Emergent phenomena in large-scale LLMs

Daniel Khashabi

Johns Hopkins University

The success we dreamed of

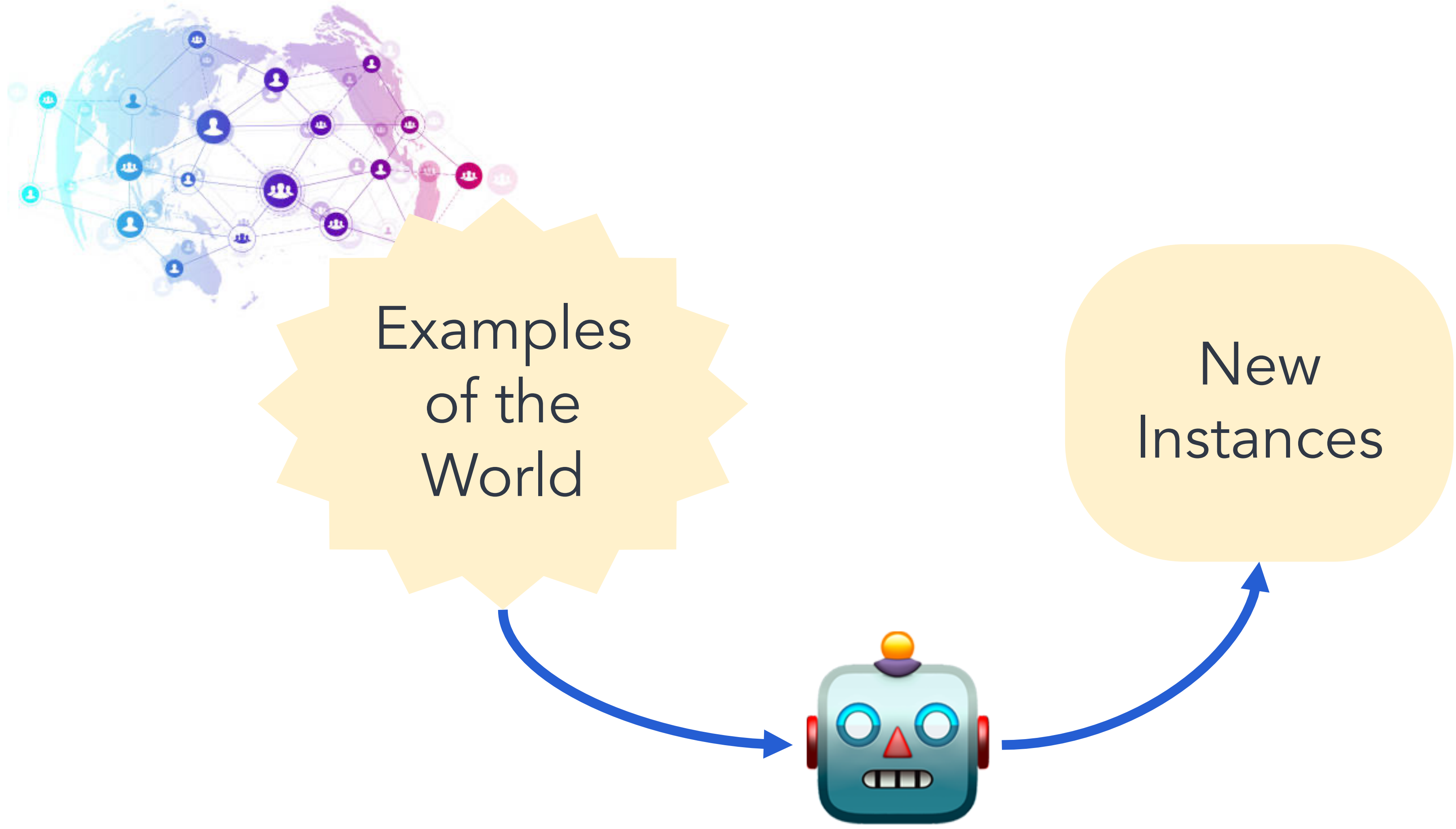
Language models that **reason**
across many tasks that were once considered aspirational.

Google Gemini

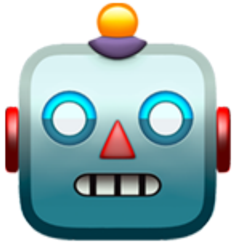
Claude
BY ANTHROP\C

How do these models work?

Learning ~ Compression + Replication

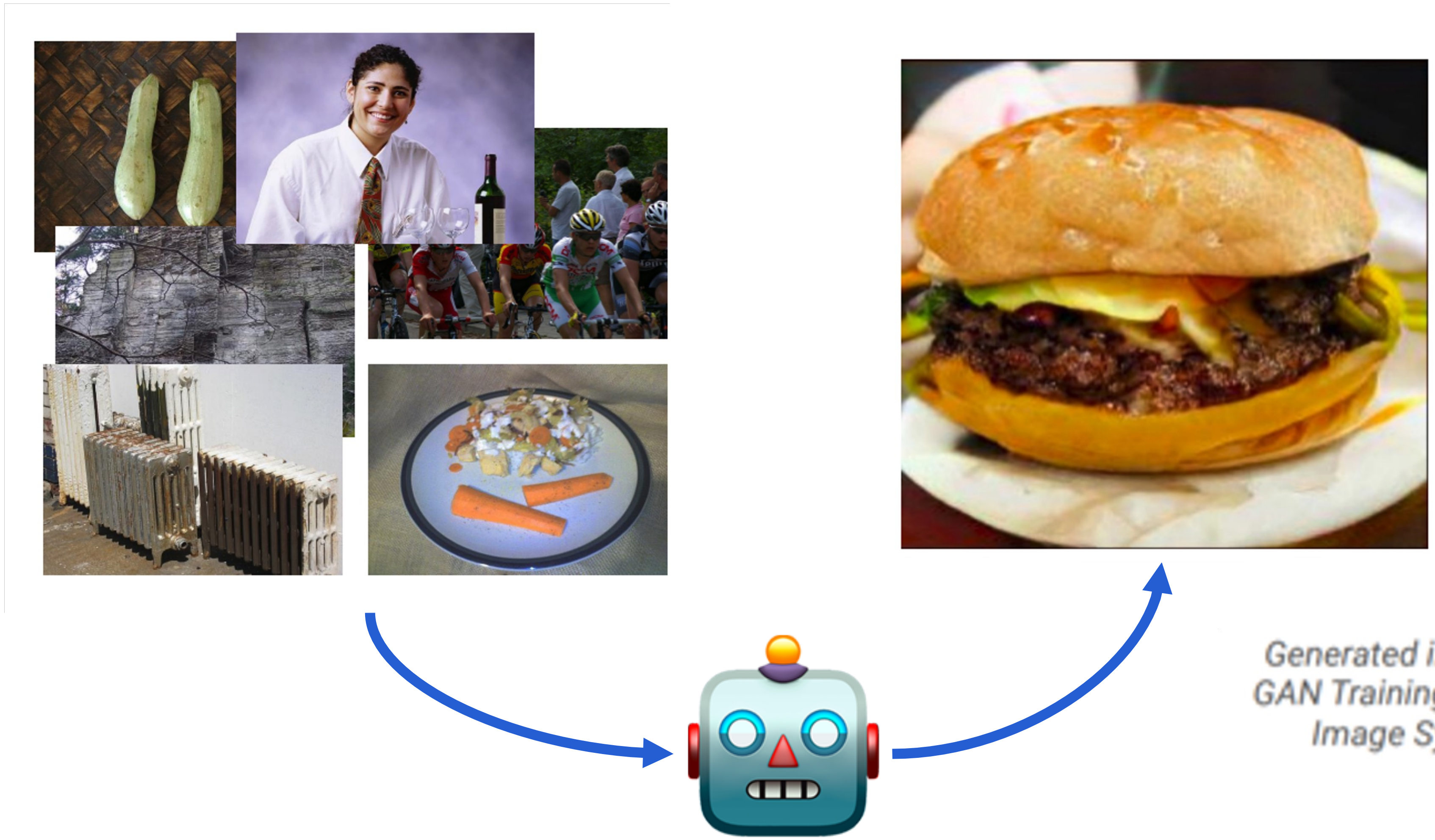


Learning ~ Compression + Replication



New Instances

Learning ~ Compression + Replication



Generated image, from "Large Scale GAN Training for High Fidelity Natural Image Synthesis", Brock et al.

Learning ~ Compression + Replication

== tr

the tr
the tr
10 fe
of gr
with
great
and s
war.

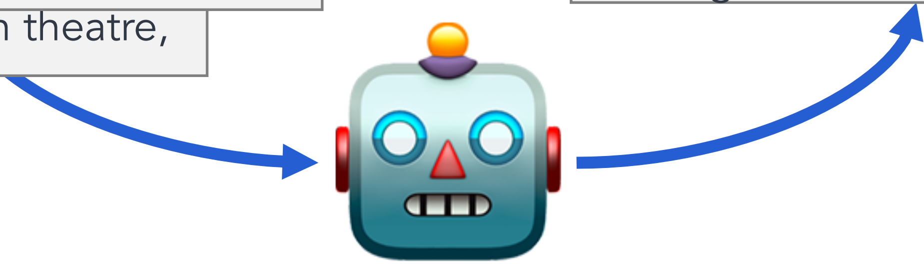
the s
ende
know
war in the north american theatre,
....

== w

A whe
prope
one w
pushe
perso
rear,
ancie
term
two w
"Barr
Old E
devic
The wheelbarrow is designed to

== lemon

The lemon (Citrus limon) is a species of small evergreen trees in the flowering plant family Rutaceae, native to Asia, primarily Northeast India (Assam), Northern Myanmar or China.[2] The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.[2] The pulp and rind are also used in cooking and baking. ....

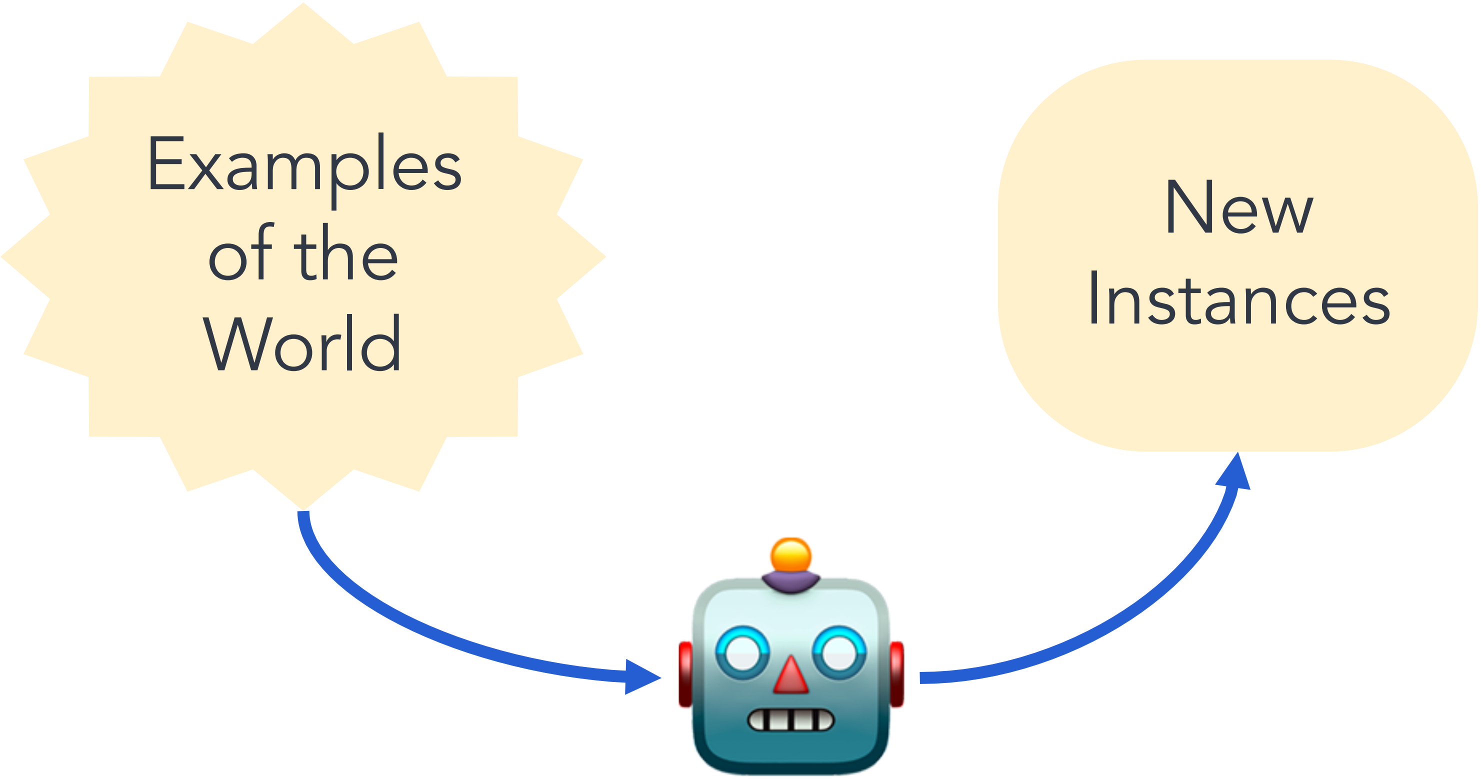


== wings over kansas

wings over kansas is the second studio album by jason ammons, john bolster and mo rosato. the album debuted at number one on the billboard 200, selling 35,000 copies in it first week at the time. it was the second highest selling album to debut at the billboard top 50 and the third highest selling album to debut at the top heatseekers, with 26,000 copies sold. this is the supremes album earning the nickname …

Learning ~ Compression + Replication

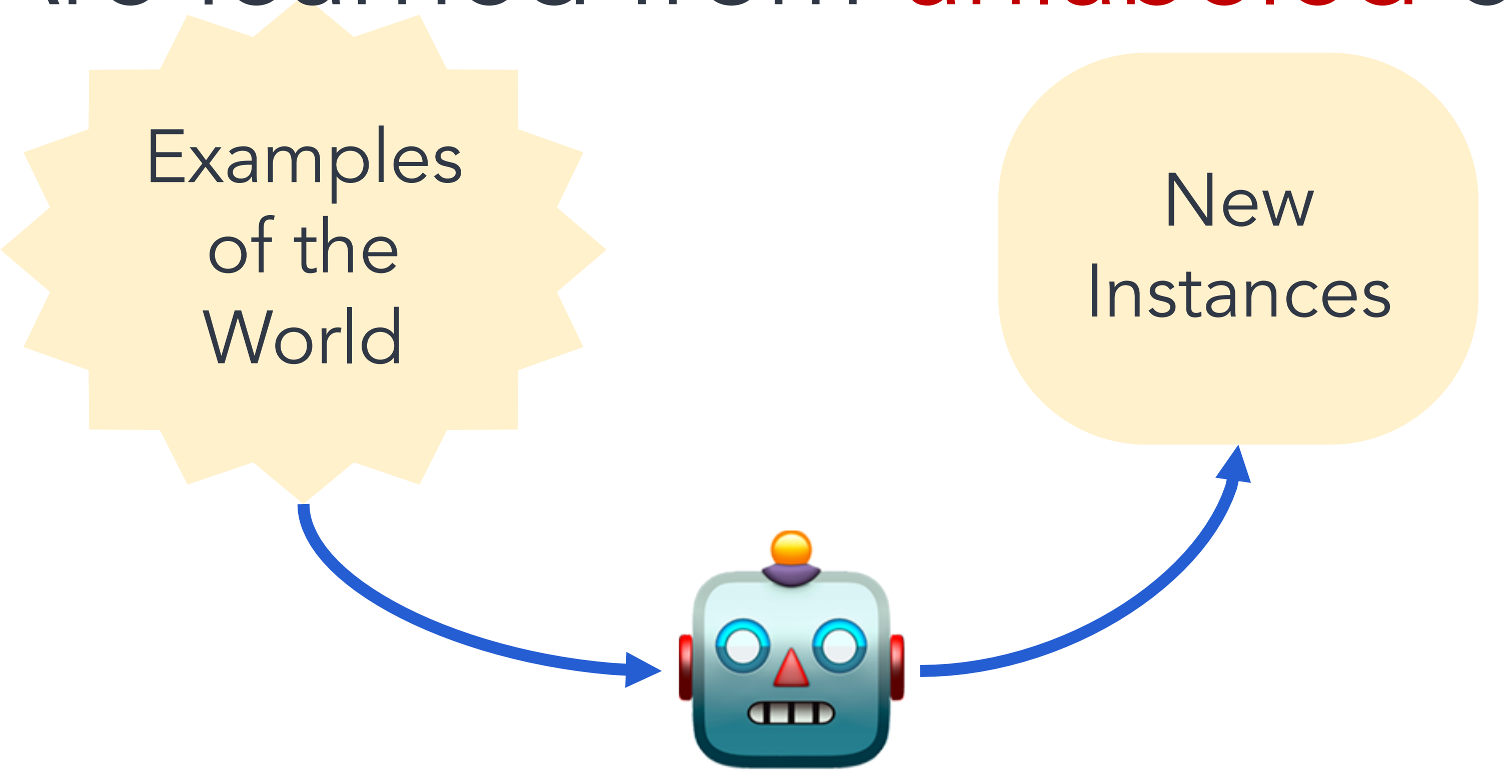
Are predictive models of the world.



Learning ~ Compression + Replication

Are predictive models of the world.

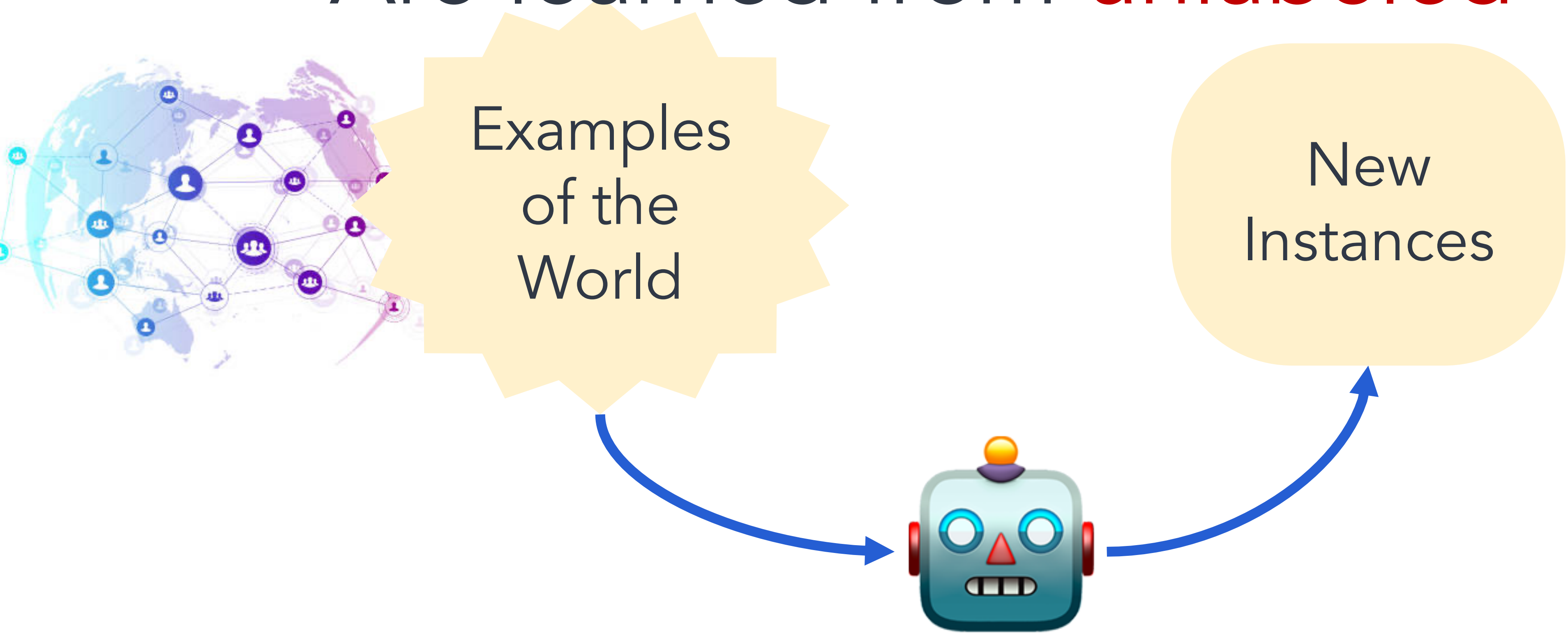
Are learned from unlabeled data.



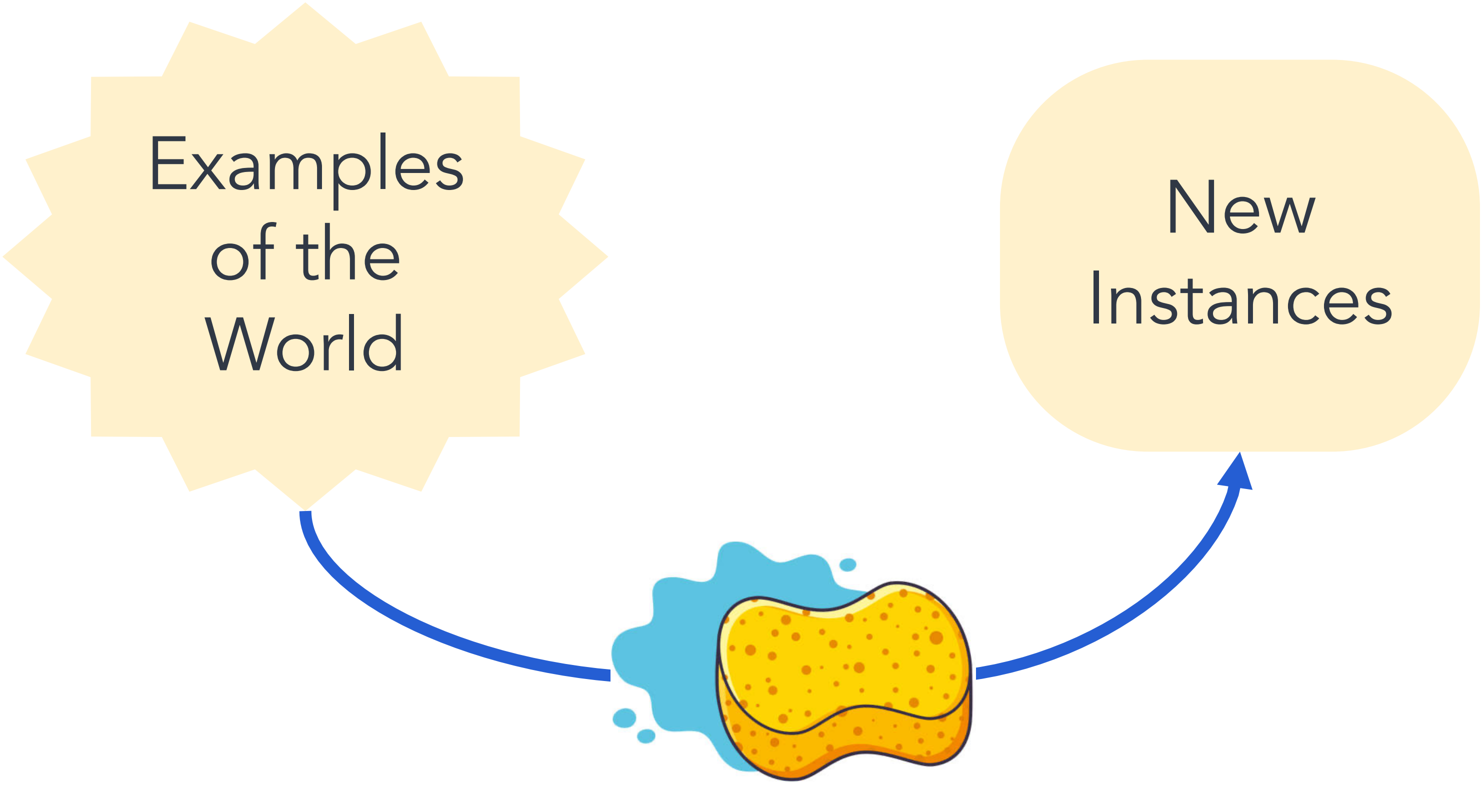
Learning ~ Compression + Replication

Are predictive models of the world.

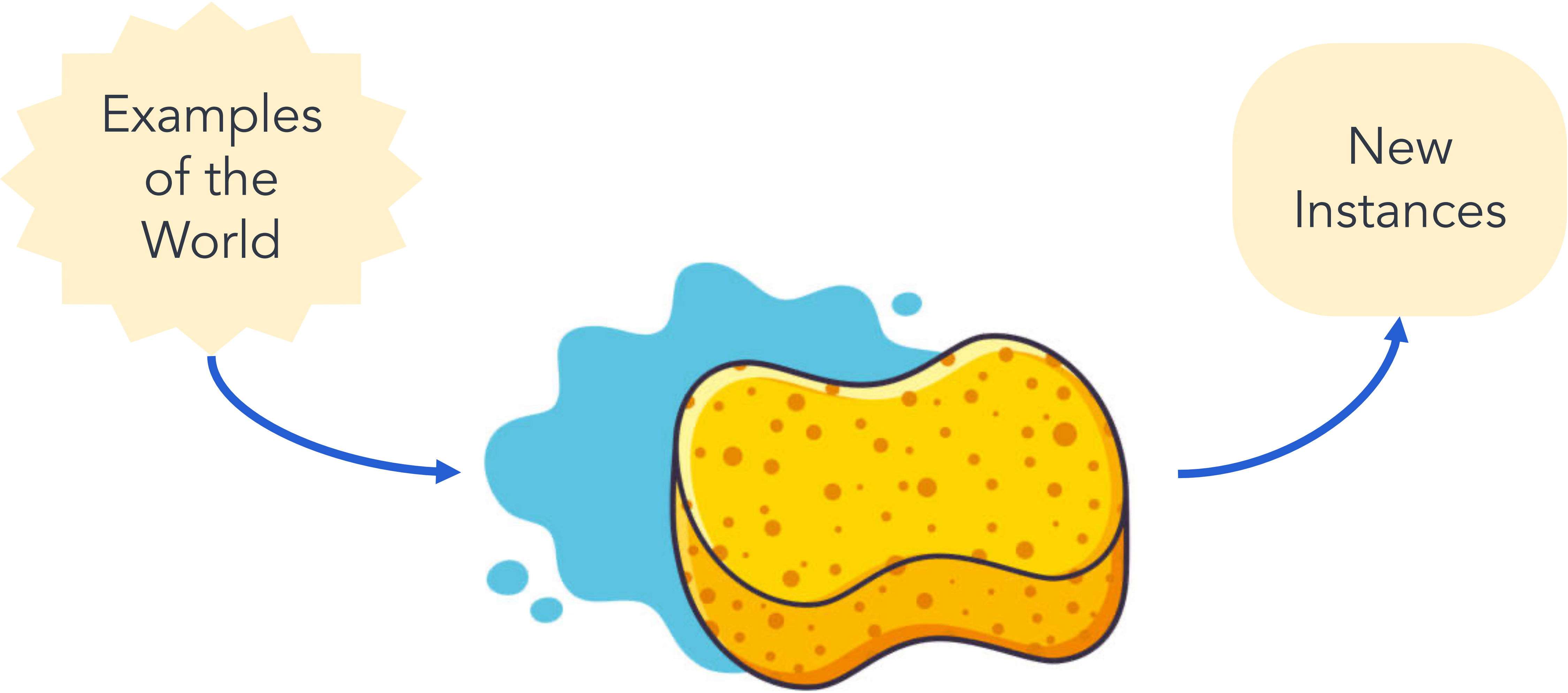
Are learned from unlabeled data



AI Models ~ Sponges

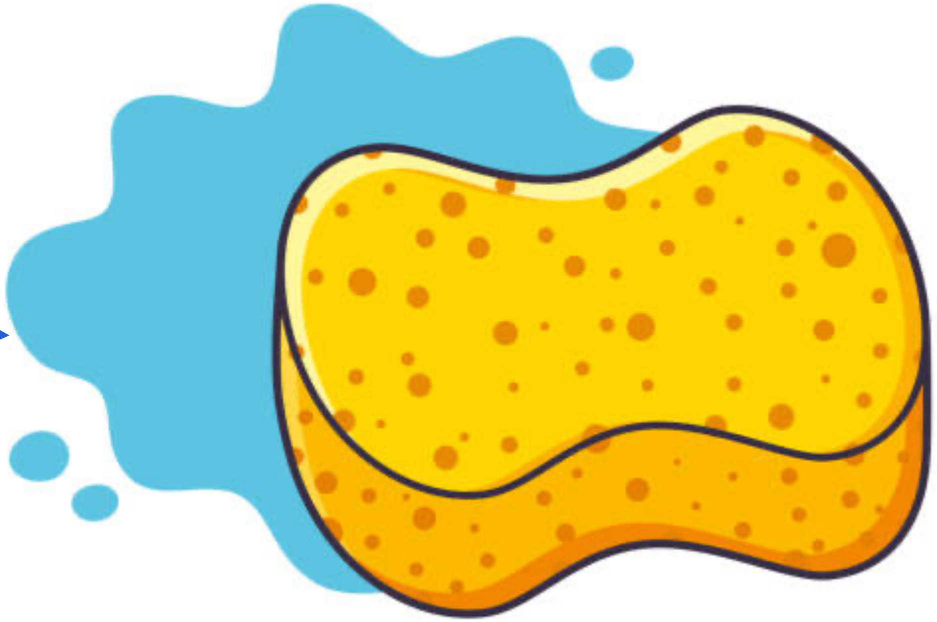


AI Models ~ Sponges



AI Models ~ Sponges

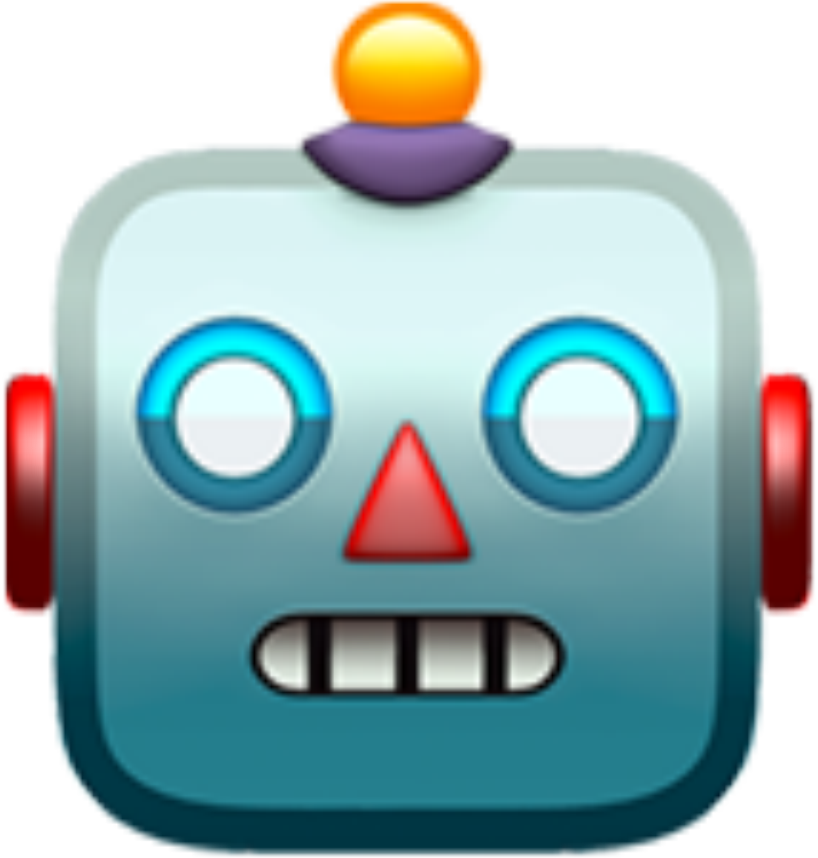
Examples of the World



New Instances

Scaling up AI Models Makes it Better

Examples of the World



New Instances

Today: Emergent phenomena

- What is happening inside language models?

- What do we mean by “emergent phenomena”?

- What do we understand about it?

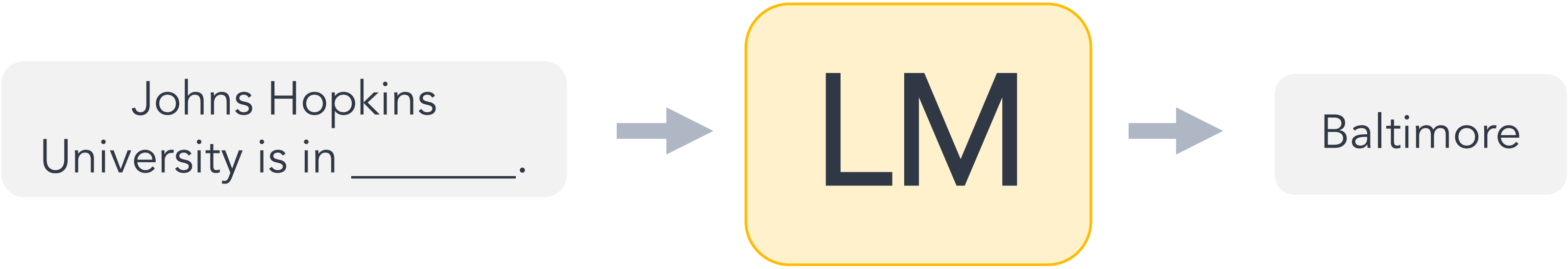
- What does the future hold?

Today: Emergent phenomena

- What is happening inside language models?
- What do we mean by “emergent phenomena”?
- What do we understand about it?
- What does the future hold?

Recipe: (1) Pre-training

- **Pre-training** := Next-Token Prediction

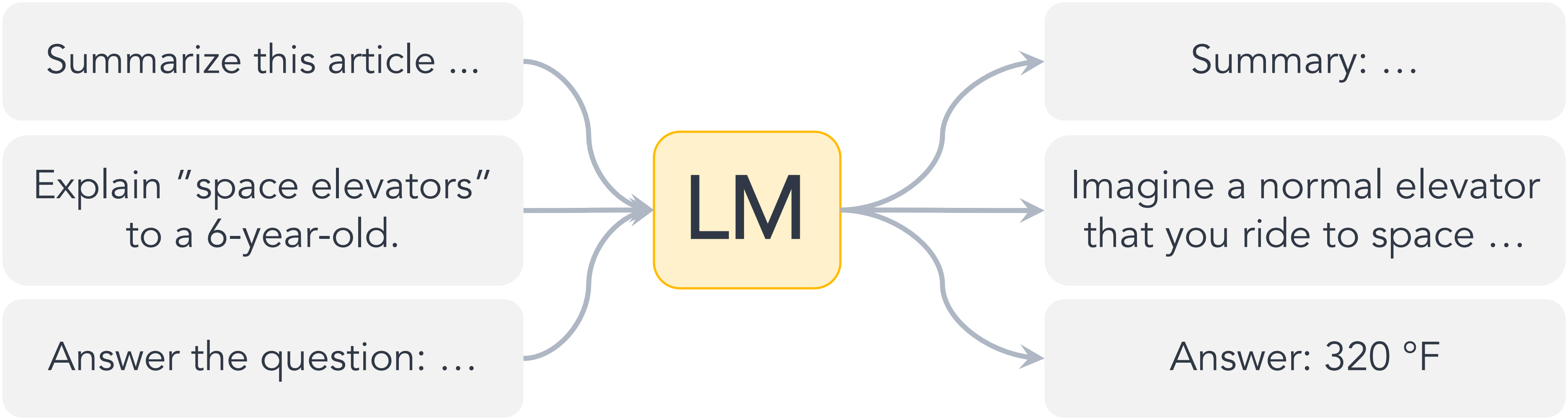


- Modern LLMs are pre-trained on many-trillions of tokens.
  - For comparison, if you read **1 book per week**, you'd need 50,000 years to cover a comparable amount of text!!!

Bengio et al. '04, Peters et al. '18, Raffel et al. '20, Brown et al. '20, many others

Recipe: (2) Post-training

- “Aligning” LMs to respect our intents embedded in instructions.
  - Supervised Fine-tuning (SFT/behavior cloning) with labeled data.

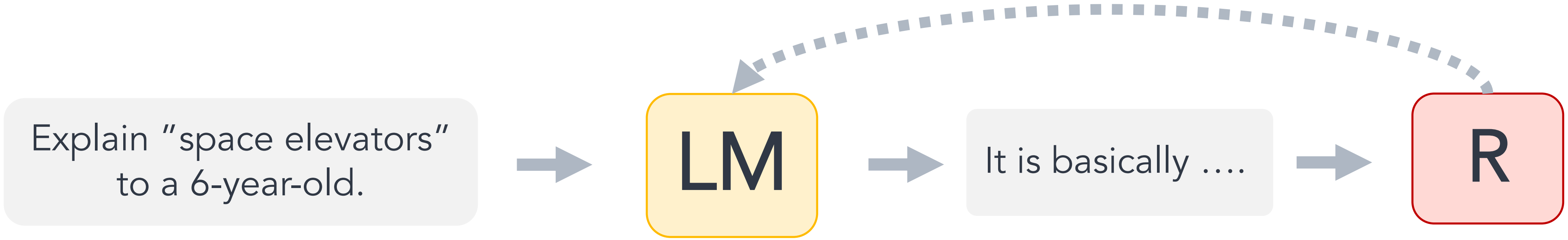


Cross-Task Generalization via Natural Language Crowdsourcing Instructions.
Mishra, Khashabi, Baral and Hajishirzi. *ACL* 2021 🥇Ai2 Lasting Impact Award!🥇

Among others: Sanh et al. 2022; Chung et al. 2022, … .

Recipe: (2) Post-training

- "Aligning" LMs to respect our intents embedded in instructions.
  - Supervised Fine-tuning (SFT/behavior cloning) with labeled data.
  - Reinforcement Learning on preference data or verifiers.



Christiano et al. 2017; Stiennon et al. 2020; Ouyang et al., 2022; many others

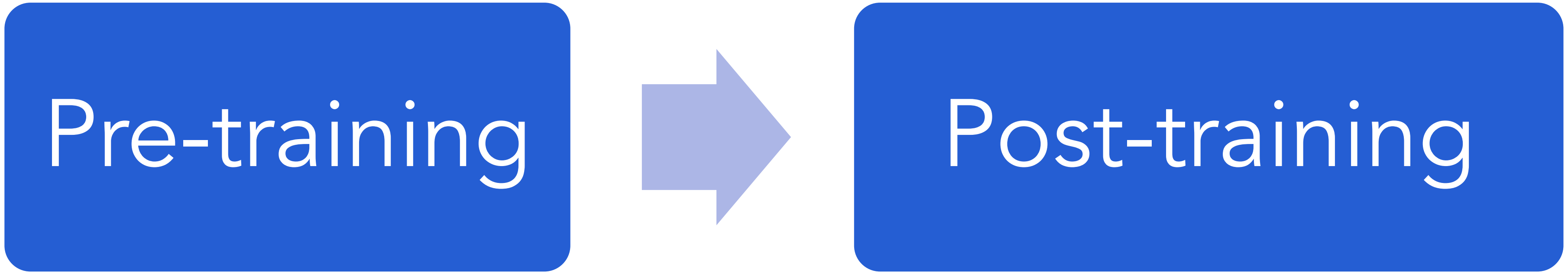
Scaling Recipe: The whole thing

- Almost all the reasoning models follow this recipe:

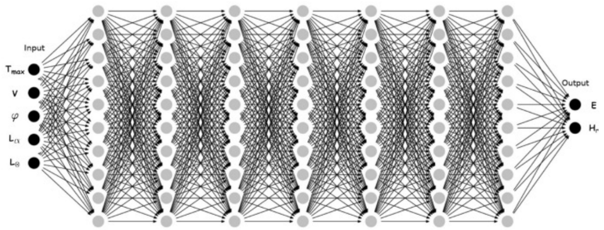
Pre-training → Post-training

Demo time: a pre-trained model

- https://gist.github.com/danyaljj/3ac2f34651d6d9dd4b422daa09dd282a



Since 1950’s



Neural Networks

Since 1980’s

Massive Data

Fast Computers

Since 1940’s
(GPUs 1990’s)

Our overnight success was 70 years in the making!

Today: Emergent phenomena

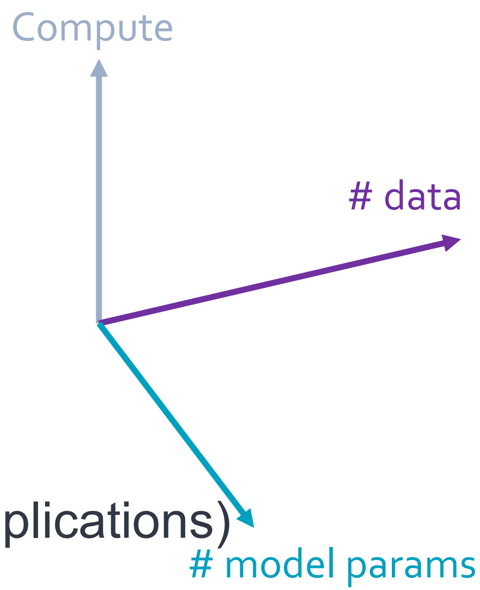
- What is happening inside language models?

- What do we mean by “emergent phenomena”?
- What do we understand about it?
- What does the future hold?

Today: Emergent phenomena

- What is happening inside language models?
- What do we mean by “emergent phenomena”?
- What do we understand about it?
- What does the future hold?

What is “Scaling” pre-training LLMs?

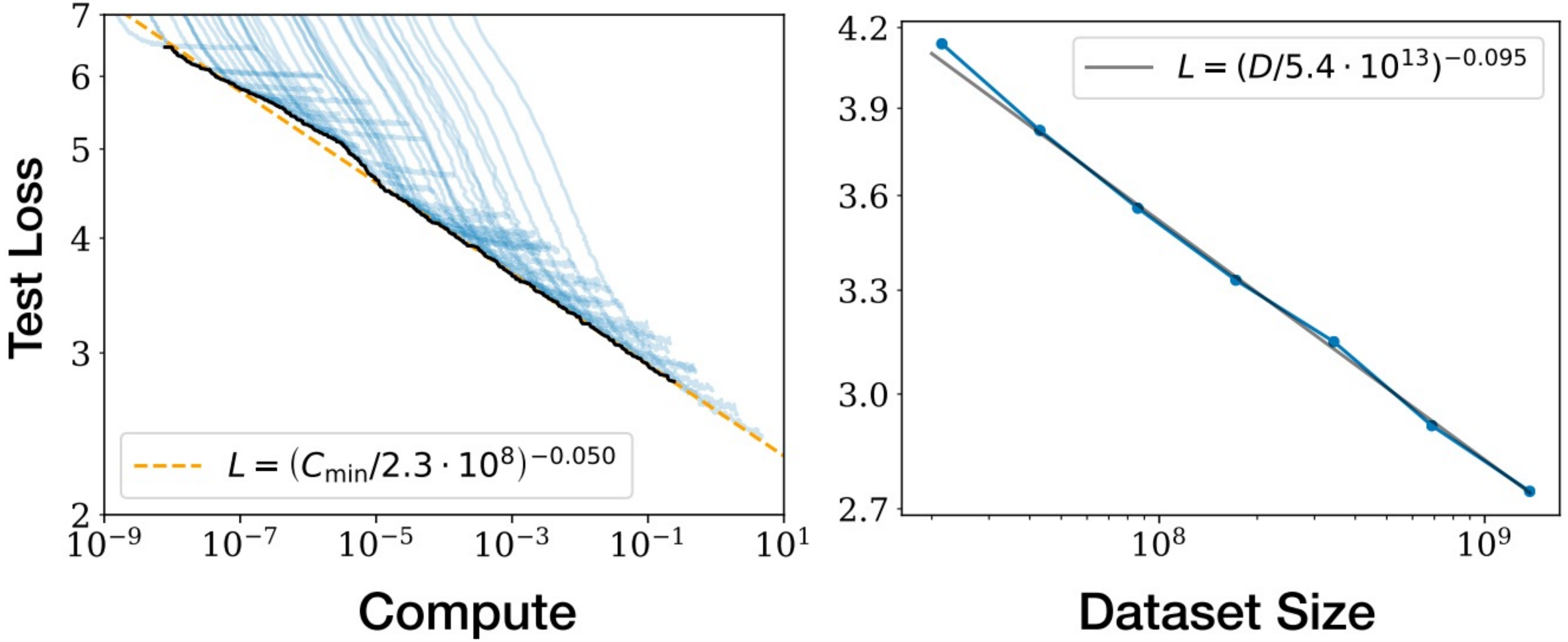


- “Scaling” models is not one thing. There are various axes.

- **Larger model size**
  - But model params may be under-utilized.
- **More data**
  - But more data might not necessarily contain more information (e.g., duplications)
- **More compute**
  - But unnecessary computation may be wasted.

- Scale means all the above: *effective compression of information*
  - Requires model capacity
  - Requires large, rich data
  - Requires compute

Pre-training: Effect of scale



Kaplan et al. 2020; among others

More data and compute ⇒ better results.

More of the same produces something qualitative different

- Anderson ’72 argued that complex systems can behave in ways that can't be understood by *the laws that govern their microscopic parts*.
- He argued that new properties appear at each level of complexity.


4 August 1972, Volume 177, Number 4047

SCIENCE

More Is Different

Broken symmetry and the nature of the hierarchical structure of science.

P. W. Anderson

The reductionist hypothesis may still be a topic for controversy among philosophers, but among the great majority of active scientists I think it is accepted

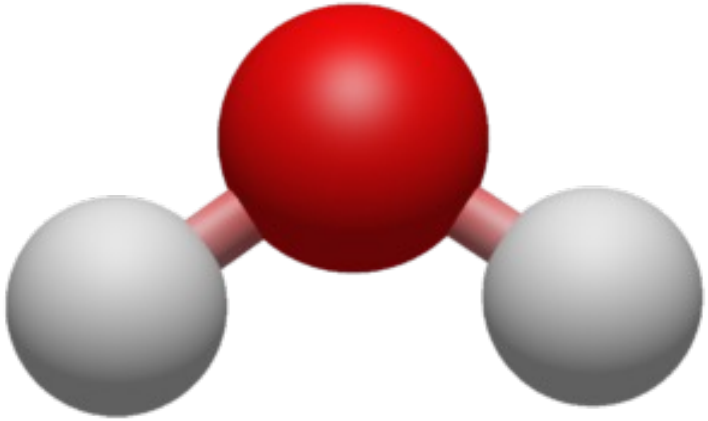
planation of phenomena in terms of known fundamental laws. As always, distinctions of this kind are not unambiguous, but they are clear in most cases. Solid

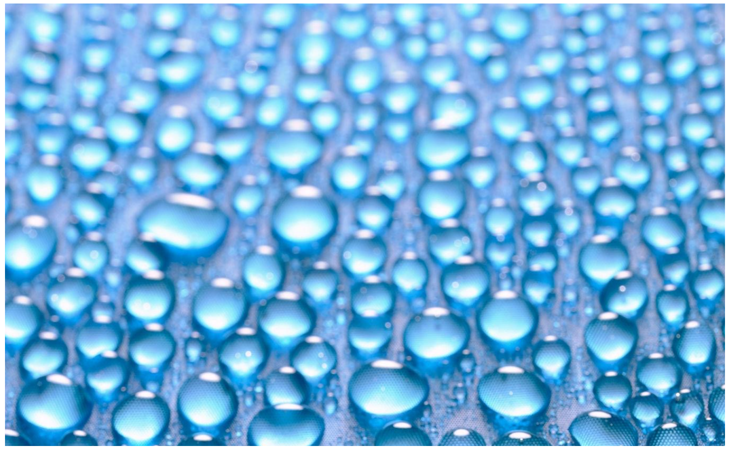
less relevance they seem to have to the very real problems of the rest of science, much less to those of society.

The constructionist hypothesis breaks down when confronted with the twin difficulties of scale and complexity. The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of a simple extrapolation of the properties of a few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other. That is, it seems to me that one may array the sciences roughly linearly in a hierarchy, according to the idea: The elementary entities of science X obey the laws of science Y.

“More is Different”: Examples

- No single molecule is wet.
- When you put billions of them together, they’re wet.
- Although chemistry is subject to the laws of physics, we can’t (easily) infer chemistry from our knowledge of physics.





“More is Different”: Examples

- No ants have grand plans. (to our knowledge)
- Their colony builds cathedrals.

Why does this matter?

- **Unpredictability:**
  - Testing a smaller model doesn’t tell you what the larger model can do.

- **Safety implications:**
  - Standard pre-deployment testing assumes you know what to test for.
  - Emergent abilities violate that assumption.

- You can’t test what you don’t know.

Today: Emergent phenomena

- What is happening inside language models?
- What do we mean by “emergent phenomena”?
- What do we understand about it?
- What does the future hold?

Today: Emergent phenomena

- What is happening inside language models?
- What do we mean by “emergent phenomena”?
- What do we understand about it?
- What does the future hold?

A working example

A working example

Input: NYU

A working example

Input: NYU   Output: NYC

A working example

Input: NYU   Output: NYC
Input: UMD

A working example

Input: NYU Output: NYC
Input: UMD Output: DC

A working example

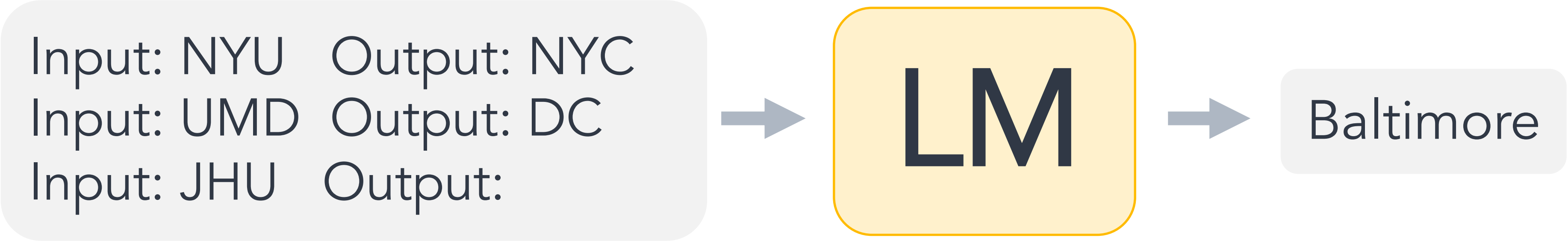
Input: NYU   Output: NYC
Input: UMD Output: DC
Input: JHU

A working example

Input: NYU Output: NYC
Input: UMD Output: DC
Input: JHU Output:

In-Context Learning

- **ICL** := learning to imitate the *implicit* pattern described by few examples provided in the context.



- This is very interesting since we didn't explicitly train our models to show this ability!! (it emerged from large-scale pre-training).
- The models *infer* the patterns from their input.

Originally coined by Brown et al., 2020

Demo time: ICL

- https://gist.github.com/danyaljj/3ac2f34651d6d9dd4b422daa09dd282a

Wait — this shouldn't work

- We only trained the model to predict the next token
- We never showed it "here are examples, learn the pattern"
- Yet at scale, it just… does
- No gradient updates. No fine-tuning. Just prompting.

The big open question:
Why (and When) does ICL emerge?

- For years, the organic emergence ICL (from 'next-token' training) is deeply tied to **human language**.
- Perhaps human language has specific properties that yields ICL.
  - Compositionality, abstraction, parallel patterns, etc.

A Theory of Emergent In-Context Learning as Implicit Structure Induction

Michael Hahn
Saarland University
i-saarland.de

Navin Goyal
Microsoft Research India
navingo@microsoft.com

Parallel Structures in Pre-training Data Yield In-Context Learning

Yanda Chen[1] Chen Zhao[2,3] Zhou Yu[1] Kathleen McKeown[1] He He[2]
[1]Columbia University, [2]New York University, [3]NYU Shanghai

{yanda.chen, kathy}@cs.columbia.edu, cz1285@nyu.edu
zy2461@columbia.edu, hehe@cs.nyu.edu

The big open question:

Why (and When) does ICL emerge?

- For years, the organic emergence ICL (from 'next-token' training) is deeply tied to **human language**.
- But … is it really about human language specifically?

Research question: Is ICL a [human] language-specific phenomenon?

Human vs. Genetic Language

Human language

```
The computer science department at JHU is known for ...
```

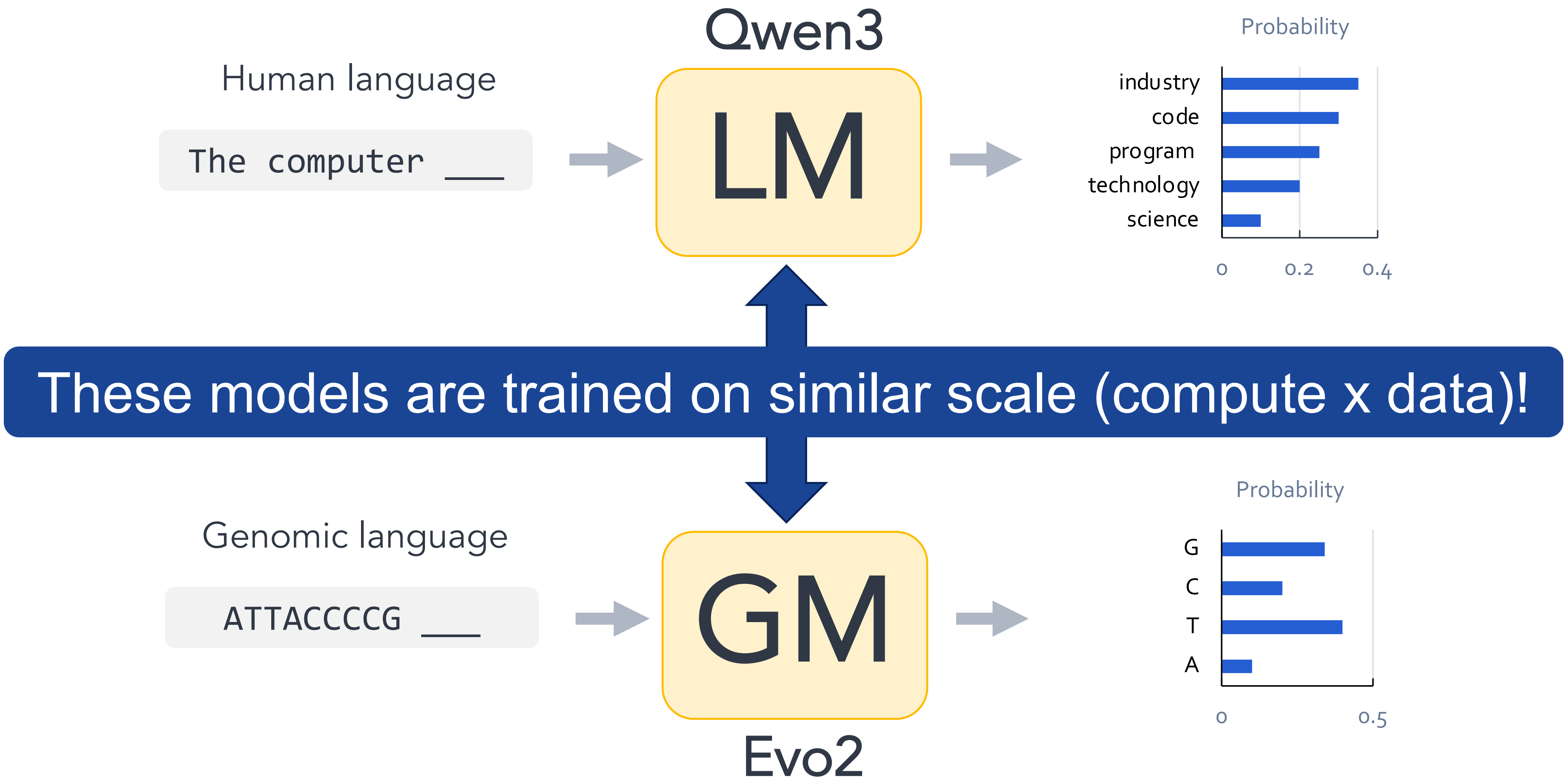
Genomic language

```
ATTACCCCGATTGCTATGCCTGAGAAGCTATTATGCCTGAGAAGCTAT ...
```

- **Similar:** Both are symbolic, short- and long-range dependencies.
- **Different:** Vocabularies, interpretability, evolutionary path.

Genomic Next-Token Predictors are In-Context Learners.
Beslow, Mishra, Revsine, Schatz, Liu, Khashabi. TMLR. 2026.

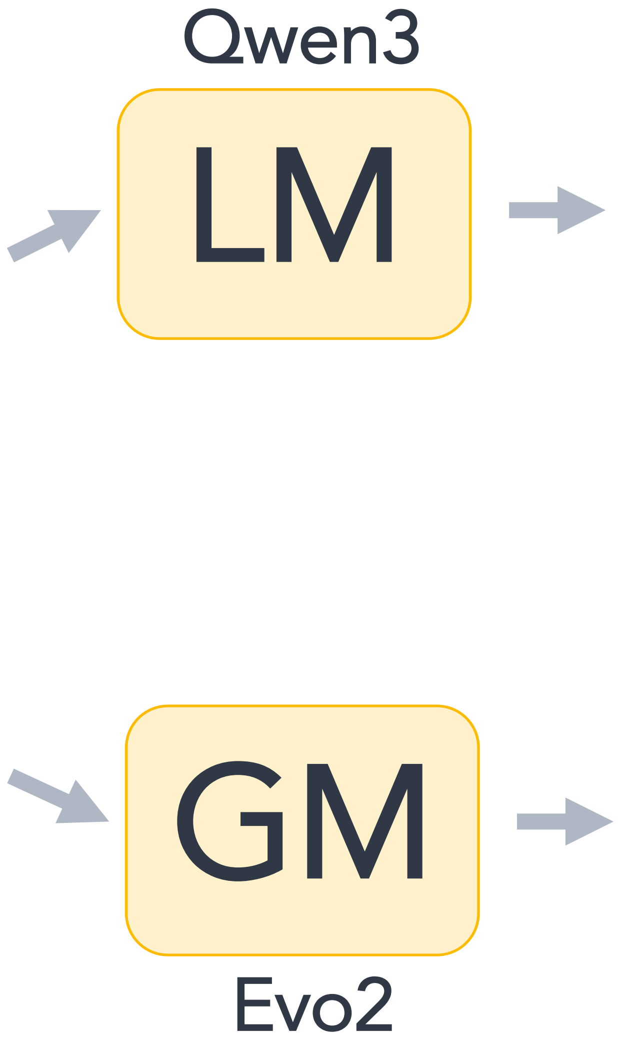
Hypothesis: If ICL appears in both settings,
it must reflect a *more general computational mechanism*.



Genomic Next-Token Predictors are In-Context Learners.
Beslow, Mishra, Revsine, Schatz, Liu, Khashabi. TMLR. 2026.

What task should we use for evaluation?

```
input1 -> Output1 SEP
input2 -> Output2 SEP
input3 -> Output3 SEP
Input4 -> ?
```



Genomic Next-Token Predictors are In-Context Learners.
Beslow, Mishra, Revsine, Schatz, Liu, Khashabi. TMLR. 2026.

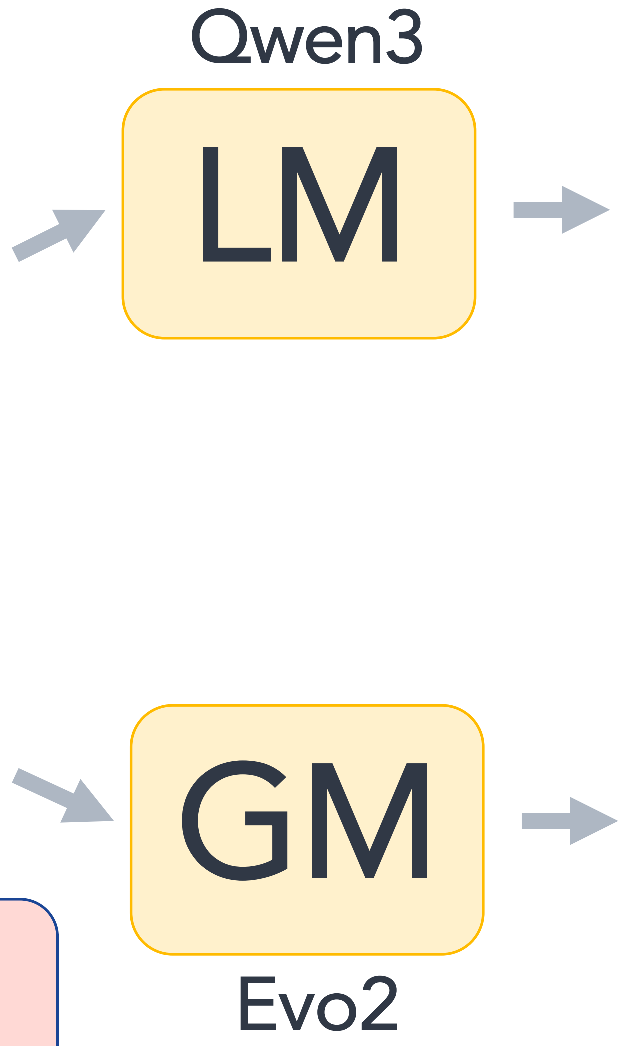
What task should we use for evaluation?

We defined 100 reasoning tasks
based on bitstrings

```
10100000 -> 00000101 SEP
11100011 -> 11000111 SEP
11001110 -> 01110011 SEP
11000000 -> ?
```

Various functions: Bitwise NOT, Reverse, etc.

But we need to express the *same underlying task* in a language each model understands.



Learners.
Beslow, Mishra, Revsine, Schatz, Liu, Khashabi. TMLR. 2026.

What task should we use for evaluation?

- *Replace "1" with "3"*
- *Replace "0" with "4"*
- *Replace "SEP" with "6"*
- *Drop "->"*

```
34344444444443436333
44433334443336334433
3443334433633444444
```

Qwen3

LM

```
10100000 -> 00000101 SEP
11100011 -> 11000111 SEP
11001110 -> 01110011 SEP
11000000 -> ?
```

- *Replace "1" with "T"*
- *Replace "0" with "A"*
- *Replace "SEP" with "G"*
- *Drop "->"*

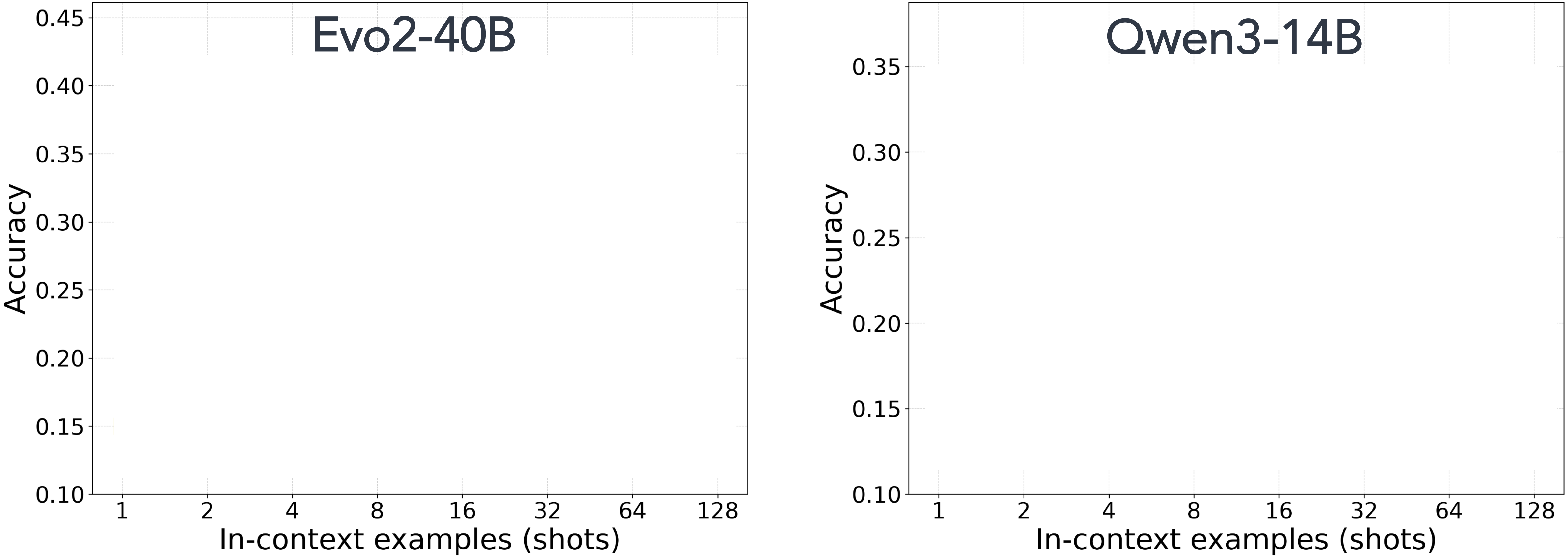
```
TATAAAAAAAAAATATGTTT
AAATTTTAAATTTGTTAATT
TAATTTAATTGTTAAAAAA
```

GM

Evo2

Genomic Next-Token Predictors are In-Context Learners.
Beslow, Mishra, Revsine, Schatz, Liu, Khashabi. TMLR. 2026.

Genomic Models are In-Context Learners



Both models exhibit log-linear gains in pattern induction as the number of in-context demonstrations increases.

Summary

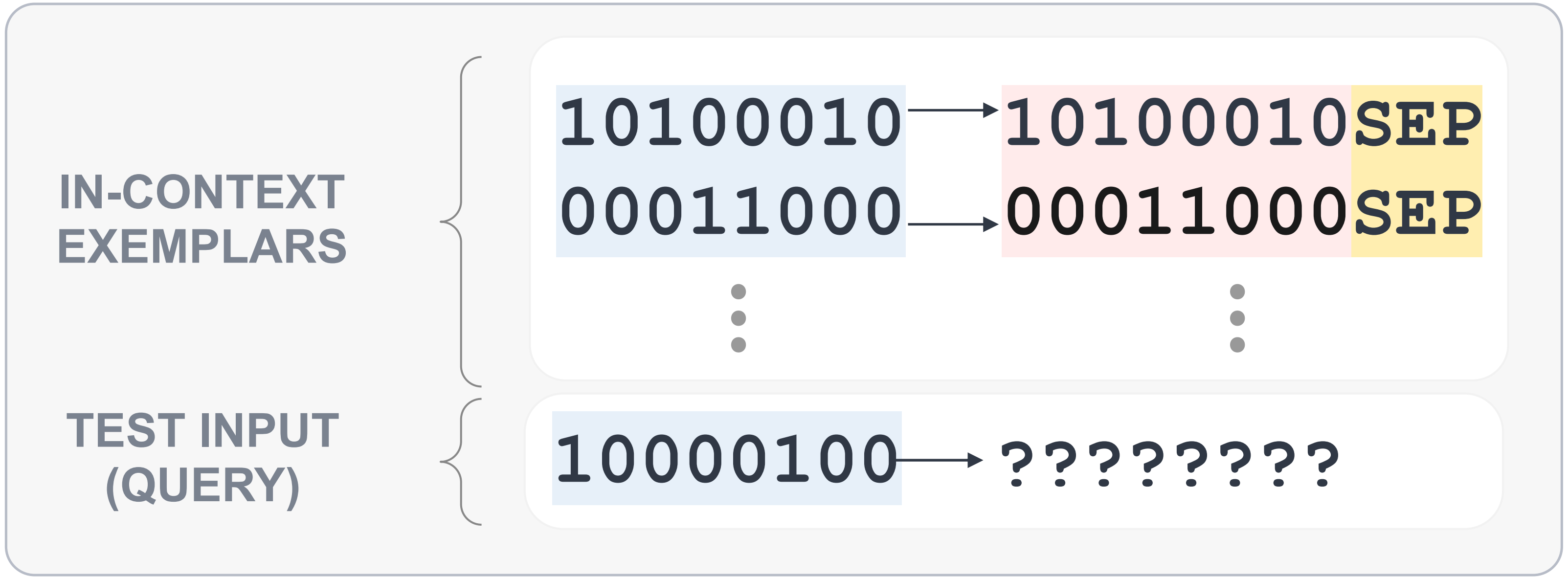
- Next-nucleotide predictors are in-context learners.
- Emergent ICL is clearly not tied to human language.
- ICL is a consequence of large autoregressive training on rich distribution.

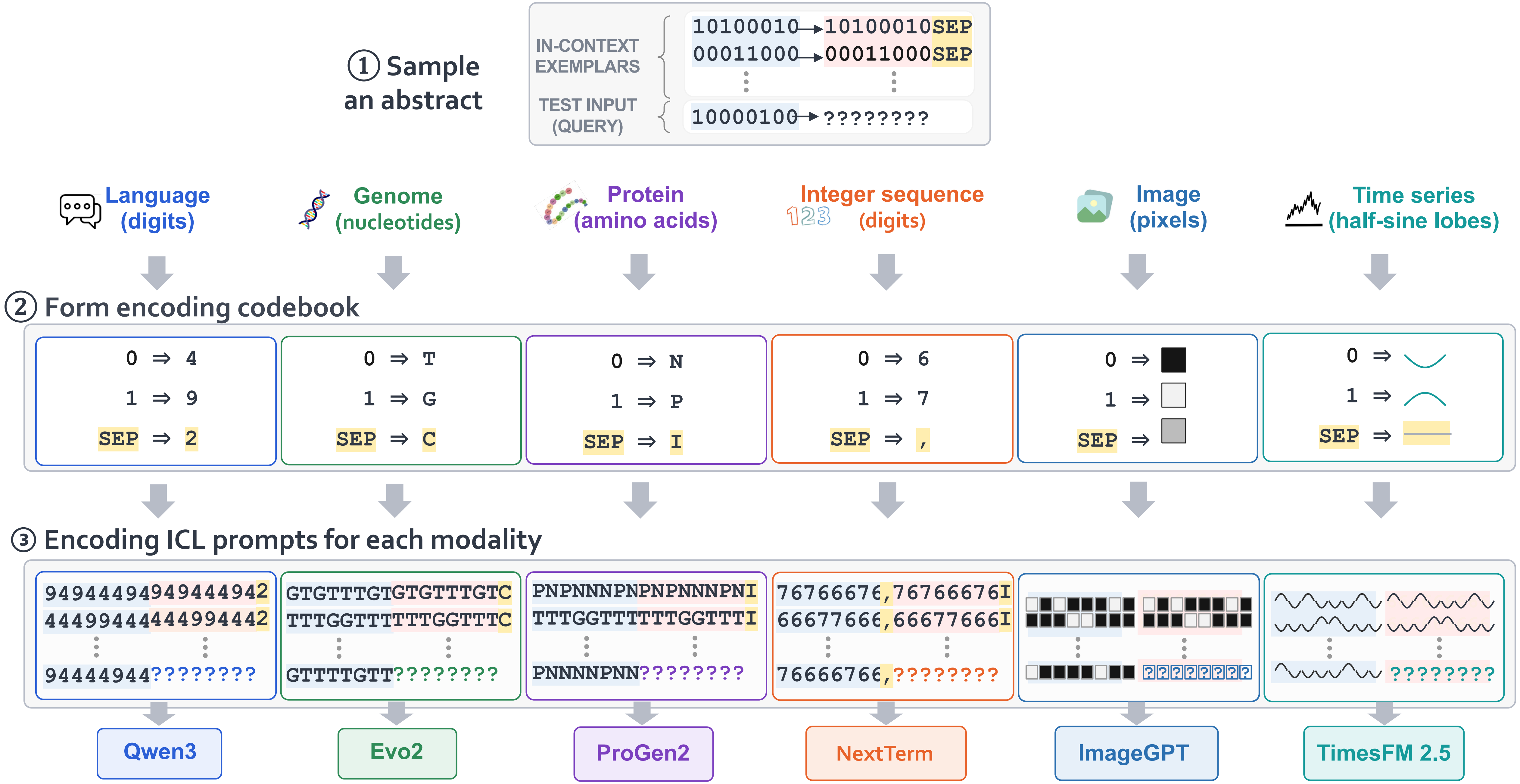
How about other modalities?

- What other data can yield ICL? And what's common in them?

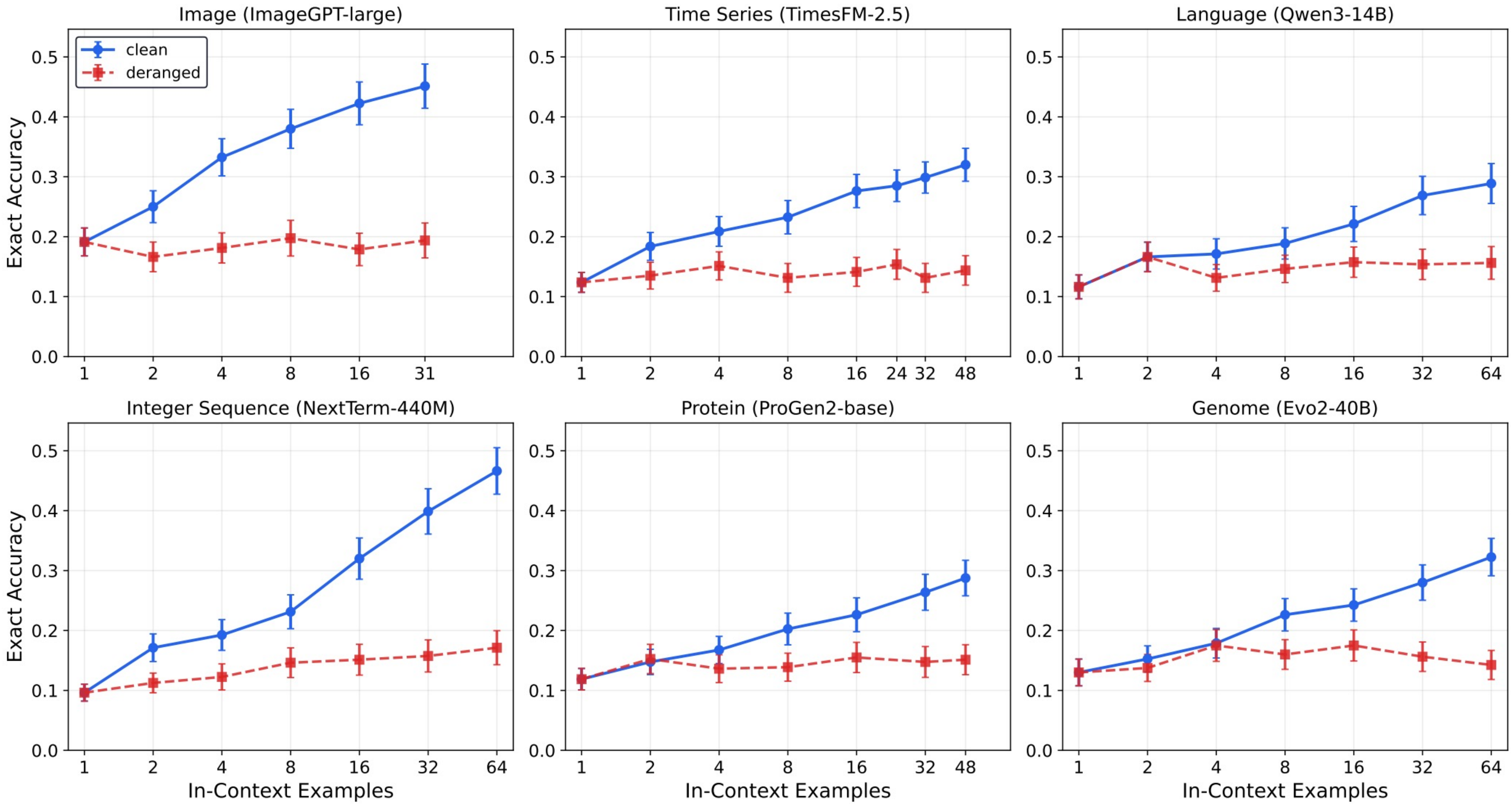
✓ Human language	Today I'm presenting …
✓ DNA sequences	T A T A A A A A A A T A T G T T T …
? Brain signals	12, 15, 18, 14, 10, 8, 11, 17, 22,…
? Time series	18934, 18951, 18920, 18885, …
? Protein sequences	M A S M T G G Q Q M G S H …

① Sample an abstract





① Sample an abstract
IN-CONTEXT EXEMPLARS
10100010 → 10100010SEP
00011000 → 00011000SEP
TEST INPUT (QUERY)
10000100 → ????????
Language (digits)
Genome (nucleotides)
Protein (amino acids)
Integer sequence (digits)
123
Image (pixels)
Time series (half-sine lobes)
② Form encoding codebook
0 ⇒ 4
1 ⇒ 9
SEP ⇒ 2
0 ⇒ T
1 ⇒ G
SEP ⇒ C
0 ⇒ N
1 ⇒ P
SEP ⇒ I
0 ⇒ 6
1 ⇒ 7
SEP ⇒ ,
0 ⇒
1 ⇒
SEP ⇒
③ Encoding ICL prompts for each modality
94944494949444942
44499444444994442
94444944????????
GTGTTTGTGTGTTTGTC
TTTGGTTTTTTGGTTTC
GTTTTGTT????????
PNPNNNPNPNPNNNPNI
TTTGGTTTTTTGGTTTI
PNNNNPNN????????
76766676,76766676I
66677666,66677666I
76666766,????????
????????
????????
Qwen3
Evo2
ProGen2
NextTerm
ImageGPT
TimesFM 2.5

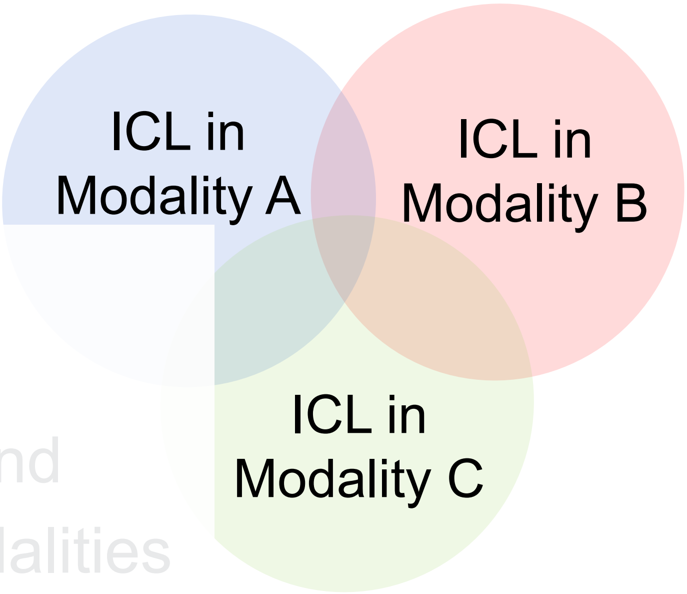


Convergent Emergence of In-Context Learning Across Modalities. Beslow et al. arXiv 2026.

Summary

- Most next-token predictors trained on natural data are in-context learners..

- ***Convergent In-Context Induction*** Hypothesis:
  1. ICL is a consequence of next-token prediction: it organically emerges in any modality with rich data, and
  2. its realizations share a **common** substrate across modalities



Open Questions

- If reasoning arises without human-interpretable tokens, what does that say about the nature of "reasoning"?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Other emergent phenomena we didn't cover

- **Generalization to unseen task instructions**

  [Mishra et al. 2021 among others]

- **Inference-time reasoning**: models improving their responses with more inference-time thinking

  [DeepSeek-R1, among others]

- **Cross-lingual transfer:** capabilities learned in English appear in low-resource languages

Today: Emergent phenomena

- What is happening inside language models?
- What do we mean by “emergent phenomena”?
- What do we understand about it?
- What does the future hold?

Today: Emergent phenomena

- What is happening inside language models?
- What do we mean by “emergent phenomena”?
- What do we understand about it?
- What does the future hold?

How far will we be able to scale?

- **Is “scaling” here to stay?**

Someone: “we won’t have enough compute”

- **For:**
  - There is simply not enough compute available.
  - Even if we have the compute (in theory), scaling the compute will be quite costly.

- **Against:**
  - Computing infrastructure continue to progress at a rapid pace.
    - E.g. renewables, space data centers, inference hardware, etc.
  - The economy of scale will just work out.
    - With more usage, cost per unit will be cheaper.

Someone: “we won’t have enough data”

- **For:**
  - Existing LLMs are severely data hungry. There is simply not much enough data in the world.
  - And you don’t need just any data; you want data on challenging and rare problems (e.g., hard math and engineering problems).
- **Against:**
  - There is massive amount of untapped data in the world.
  - As we deploy AI in more applications, allowing continuous stream of fresh data.

How far will we be able to scale?

- **Is “scaling” here to stay?**

Putting things together

- **Emergence is real:** Large-scale pre-training on *simple* objectives lead to capabilities that appear at scale.

- **In-context learning** is a prominent manifestation of emergence.
  - Not specific to human language.

- **It is important to understand emergence:** It is not just an academic endeavor. It is how a foundation of how we form trustworthy AI.