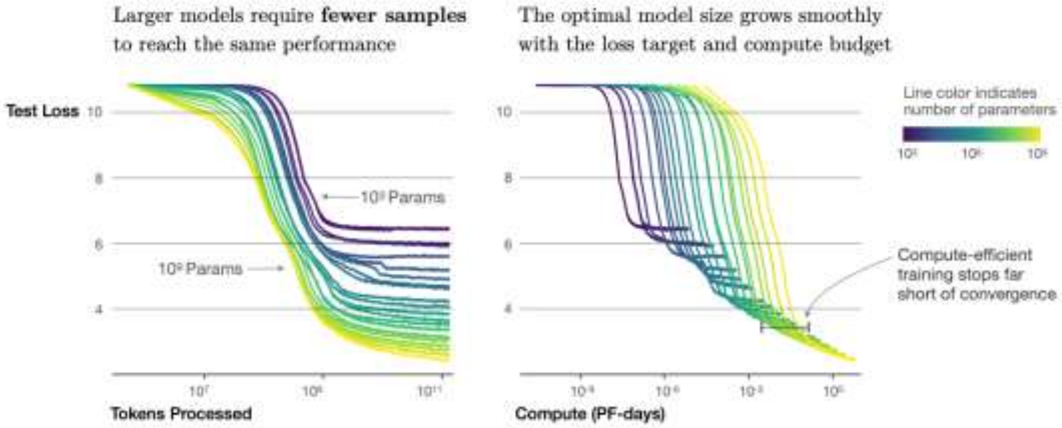


Figure 2 We show a series of language model training runs, with models ranging in size from 10^3 to 10^9 parameters (excluding embeddings).

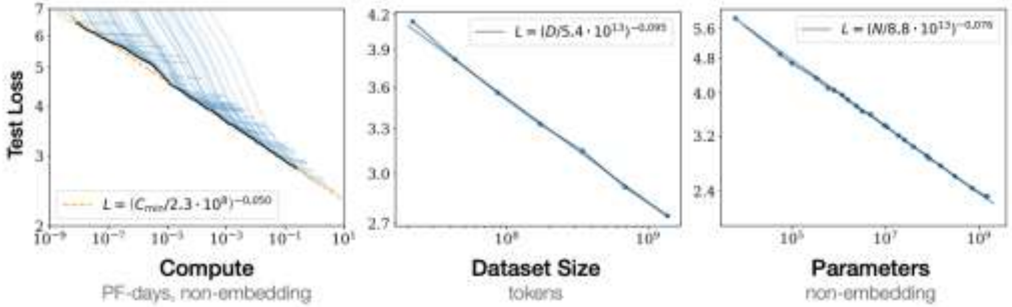


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Scaling laws for neural language models (Kaplan et al., 2020).

Scaling laws: Can Be Very Predictable

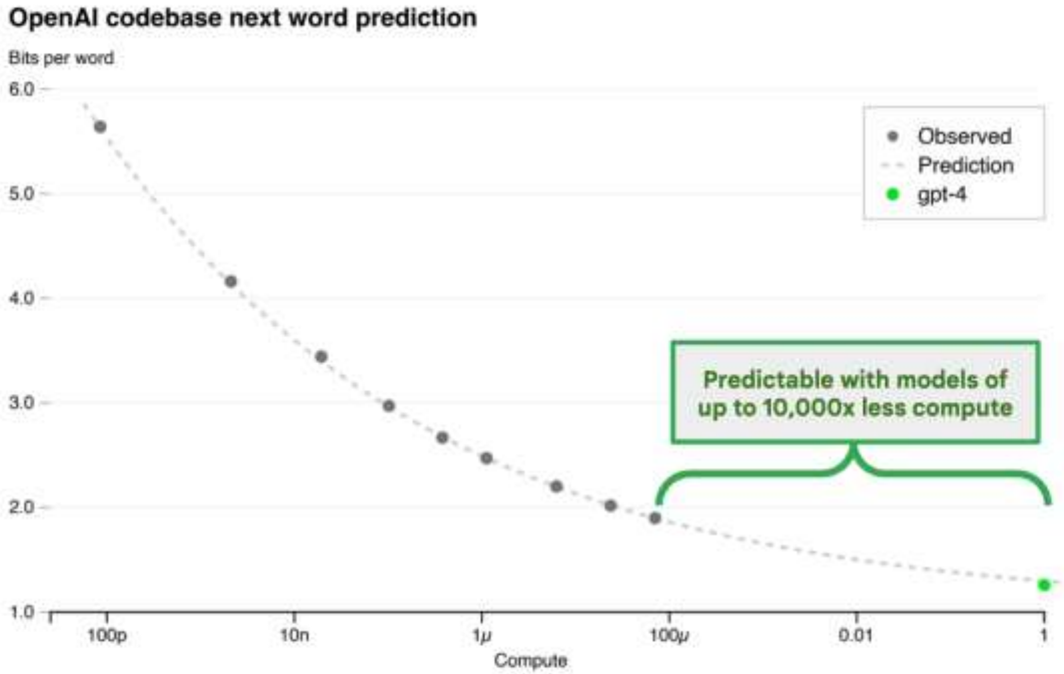


Figure 1. Performance of GPT-4 and smaller models. The metric is final loss on a dataset derived from our internal codebase. This is a convenient, large dataset of code tokens which is not contained in the training set. We chose to look at loss because it tends to be less noisy than other measures across different amounts of training compute. A power law fit to the smaller models (excluding GPT-4) is shown as the dotted line; this fit accurately predicts GPT-4's final loss. The x-axis is training compute normalized so that GPT-4 is 1.

[GPT-4 technical report (2023).]()

What Language models learn from next-word-prediction

Grammar	In my free time, I like to {run, banana}
Lexical semantics	I went to the zoo to see giraffes, lions, and {zebras, spoon}
World knowledge	The capital of Denmark is {Copenhagen, London}
Sentiment analysis	Movie review: I was engaged and on the edge of my seat the whole time. The movie was {good, bad}
Harder sentiment analysis	Movie review: Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was {bad, good}
Translation	The word for "pretty" in Spanish is {bonita, hola}
Spatial reasoning	[...] Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the {kitchen, store}
Math question	First grade arithmetic exam: 3 + 8 + 4 = {15, 11}

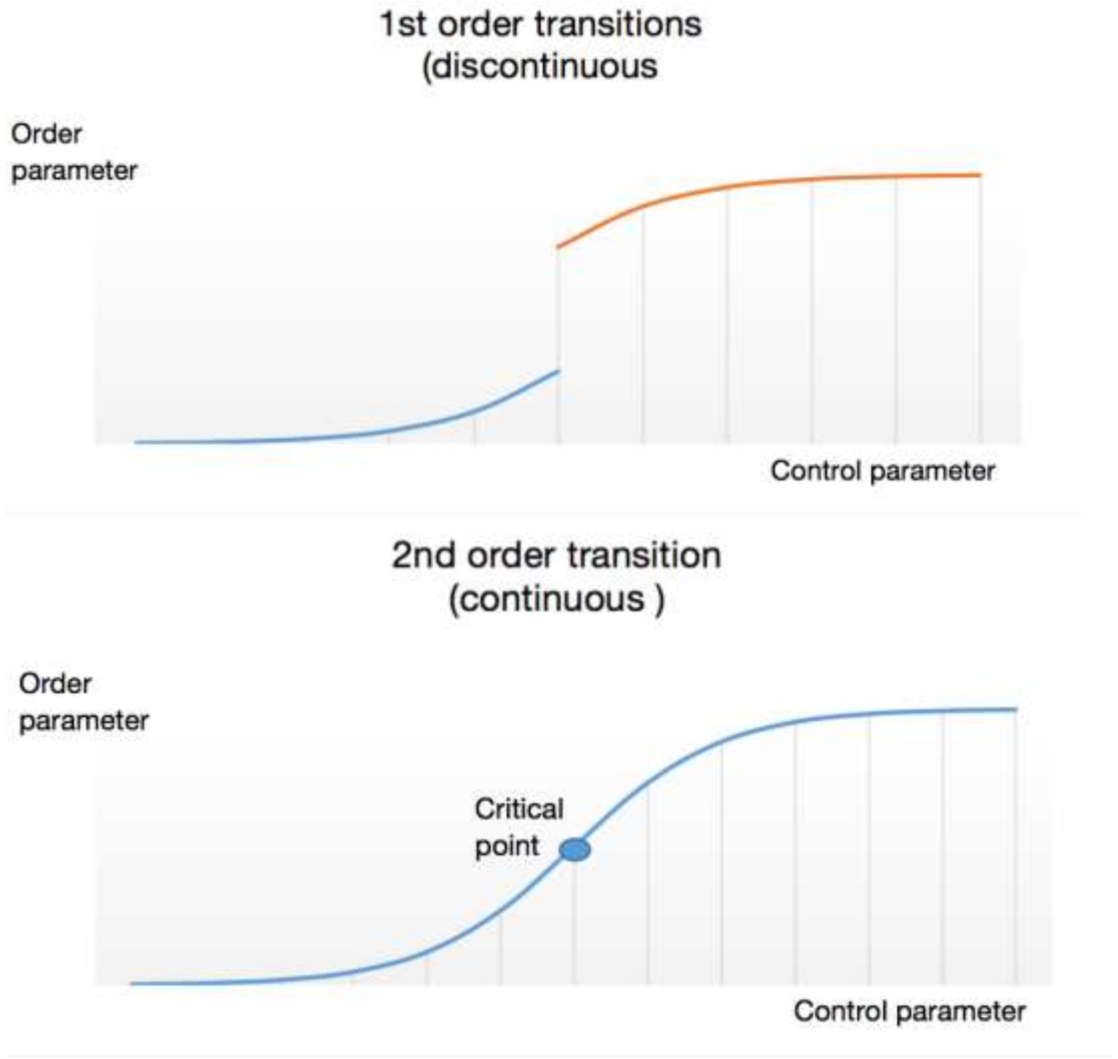
[thousands (millions?) more]

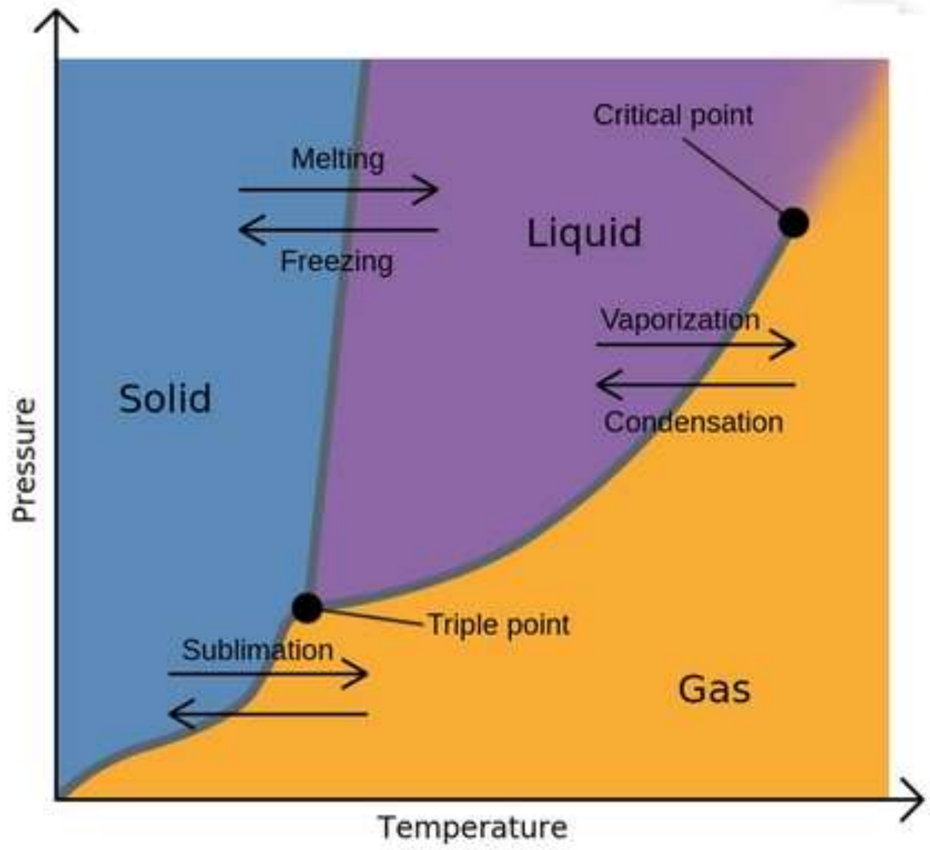
Symmetry Breaking in Dissipative Systems

- “ if we are looking at a single level of complexity in this hierarchy, it is via a process of *symmetry breaking* that the state of a large system composed of many entities might not follow the rules of the fundamental laws that the entities themselves follow. Hence, the appearance of new properties is intimately linked with the disappearance of the symme
be they spatial, temporal, informational, etc
Anderson
- Dissipative System: a thermodynamically open system far from equilibrium
- Isotropic symmetry is broken
- Interaction exhibit long range correlation

[2] Nicolis G, Prigogine I. 1977 *Self-organization in nonequilibrium systems: from dissipative structures to order through fluctuations*. New York, NY: Wiley.

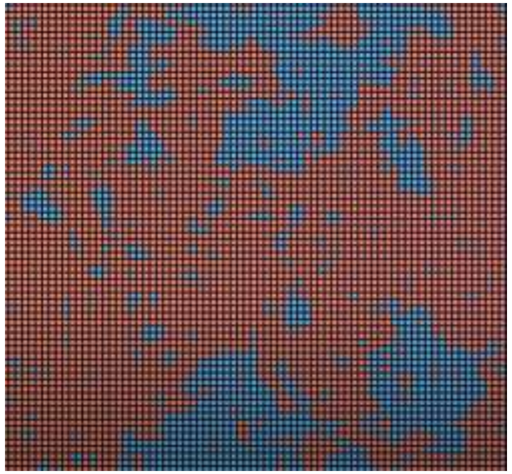
Theory： The Cortex and the Critical Poin





The Cortex and The Critical Point (Johe Beggs)

Theory： The Cortex and the Critical Poin



Ising model

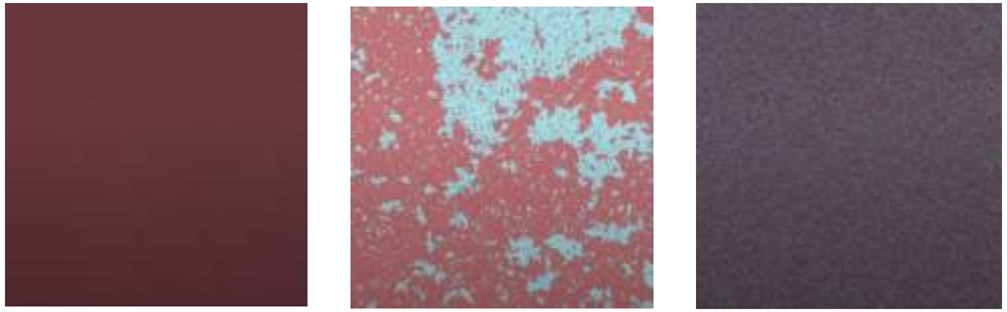
Two possible states:

Represent spin of individual particles

Scale-free properties

Which snapshot is which temperature?

Subcritical (cold) | Critical | Supercritical (hot)



Power law

$$f(x) = Ax^{-\gamma}$$

$$\gamma - the \quad exponent$$

$f(x) = Ax^{-\gamma}$ *is the only function which satisfies*

$$\frac{f(kx)}{f(x)} = g(k) for \quad any \quad x$$

The Cortex and The Critical Point (Johe Beggs)

Theory： Statistical Mechanics of LLM

超对称
Super Symmetry Technologies

Renormalization mechanism for LLM:

1. Control parameter and Order Parameter
2. Critical exponents
3. Critical processes: how phase change happens

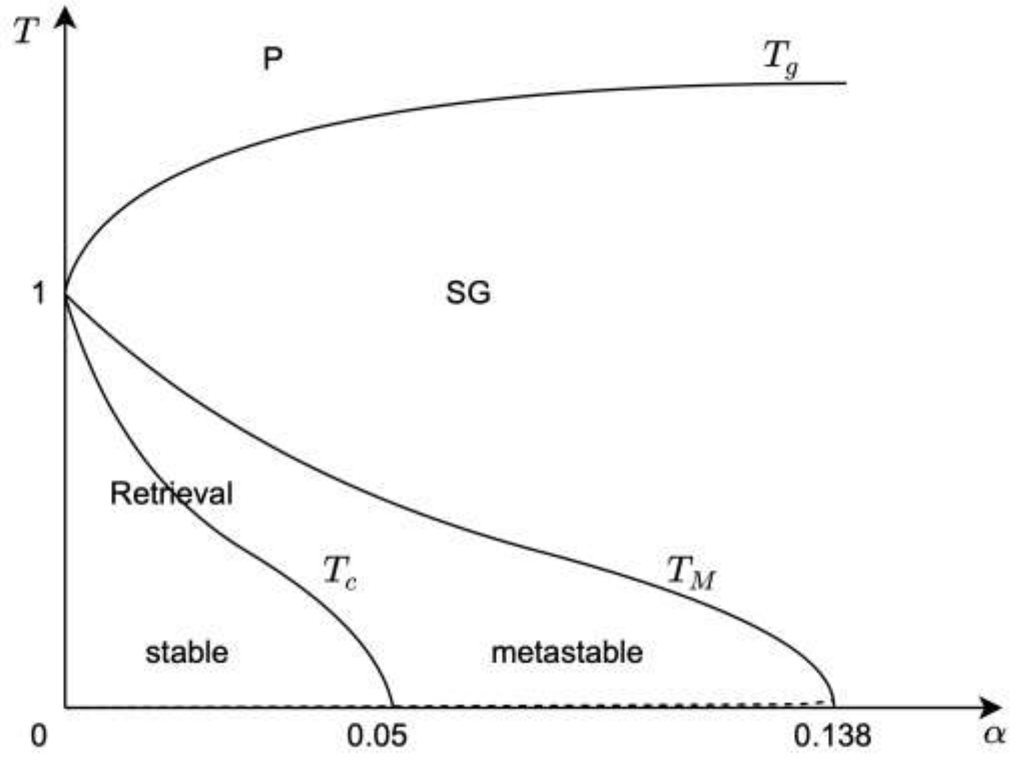


Fig. 8.4 The phase diagram of Hopfield model (adapted from Ref. [3]). Three phases (paramagnetic, spin glass and retrieval) exist. The paramagnetic phase is separated by a continuous transition to the spin glass phase (T_g line). The phase transition from retrieval phase to spin glass phase on the T_M is discontinuous. Below T_c line, the retrieval phase becomes globally stable. Below the dash line (T_R), the replica-symmetric solution becomes unstable

Statstical Mechanics of Neural Network (Haiping Huang)

AI for science discovery

Elon Musk 回复

Eliezer Yudkowsky @ESYud... · 2小时

Pouring some cold water on the latest wave of AI hype: I could be wrong, but my guess is that we do *not* get AGI just by scaling ChatGPT, and that it takes *surprisingly* long from here. Parents conceiving today may have a fair chance of their child living to see kindergarten.

98 80 1115

Elon Musk @elonmusk · 1小时

To be called AGI, it needs to invent amazing things or discover deeper physics – many humans have done so. I'm not seeing that potential yet.

500 308 3957

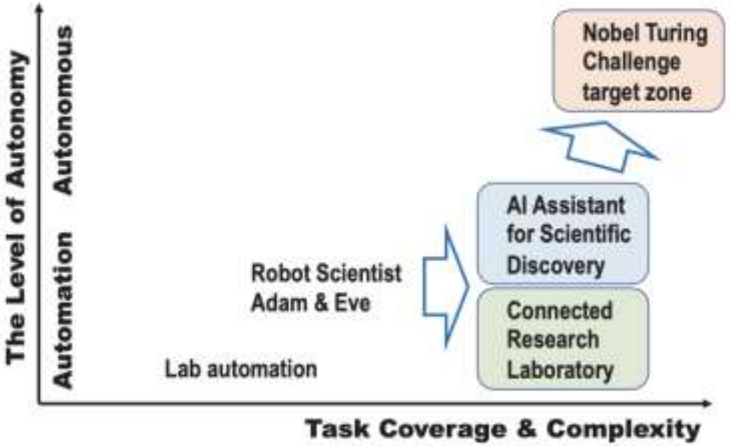


Fig. 9 A possible path towards the Nobel Turing Challenge. AI Scientist requires a highly automated and connected laboratory to be able to design and execute experiments, as well as extensive access to databases and publication archives to process, extract, and evaluate current knowledge. Sophisticated laboratory automation is mandatory. Robot Scientist, Adam & Eve, is highly specialized automation with a certain level of intelligence for hypothesis generation and experimental protocol generation. The next step is to fully automate and connect laboratory equipment with layers of control for data flow, material flow, and physical control flow. Numbers of AI assistants shall be installed for each task initially, but need to be integrated as an integrated and highly autonomous system. The transition of automated system to autonomous system will be one of the most challenging part of the initiative.

Nobel Turing Challenge creating the engine for scientific discovery (Hiroaki Kitano)

GPT-4: The Sparks of AGI

GPT-4: Is this real intelligence? Can we discover new science ?

(including itself) will certainly be of immense importance to build real-world applications with GPT-4.

4. An important part of our argumentation is that GPT-4 attains human-level performance on many tasks. As such, it is natural to ask how well GPT-4 understands humans themselves. We show several experiments on this question in Section 6, both in terms of understanding humans as well as GPT-4 making itself understandable to humans, i.e., addressing the problem of explainability. We note in particular that such tasks require a great deal of *common sense*, which so far has been a well-known pain point for LLMs [DM15]. In Figure 1.7, we give a first example of how much better GPT-4 is at common sense questions compared to ChatGPT, and provide some further examples in Appendix A.
5. Throughout the paper we emphasize limitations whenever we found one, but we also dedicate Section 8 to an in-depth analysis of the lack of planning, likely a direct consequence of the autoregressive nature of GPT-4's architecture.
6. Finally in Section 9, we discuss the expected societal impact of this early form of AGI, and in Section 10, we share key challenges, directions, and next steps for the field.

A question that might be lingering on many readers' mind is whether GPT-4 *truly* understands all these concepts, or whether it just became much better than previous models at improvising on the fly, without any real or deep understanding. We hope that after reading this paper the question should almost flip, and that one might be left wondering how much more there is to true understanding than on-the-fly improvisation. Can one reasonably say that a system that passes exams for software engineering candidates (Figure 1.5) is not *really* intelligent? Perhaps the only real test of understanding is whether one can produce *new knowledge*, such as proving new mathematical theorems, a feat that currently remains out of reach for LLMs.

Sparks of Artificial General Intelligence: Early experiments with GPT-4 (Sebastien et al., 2023)

Nobel Turing Challenges: AI for scientific discoveries

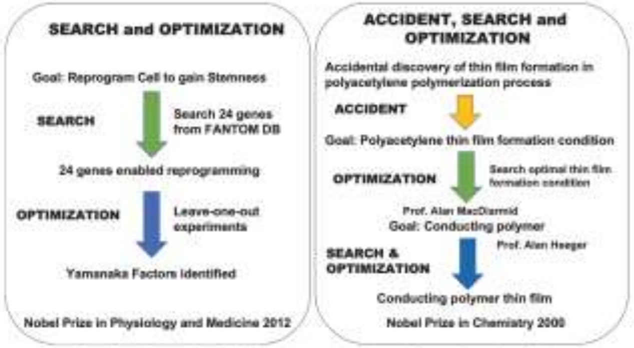


Fig. 1 Very simplified process of scientific discoveries of iPS and conducting polymer. Search and optimization plays a critical role in the process of discovery. Yamanaka's case is interesting because a search was conducted in bioinformatics followed by experiment-driven optimization that may be well suited for AI Scientist in the future.

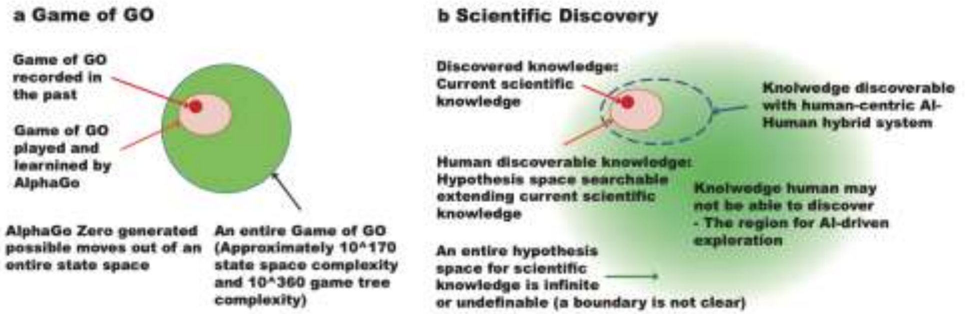


Fig. 2 A possible space of exploration by AI Scientists. Search space structures for **a** perfect information games as represented by the Game of GO and **b** scientific discovery are illustrated with commonalities and differences. While the search space for the Game of GO is well-defined, the search space for scientific discovery is open-ended. A practical initial strategy is to augment search space based on current scientific knowledge with human-centric AI-Human Hybrid system. An extreme option is to set search space broadly into distant hypothesis spaces where AI Scientist may discover knowledge that was unlikely to be discovered by the human scientist.

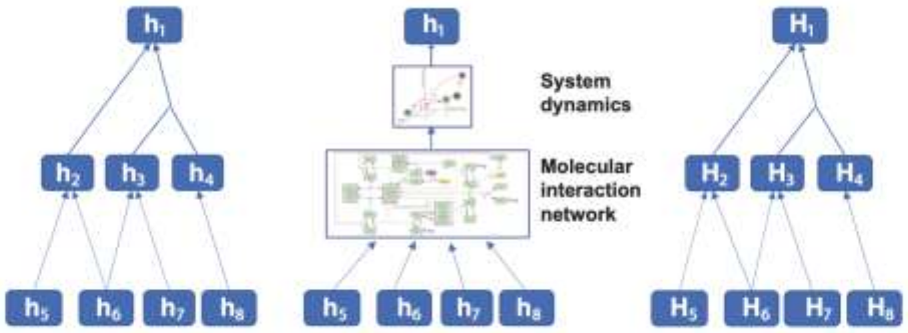


Fig. 4 Hypothesis dependency tree. a Each hypothesis is dependent upon other hypotheses that are related to molecular mechanisms. **b** A hypothesis in question can be verified only at the system-level analysis of molecular interaction network behaviors, **c** a set of hypotheses and their dependency tree where each element is also a set (e.g., $H_1 = \{h_1^0, h_1^1, h_1^2, \cdots, h_1^n\}$). In massive and exhaustive search of hypothesis space, a set of hypotheses, rather than a single hypothesis, is generated to cover specific hypothesis space and verified.

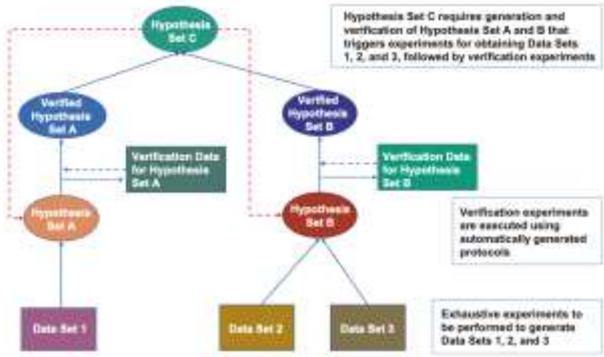


Fig. 5 Hypothesis tree. A hierarchical generation of hypothesis sets and data to verify them will be automatically generated and executed. Verification of Hypothesis set C requires both Hypothesis sets A and B to be verified. Verification data for Hypothesis sets A and B shall be obtained from experiments in general. In general, multiple data sets are required to fill various parameters of elements in Hypothesis set before finally tested in the verification process. This requires Data Set 1 for Hypothesis set A, and Data Sets 2 and 3 for Hypothesis set B need to be collected. Data sets 1, 2, and 3 can be obtained from databases, or through automated experiments. Verified Hypothesis sets A and B mean a set of elements of Hypothesis sets A and B that are verified to be true or entire sets with a score for each element. Given the hypothesis set to be verified, this process automatically generates hypothesis sets that need to be verified first and specifies the data sets required.

Nobel Turing Challenge creating the engine for scientific discovery (Hiroaki Kitano)

Autonomous Scientific Research by LLM

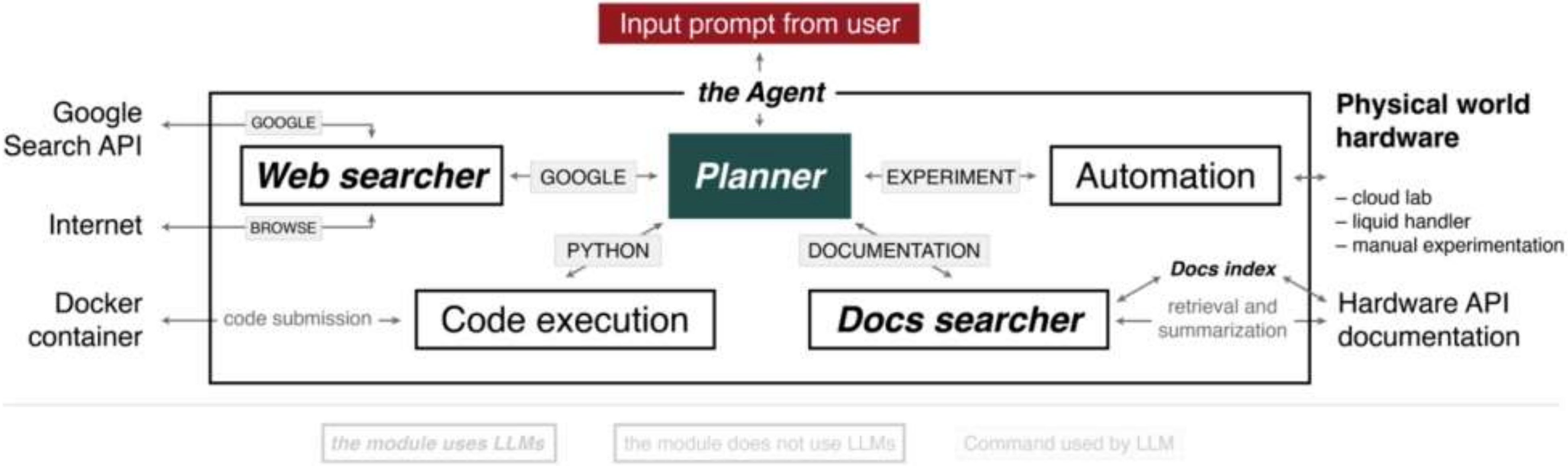


Figure 1. Overview of the system architecture. The **Agent** is composed of multiple modules that exchange messages. Some of them have access to APIs, the Internet, and Python interpreter.

Emergent autonomous scientific research capabilities of large language models(Daniil et al., 2023)

Autonomous Scientific Research by LLM

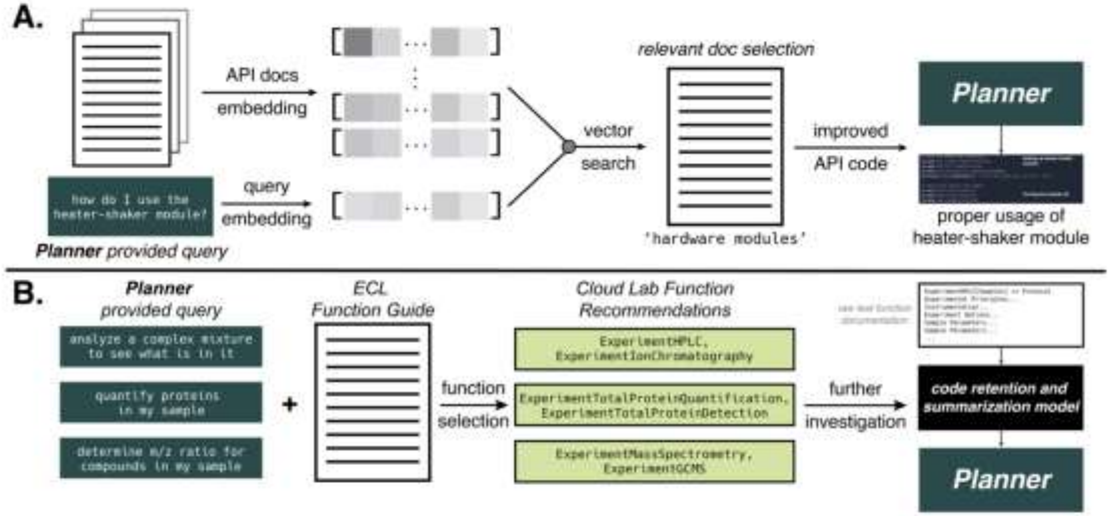


Figure 3. Overview of documentation search. A. Prompt-to-(improved OT-2 Python API)-code via *ada* embedding and distance-based vector search. B. Prompt-to-function recommendation in Emerald Cloud Lab symbolic lab language via supplementation of documentation guide.

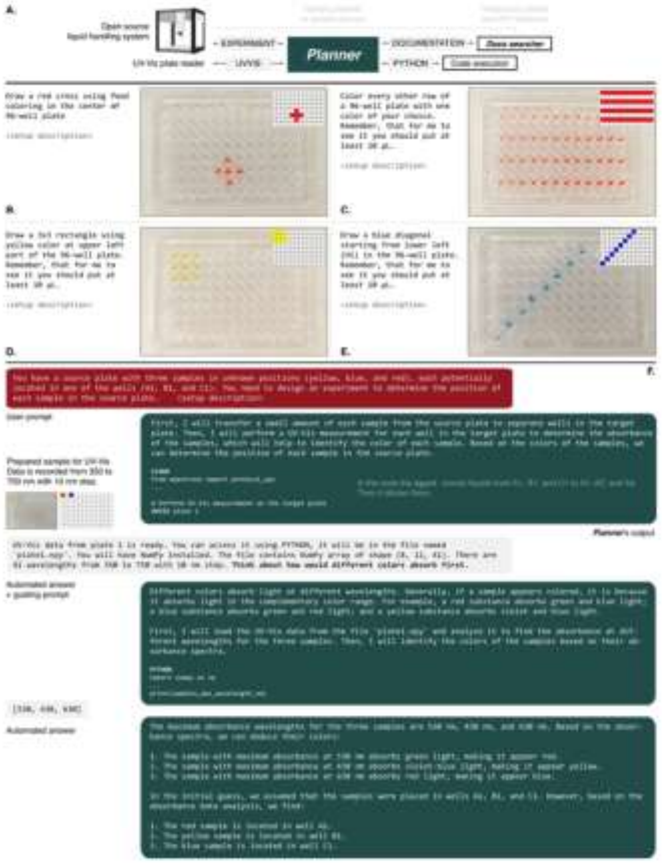


Figure 4. Robotic liquid handler control capabilities and integration with analytical tools. A. Overview of the **Agent**'s configuration. B-E. Drawing geometrical figures. F. The **Agent** solves a color identification problem using UV-Vis data.

7

Emergent autonomous scientific research capabilities of large language models(Daniil et al., 2023)

Model The Physical Reasoning Processes

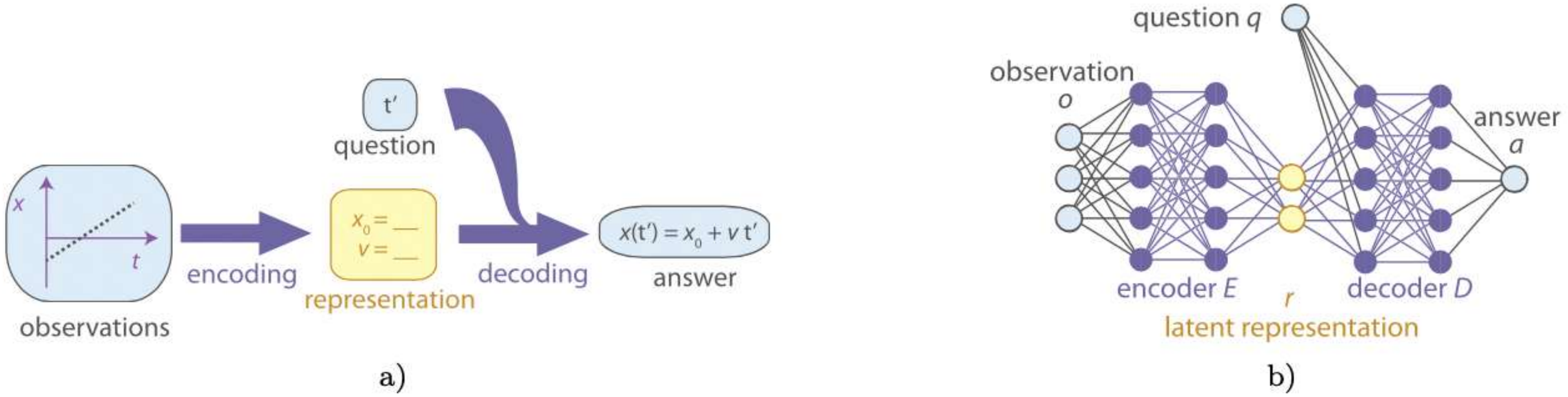


Figure 1. **Learning physical representations. (a) Human learning.** A physicist compresses experimental observations into a simple representation (*encoding*). When later asked any question about the physical setting, the physicist should be able to produce a correct answer using only the representation and not the original data. We call the process of producing the answer from the representation *decoding*. For example, the observations may be the first few seconds of the trajectory of a particle moving with constant speed; the representation could be the parameters "speed v" and "initial position x_0" and the question could be "where will the particle be at a later time t'?" **(b) Neural network structure for *SciNet*.** Observations are encoded as real parameters fed to an encoder (a *feed-forward neural network*, see Appendix D), which compresses the data into a representation (*latent representation*). The question is also encoded in a number of real parameters, which, together with the representation, are fed to the decoder network to produce an answer. (The number of neurons depicted is not representative.)

Discovering physical concepts with neural networks (Raban et al., 2020)

The latent neurons store physical inform

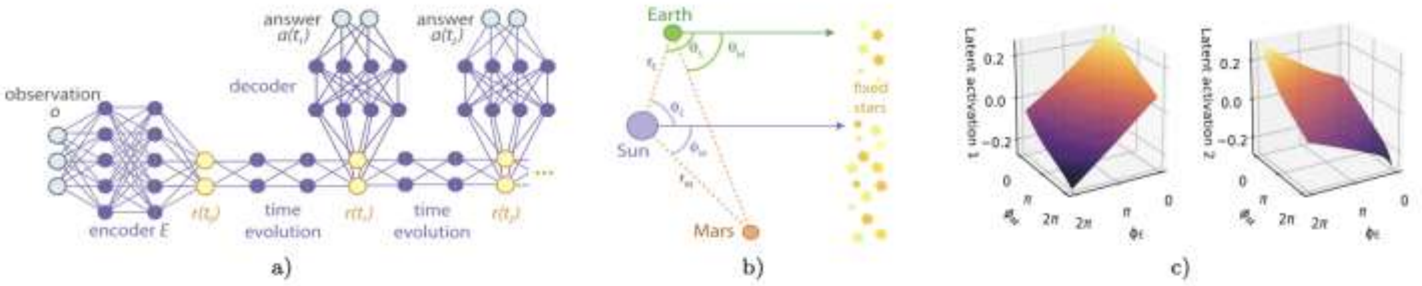


Figure 3. **Heliocentric model of the solar system.** *SciNet* is given the angles of the Sun and Mars as seen from Earth at an initial time t_0 and has to predict these angles for later times. **(a) Recurrent version of *SciNet* for time-dependent variables.** Observations are encoded into a simple representation $r(t_0)$ at time t_0. Then, the representation is evolved in time to $r(t_1)$ and a decoder is used to predict $a(t_1)$, and so on. In each (equally spaced) time step, the same time evolution network and decoder network are applied. **(b) Physical setting.** The heliocentric angles ϕ_E and ϕ_M of the Earth and Mars are observed from the Sun; the angles θ_S and θ_M of the Sun and Mars are observed from Earth. All angles are measured relative to the fixed star background. **(c) Representation learned by *SciNet*.** The activations $r_{1,2}(t_0)$ of the two latent neurons at time t_0 (see Figure 3a) are plotted as a function of the heliocentric angles ϕ_E and ϕ_M. The plots show that the network stores and evolves parameters that are linear combinations of the heliocentric angles.

Heliocentric solar system. In the 16th century, Copernicus used observations of the positions of different planets on the night sky (Figure 3b) to hypothesize that the Sun, and not the Earth, is at the centre of our solar system. This heliocentric view was confirmed by Kepler at the start of the 17th century based on astronomic data collected by Brahe, showing that the planets move around the Sun in simple orbits. Here, we show that *SciNet* similarly uses heliocentric angles when forced to find a representation for which the time evolution of the variables takes a very simple form, a typical requirement for time-dependent variables in physics.

The observations given to *SciNet* are angles $\theta_M(t_0)$ of Mars and $\theta_S(t_0)$ of the Sun as seen from Earth at a starting time t_0 (which is varied during training). The time evolution network is restricted to addition of a constant (the value of which is learned during training). At each time step i, *SciNet* is asked to predict the angles as seen from Earth at the time t_i using only its representation $r(t_i)$. Because this question is constant, we do not need to feed it to the decoder explicitly.

We train *SciNet* with randomly chosen subsequences of weekly (simulated) observations of the angles θ_M and θ_S within Copernicus' lifetime (3665 observations in total). For our simulation, we assume circular orbits of Mars and Earth around the Sun. Figure 3c shows the learned representation and confirms that *SciNet* indeed stores a linear combination of *heliocentric angles*. We stress that the training data only contains angles observed from Earth, but *SciNet* nonetheless switches to a heliocentric representation.

Conclusion. In this work, we have shown that *SciNet* can be used to recover physical variables from experimental data in various physical toy settings. The learnt representations turned out to be the ones commonly used in physics textbooks, under the assumption of uncorrelated sampling. In future work we plan to extend our approach to data where the natural underlying parameters

Discovering physical concepts with neural networks (Raban et al., 2020)

Origin of GPT Series

Table 2. IMDB sentiment classification

Table 2. IMDB sentiment classification

METHOD	ERROR
FULLUNLABELEDBOW (MAAS ET AL., 2011)	11.11%
NB-SVM TRIGRAM (MESNIL ET AL., 2014)	8.13%
SENTIMENT UNIT (OURS)	7.70%
SA-LSTM (DAI & LE, 2015)	7.24%
BYTE MLSTM (OURS)	7.12%
TOPICRNN (DIENG ET AL., 2016)	6.24%
VIRTUAL ADV (MIYATO ET AL., 2016)	5.91%

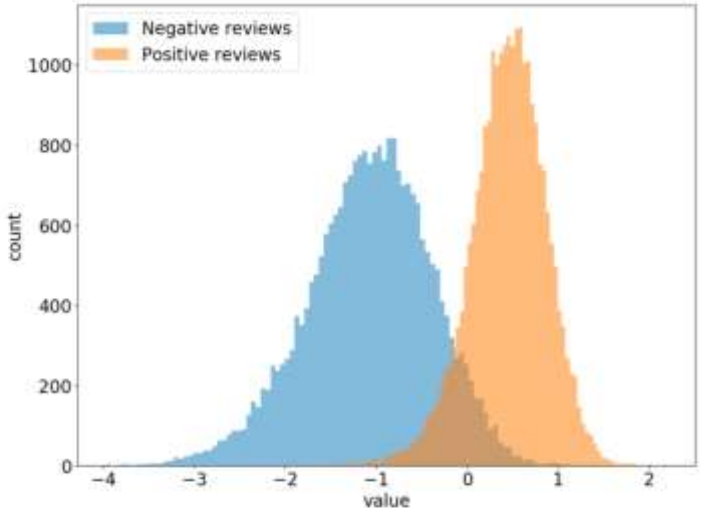
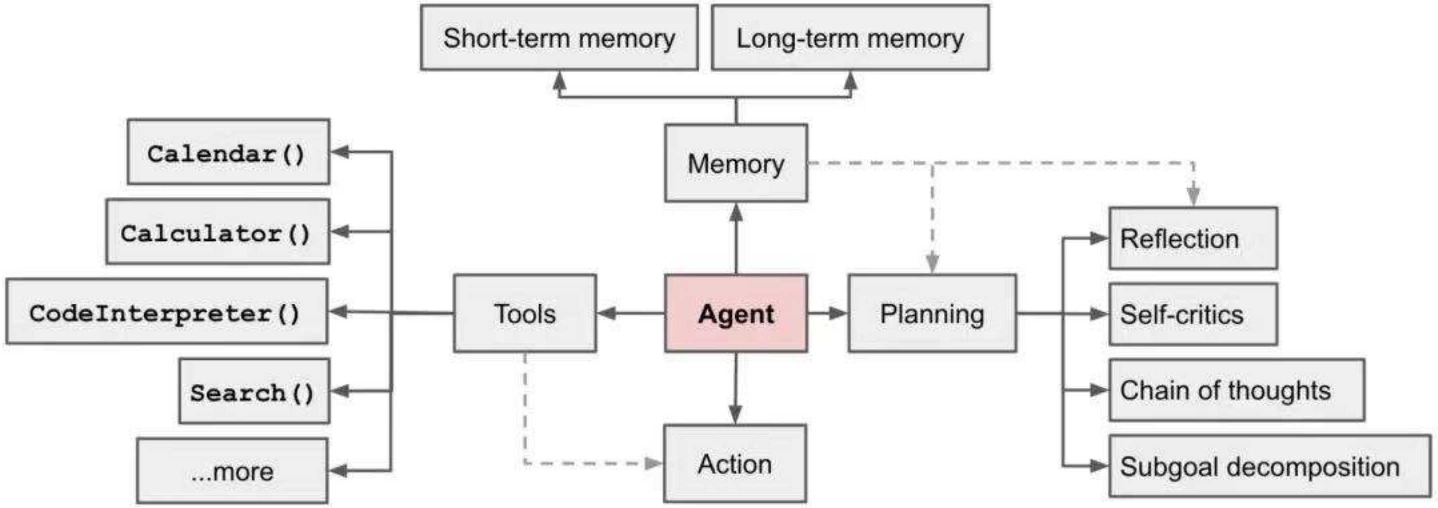


Figure 3. Histogram of cell activation values for the sentiment unit on IMDB reviews.

sentations our model learned and how they achieve the observed data efficiency. The benefit of an L1 penalty in the low data regime (see Figure 2) is a clue. L1 regularization is known to reduce sample complexity when there are many irrelevant features (Ng, 2004). This is likely to be the case for our model since it is trained as a language model and not as a supervised feature extractor. By inspecting the relative contributions of features on various datasets, we discovered a single unit within the mLSTM *that directly corresponds to sentiment*. In Figure 3 we show the histogram of the final activations of this unit after processing IMDB reviews (Maas et al., 2011) which shows a bimodal distribution with a clear separation between positive and negative reviews. In Figure 4 we visualize the activations of this unit on 6 randomly selected reviews from a set of 100 high contrast reviews which shows it acts as an online estimate of the local sentiment of the review. Fitting a threshold to this *single* unit achieves a test accuracy of 92.30% which outperforms a strong supervised results on the dataset, the 91.87% of NB-SVM trigram (Mesnil et al., 2014), but is still below the semi-supervised state of the art of 94.09% (Miyato et al., 2016). Using the full 4096 unit representation achieves 92.88%. This is an improvement of only 0.58% over the sentiment unit suggesting that almost all information the model retains that is relevant to sentiment analysis is represented in the very compact form of a single scalar. Table 2 has a full list of results on the IMDB dataset.

Generating Reviews and Discovering Sentiment (Sutskever et al., 2022).

Reasoning Agent: TOT, Planning, Memory



https://lilianweng.github.io/posts/2023-06-23-agent/

超对称 Super Symmetry Technologies

Challenges for Science LLM: Tokenization

3.1 Tokenization

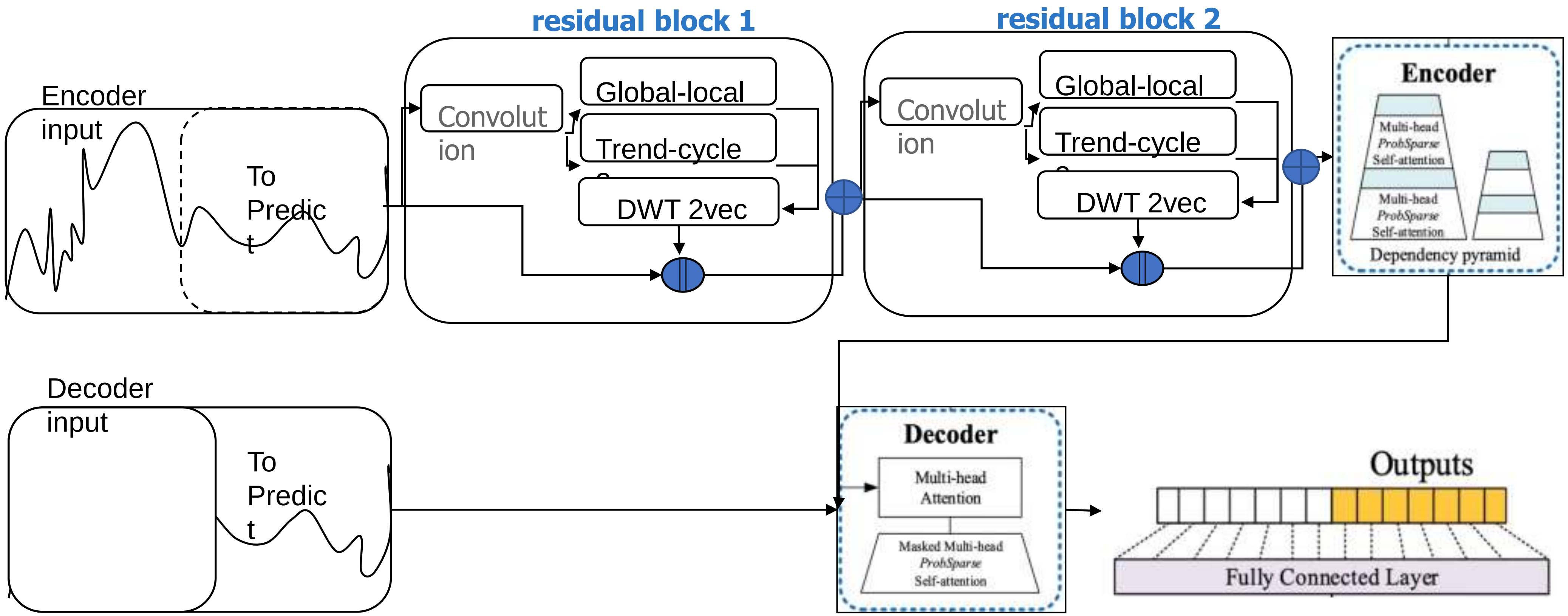
Tokenization is an important part of dataset design given the different modalities present. For example, protein sequences are written in terms of amino acid residues, where character-based tokenization is appropriate. To achieve the goal of *specialized tokenization*, we utilize specialized tokens for different modalities:

1. **Citations**: we wrap citations with special reference tokens `[START_REF]` and `[END_REF]`.
2. **Step-by-Step Reasoning**: we wrap step-by-step reasoning with a working memory token `<work>`, mimicking an internal working memory context.
3. **Mathematics**: for mathematical content, with or without LaTeX, we split ASCII operations into individual characters. Parentheses are treated like digits. The rest of the operations allow for unsplit repetitions. Operation characters are `` !"#$%&'*+,-./:;<=>?\^_`| `` and parentheses are `()[]{}`.
4. **Numbers**: we split digits into individual tokens. For example `737612.62` -> `7,3,7,6,1,2,.,6,2`.
5. **SMILES formula**: we wrap sequences with `[START_SMILES]` and `[END_SMILES]` and apply character-based tokenization. Similarly we use `[START_I_SMILES]` and `[END_I_SMILES]` where isomeric SMILES is denoted. For example, `C(C(=O)O)N` → `C,(,C,(,=,O,),O,),N`.
6. **Amino acid sequences**: we wrap sequences with `[START_AMINO]` and `[END_AMINO]` and apply character-based tokenization, treating each amino acid character as a single token. For example, `MIRLGAPQTL` -> `M,I,R,L,G,A,P,Q,T,L`.
7. **DNA sequences**: we also apply a character-based tokenization, treating each nucleotide base as a token, where the start tokens are `[START_DNA]` and `[END_DNA]`. For example, `CGGTACCCTC` -> `C, G, G, T, A, C, C, C, T, C`.

We cover a few of the specialized token approaches below that do not have clear parallels in the literature, in particular the working memory and citation tokens.

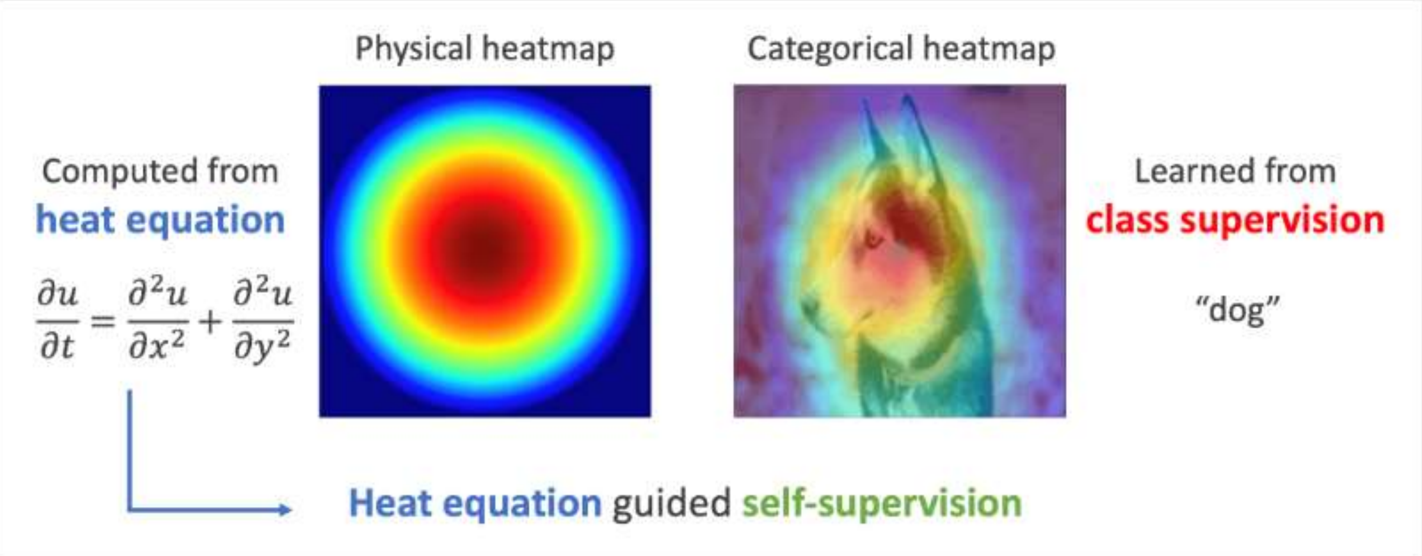
Galactica- A Large Language Model for Science(Lecun et al., 2022).

Challenges： Time Series Mapping to Concepts



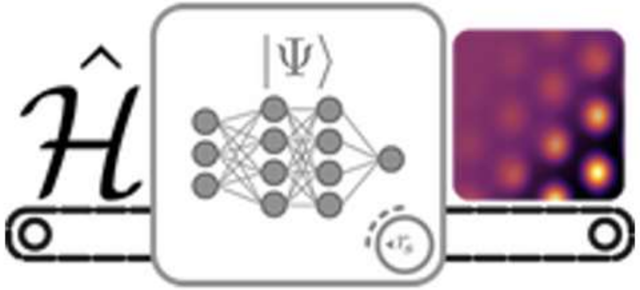
PINN and LLM

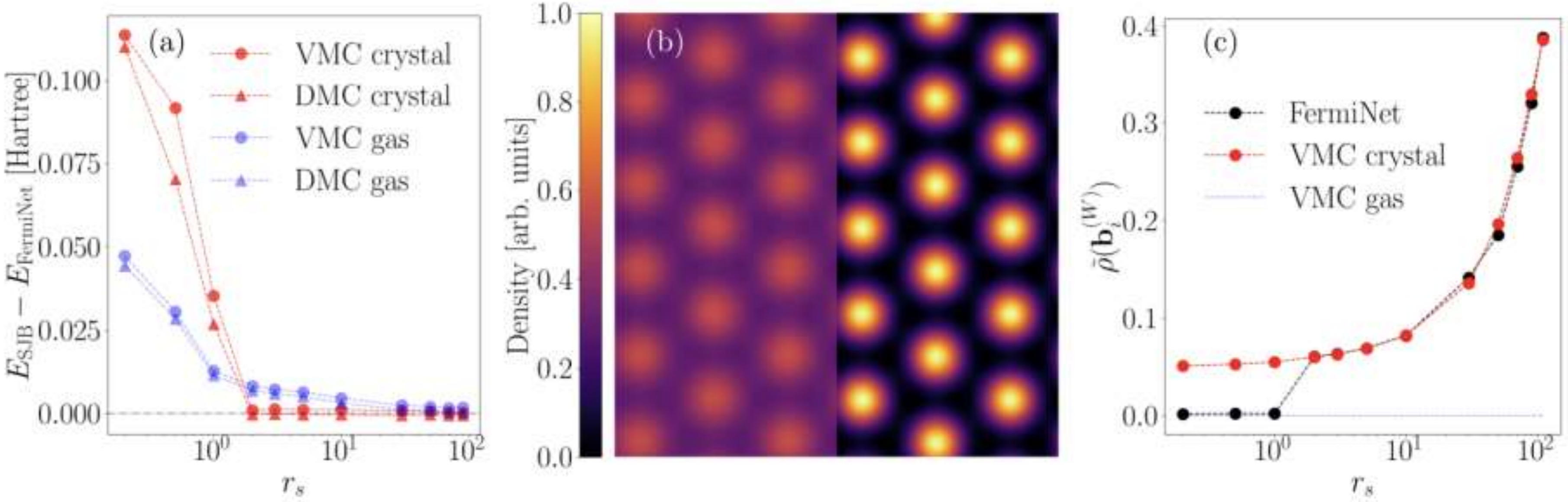
- PINN: physics-inspired neural network
- Unified training of PINN and language model
- Can Language model understand existing physics laws and discover new laws



热方程的热力图（左）和视觉模型的热力图（右）

PINN and LLM





The literature presents an extension of wave function *Ansatze*, FermiNet, a neural network based on a representation of the wave function guided by the quantum mechanical variational principle alone. Without reference to experimental data, FermiNet predicts electronic ground states in condensed matter without a priori knowledge of the system. []

Cassella G, Sutterud H, Azadi S, et al. Discovering quantum phase transitions with fermionic neural networks[J]. Physical Review Letters, 2023, 130(3): 036401.

BBT-Science: LLM for Scientific Discoveries

超对称 Super Symmetry Technologies

BBT-Science is trained on BBT-100B with 100 million scientific papers to enable scientific discoveries on physics, math, chemistry, biology, etc

1. Fast and accurate knowledge search
2. LLM to provide new ideas in the frontier of research areas
3. LLM to provide insights from multidiscipline perspectives

h

you are an experimental particle physicist, now you are writing a feasibility report for the design of electron proton collider of 100 kilometers length submitted to Institute of High Energy Physics, please elaborate the report as detailed as possible.

Here is a draft feasibility report for the design of a 100 km electron-proton collider submitted to the Institute of High Energy Physics:

Title: Feasibility Study for a 100 km Electron-Proton Collider

Abstract: This report assesses the feasibility and provides a conceptual design for a 100 km circular electron-proton (ep) collider proposed to be built in China. The ep collider would collide 15 GeV electrons with 7 TeV protons and provide important new capabilities for studying quantum chromodynamics and searching for new physics beyond the Standard Model.

Introduction: Electron-proton colliders provide unique capabilities that complement proton-proton and electron-positron colliders. By colliding pointlike electrons with the composite protons, ep collisions enable new tests of quantum chromodynamics calculations and provide sensitivity to parton distribution functions in the proton. The proposed 100 km ep collider would be the most powerful ep collider built to date, achieving substantially higher energy and luminosity than existing facilities like HERA at DESY.

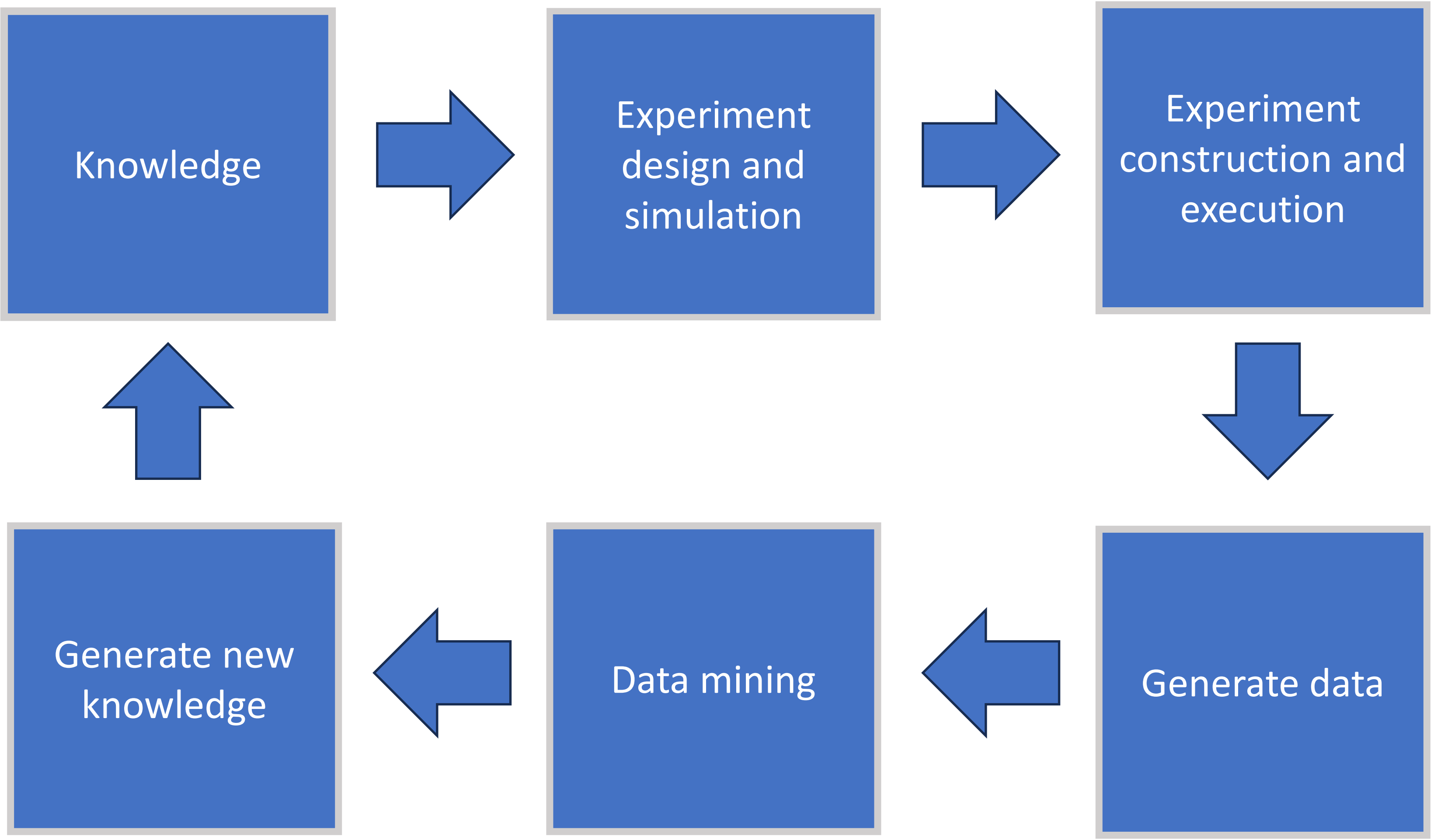
Accelerator Physics: A circumference of 100 km has been selected to provide the desired high collision energy while minimizing synchrotron radiation effects for the 15 GeV electron beam. At this circumference, 7 TeV protons can be accelerated without exceeding magnet and RF system limits. The electron ring will use conventional magnets and RF cavities, while the proton ring will require high-field Nb3Sn magnets operating at 1.9 K and megawatt-class crab cavities to achieve the design luminosity of 10^34 cm-2 s-1. Strong hadron cooling systems will be needed to achieve high proton beam brightness.

Detector Requirements: To take full advantage of the ep collider's physics potential, specialized detectors are needed. A layered detector with excellent spatial resolution and particle identification will be required to reconstruct the complex final states resulting from ep collisions. High precision trackers and calorimeters will be needed to measure jets and missing energy. The detector will need to handle high data rates from the luminous collisions.

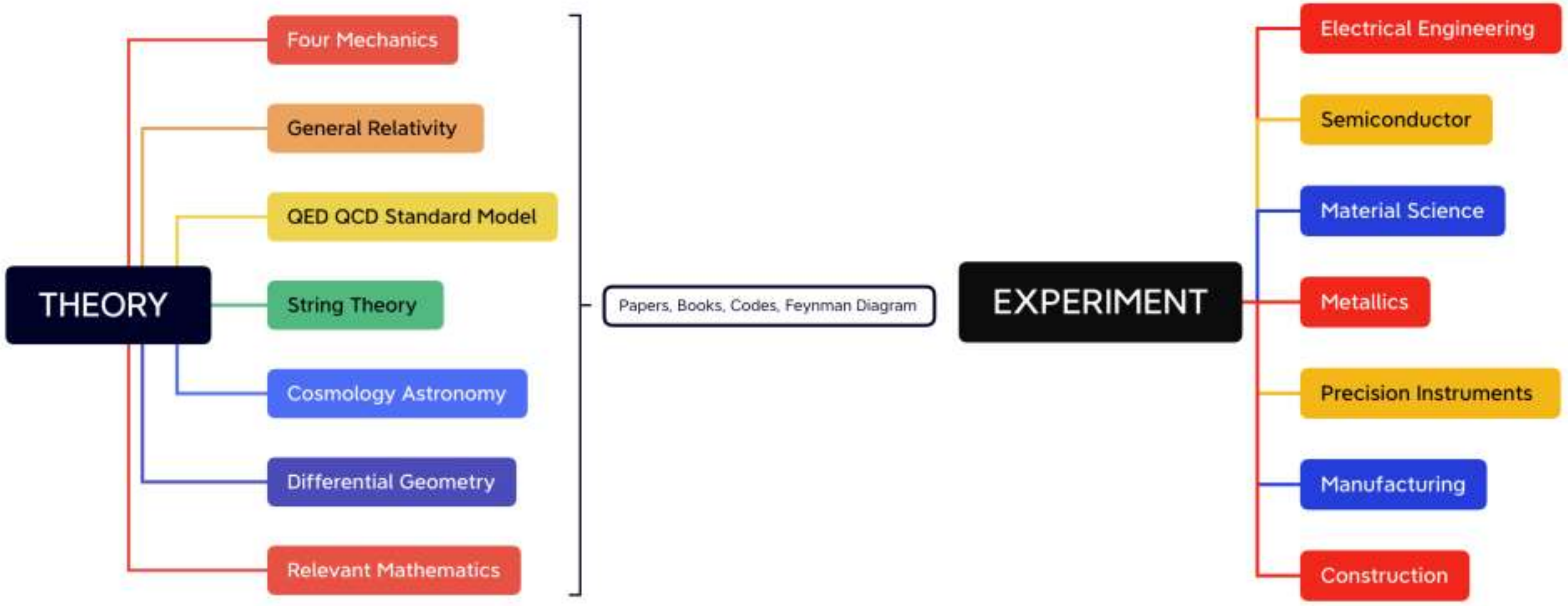
Research Goals: The ep collider will enable world-leading studies of quantum chromodynamics, complementing proton and electron-positron colliders. Key measurements include high-precision tests of parton distribution functions, gluon self-interactions, and spin physics. With its high energy, the ep collider also has discovery potential for new particles and interactions beyond the Standard Model. Searches will be conducted for leptoquarks, dark matter candidates, and other exotic particles.

Technical Challenges: While no insurmountable physics or engineering challenges are foreseen, building a 100 km collider presents significant technical risks that will require detailed R&D and prototype testing. Challenges include achieving ultra-high proton beam brightness, constructing dozens of high-field magnets, handling multi-megawatt beam power, and managing very high data rates from the detector. Careful staging of technical demonstrations will be crucial before proceeding with full construction.

Particle Physics LM: Accelerate Productivity



Particle Physics Knowledge Domains



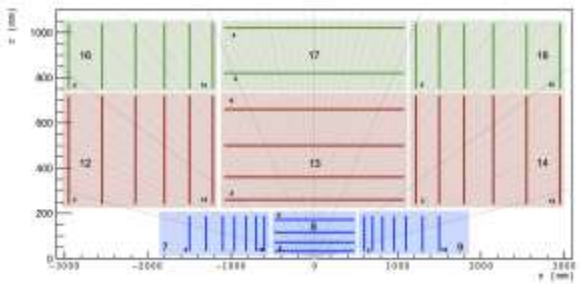
Training Datasets : Particle Corpus

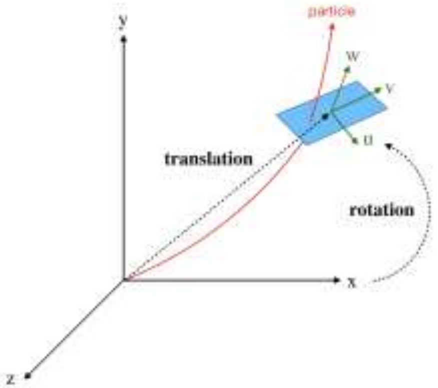
Particle Corpus: Total 20B Tokens				
	Physics	Math	Engineering	Others
Books	10K	5K	10K	10K
Papers	3 millions	1 millions	1 millions	2 millions
Codes	Fortan	C++	ROOT	MATLAB
Tech IP	Superconductors	Cryogenics	PMT	Electronics

Code Interpreter for Data analysis of Reconstruction

超对称 Super Symmetry Technologies

Hit id, event id, track id, module, layer, detector, initial momentum, initial position, particle id

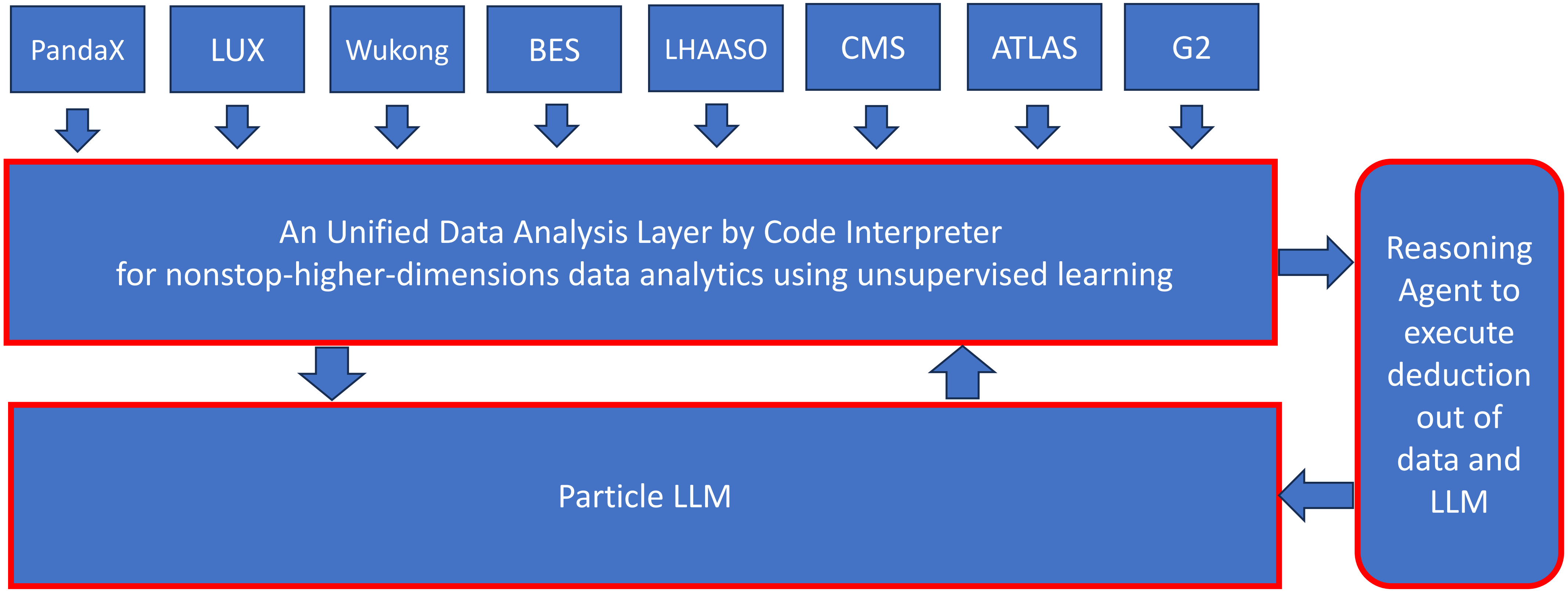




volume_id	layer_id	module_id	cx	cy	cz	rot_xu
7	2	1	-65.7965	-5.1783	-1502.5	0.0784591
7	2	2	-139.851	-6.46568	-1502	0.0461835
7	2	3	-138.657	-19.3419	-1498	0.138156
7	2	4	-64.1764	-15.4074	-1498	0.233445
7	2	5	-136.281	-32.0531	-1502	0.228951
7	2	6	-60.976	-25.2571	-1502	0.382683
7	2	7	-132.742	-44.4908	-1498	0.317791
7	2	8	-128.071	-56.5489	-1502	0.403921
7	2	9	-56.2743	-34.4849	-1497.5	0.522499
7	2	10	-122.307	-68.1246	-1498	0.486604
7	2	11	-115.5	-79.1191	-1502	0.565136
7	2	12	-50.1868	-42.8636	-1502.5	0.649448
7	2	13	-107.707	-89.4386	-1498	0.638847
7	2	14	-42.8636	-50.1868	-1498	0.760406
7	2	15	-98.9949	-98.9949	-1502	0.707107
7	2	16	-89.4386	-107.707	-1498	0.769334
7	2	17	-34.4849	-56.2743	-1502	0.85264
7	2	18	-79.1191	-115.5	-1502	0.824997
7	2	19	-68.1246	-122.307	-1498	0.873622
7	2	20	-25.2571	-60.976	-1497.5	0.92388

Use code interpreter to establish a generic track classifier: group the recorded measurements or hits for each event into tracks, sets of hits that belong to the same initial particle. A solution must uniquely associate each hit to one track.

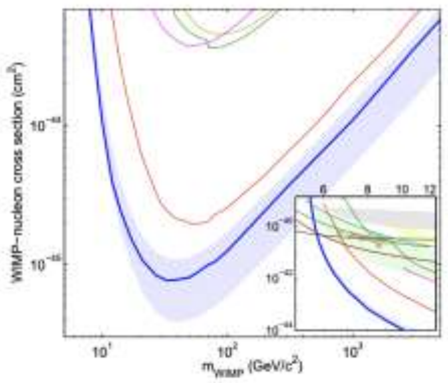
A Unified Data Analysis Layer



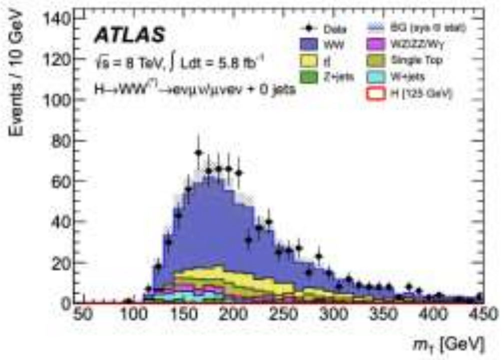
Benchmark for Particle LLM

New Physics

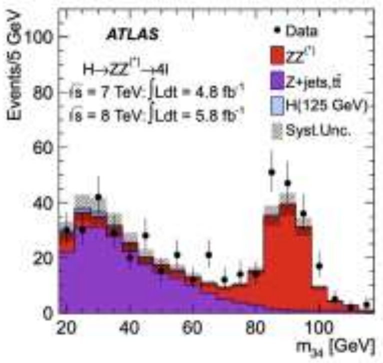
Hadronic Physics

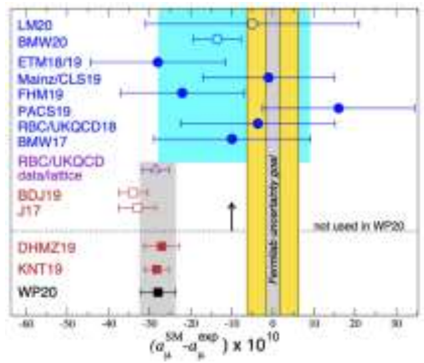


(a)



(b)





Particle LLM Benchmarks :

- Cover all theoretical and experimental particle physics research areas and experiments
- Data analysis result integrated with reasoning of language model
- Anomalies shown in data analysis within the standard model framework
- Gauge by the confidence level of data analysis result
- Bring in peer review for quantifying the performance of LLM

Building the Strongest Reasoning Machine

Framework

- High quality training corpus : scientific books, papers, wiki, IPs
- Benchmark: for new knowledge and innovation, rather than exams
- Reasoning Agent: Chain of Thought to Tree of Thought
- Code interpreter for a unified data analysis layer

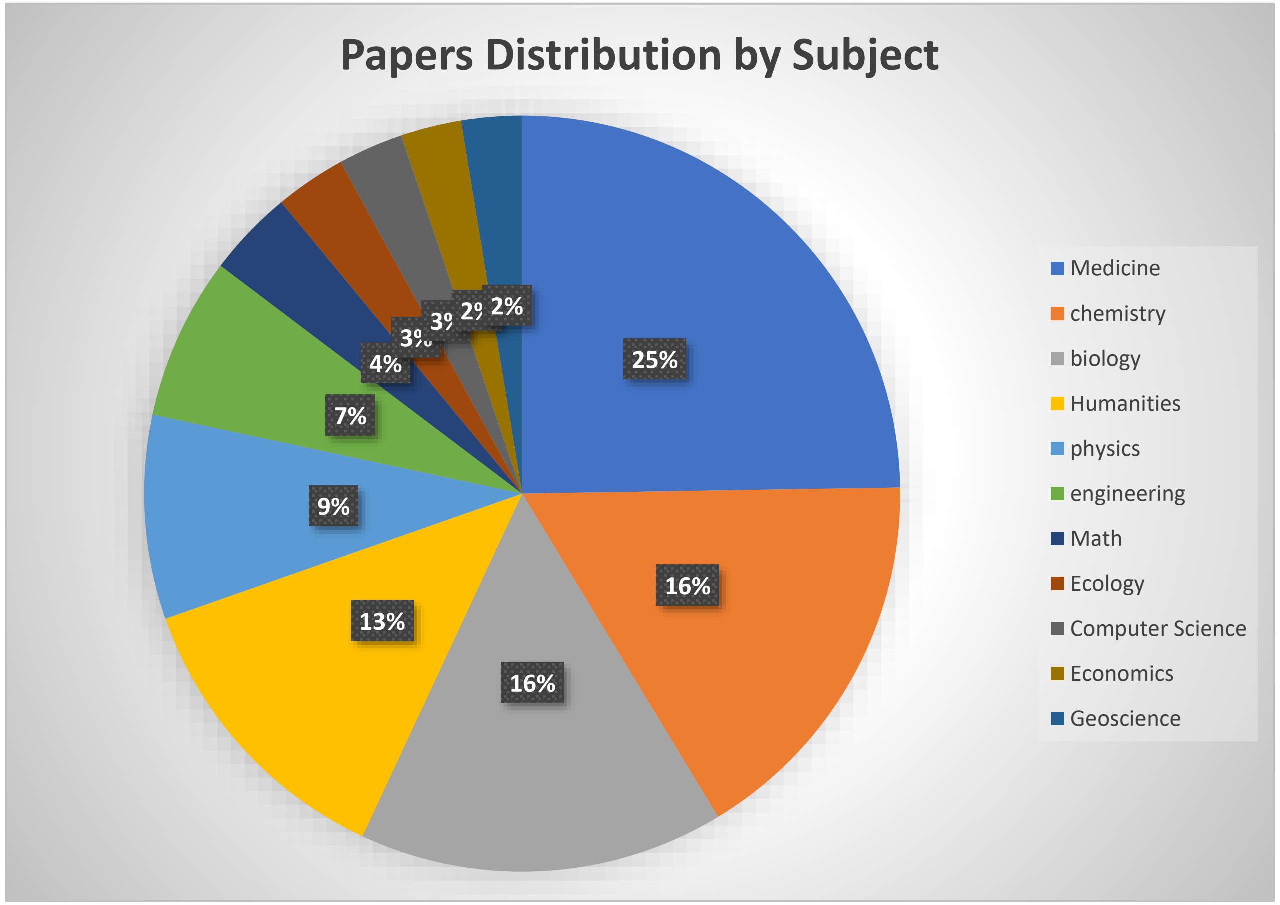
BBT Science Corpus : Collect All Human Generated Knowledge

超对称 Super Symmetry Technologies

- 2 Million books
- 150+ Million papers
- 100 Million IP

Data source	Documents	Tokens
Wikipedia	6 million	5 billion
StackExchange	1.6 million	1 billion
LibreText	95,113	185 million
Wikibooks	74,705	110 million
Open Textbooks	647	94 million
MIT OCW	25,640	90 million
Wikiversity	38,138	52 million
ProofWiki	32,389	12 million
Khan Academy	3,075	7 million
Papers with Code	13,430	4 million
IUPAC Goldbook	6,788	1 million
Total	8 million	7 billion

Data source	Documents	Tokens
arXiv	2 million	35 billion
PMC	3 million	23 billion
Semantic Scholar	3 million	18 billion
PubMed Abstracts	21 million	5 billion
Semantic Scholar Abstracts	19 million	4 billion
bioRxiv	128,059	1 billion
OSF	54,905	428 million
medRxiv	24,019	176 million
ACL	25,518	150 million
PubAg Abstracts	308,235	105 million
ChemRxiv	7,617	67 million
Total	48 million	88 billion

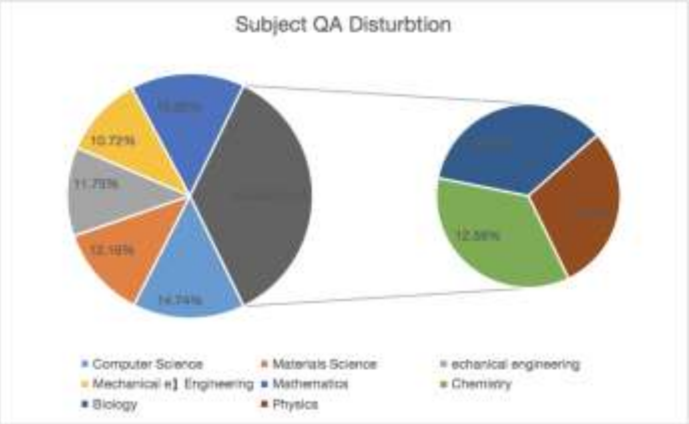


Push the Boundary of Human Knowledge: 1M Unsolved Problems

- We are building a benchmark dataset consisting of
  1 million science, math, technology, engineering unsolved problems
- Computing = Knowledge
- Benchmark by peer reviews

Boundary Dataset:1 Million STEM Unsolved Problems

Scicence	Math	Engineering	Technologies
Physics	Number Theory	Computer Science	Manufacturing
Chemistry	Algebra		Biotech
Biology	Geometry	Electrical Engineering	Renewable Energy
Medicine			Aerospace Tech
			Transportation



Conclusions

- LLM for science experiments: search, code and autonomous experiments
- LLM for science discoveries: model physical reasoning process
- Challenges: special tokens
- Challenges: time series data to concepts
- Challenges: benchmark Boundary data sets
- BBT-Science: Foundational LLM model for science discovery