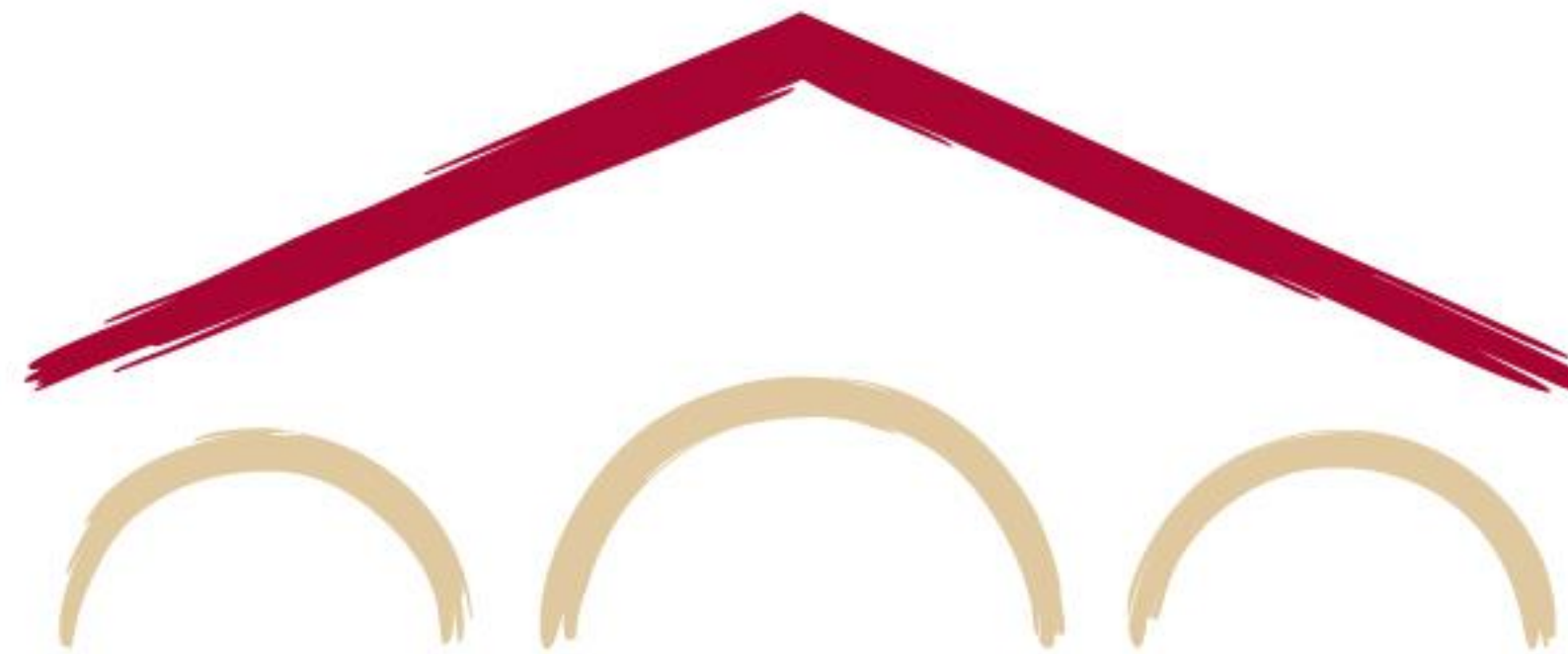


The Surprising Victory of NLP: From Philosophy to Agentic Language Models

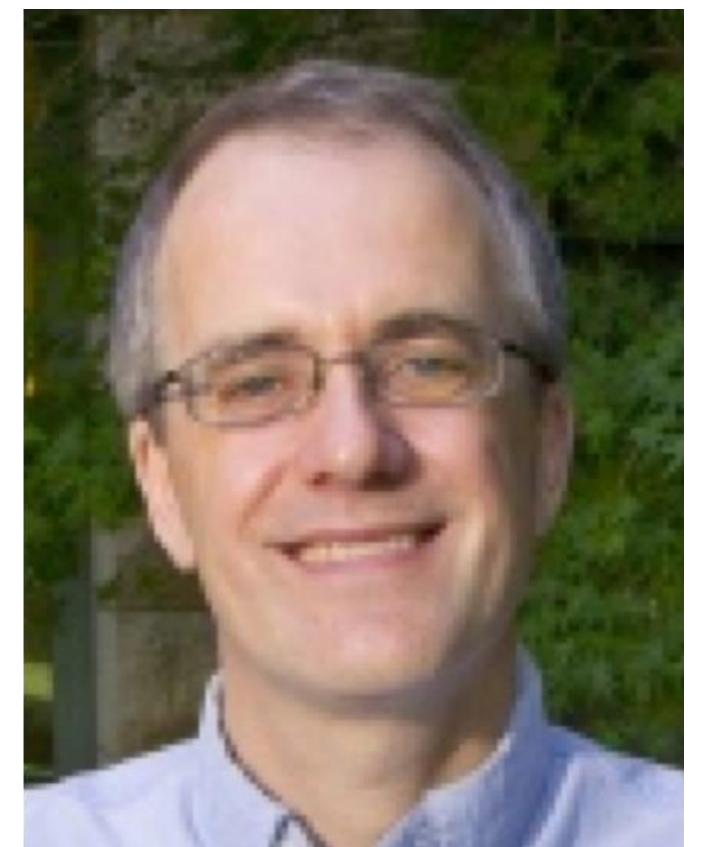


Christopher Manning

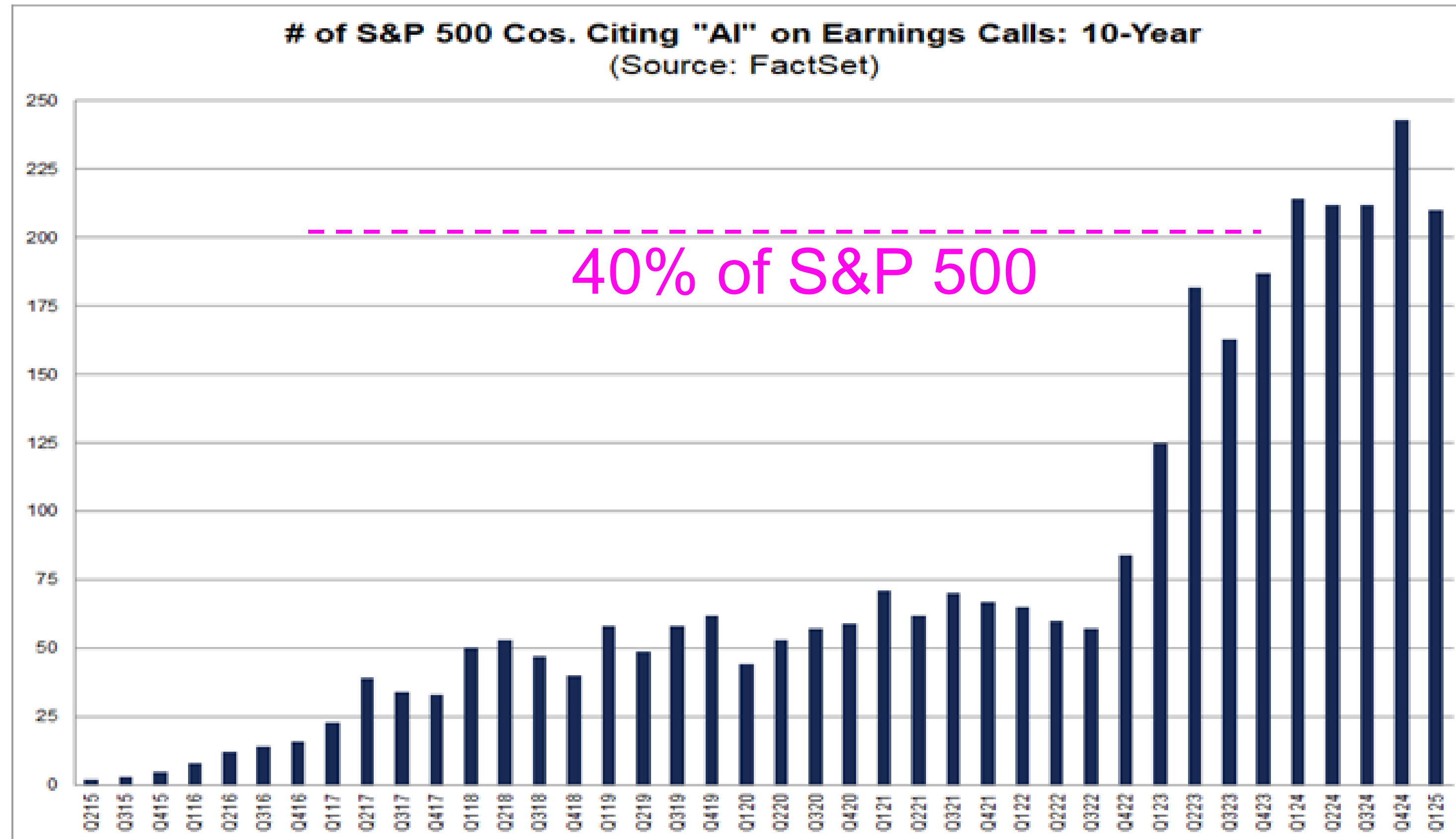
@chrmanning ✿ manning@stanford.edu

Founder, Stanford NLP Group ✿ Senior Fellow, HAI

KDD 2025 ✿ Toronto

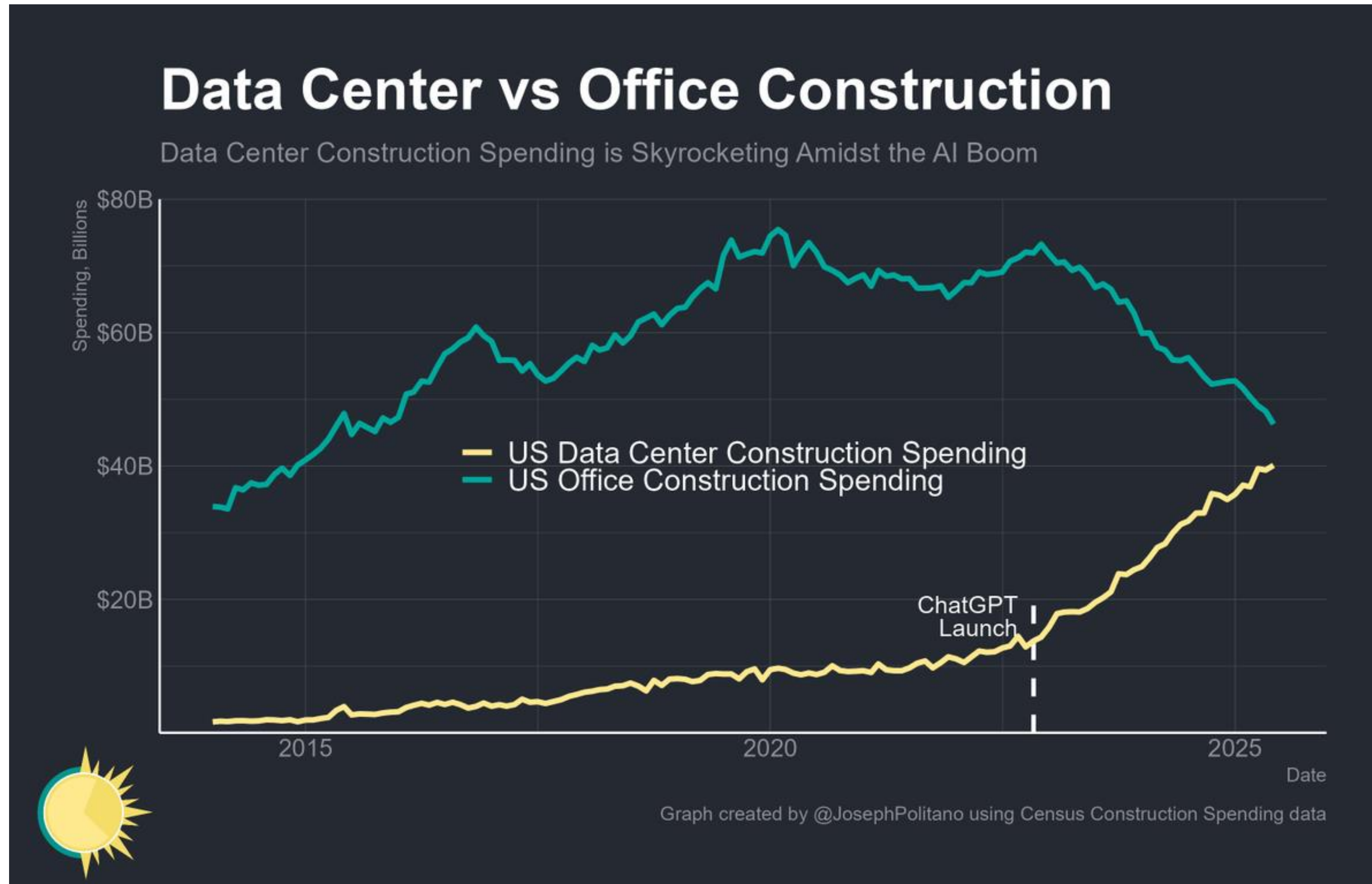


The Rise of Generative AI: S&P 500 Companies Discussing “AI”



ChatGPT launch

The Rise of Generative AI: Data Center Construction for AI



Who should you take your LLM wisdom from in 2025?

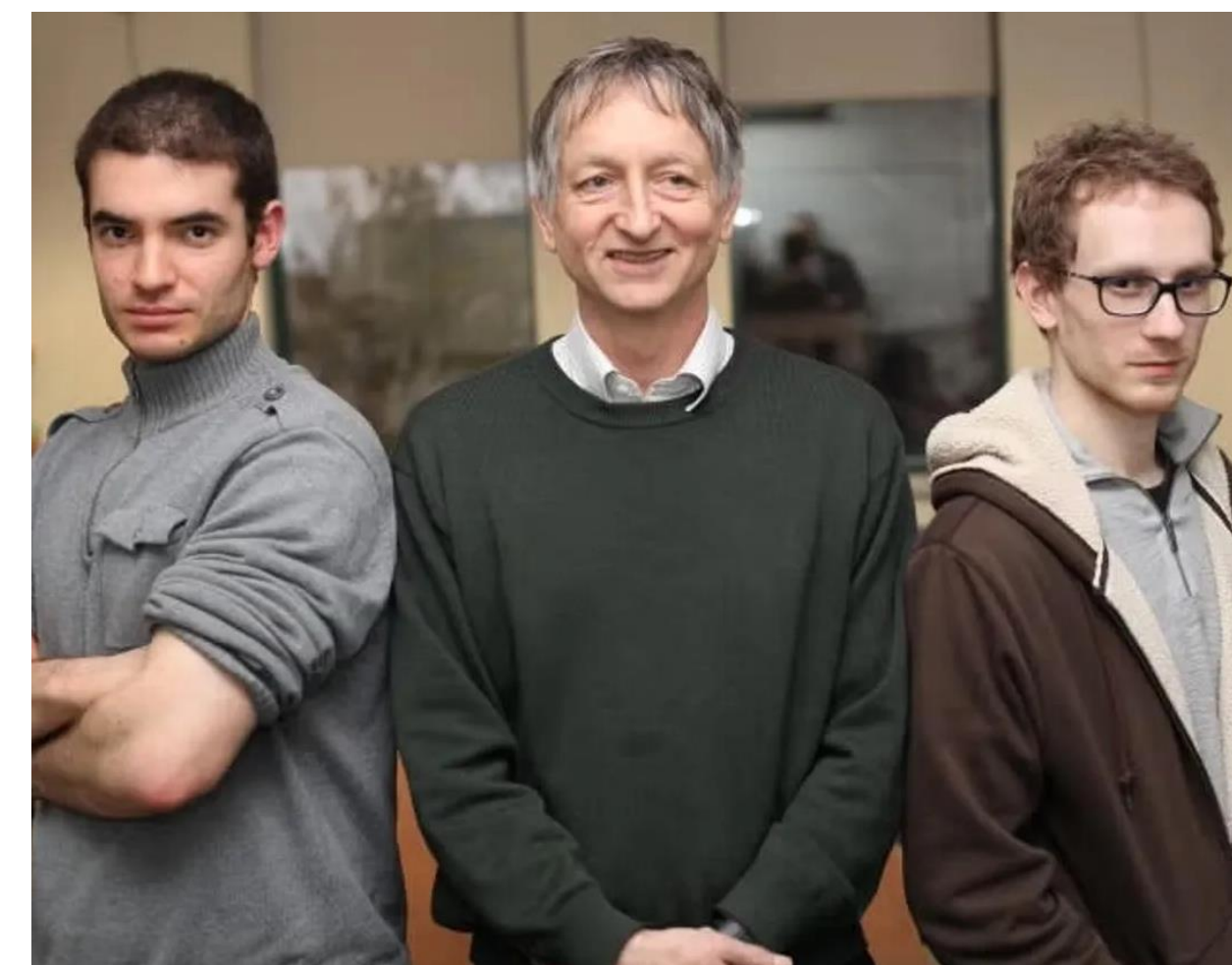
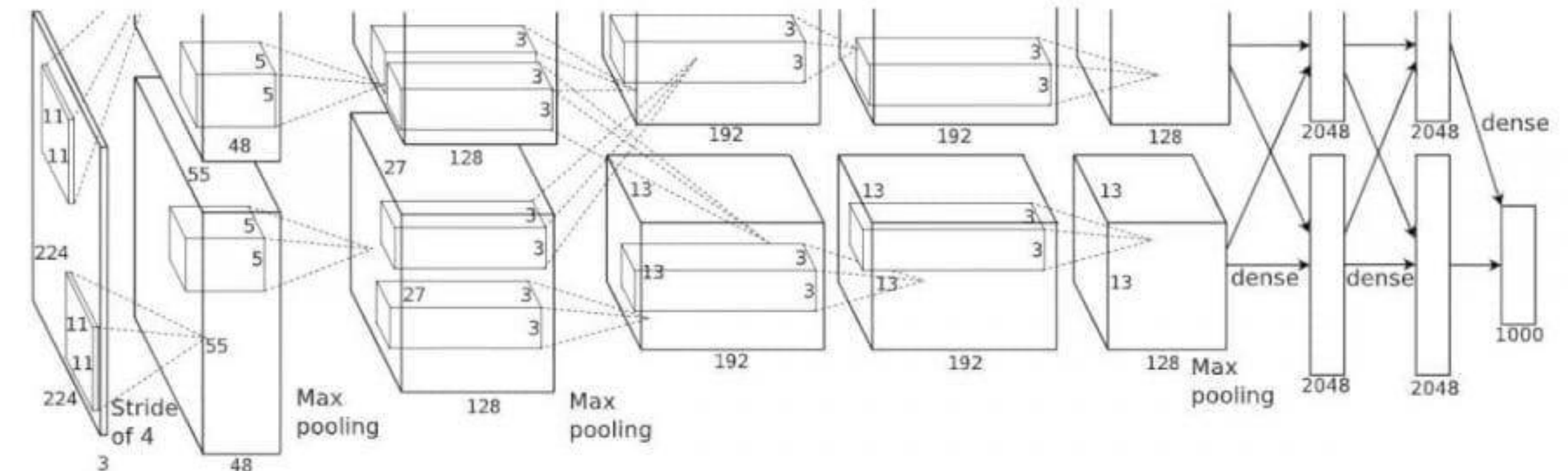
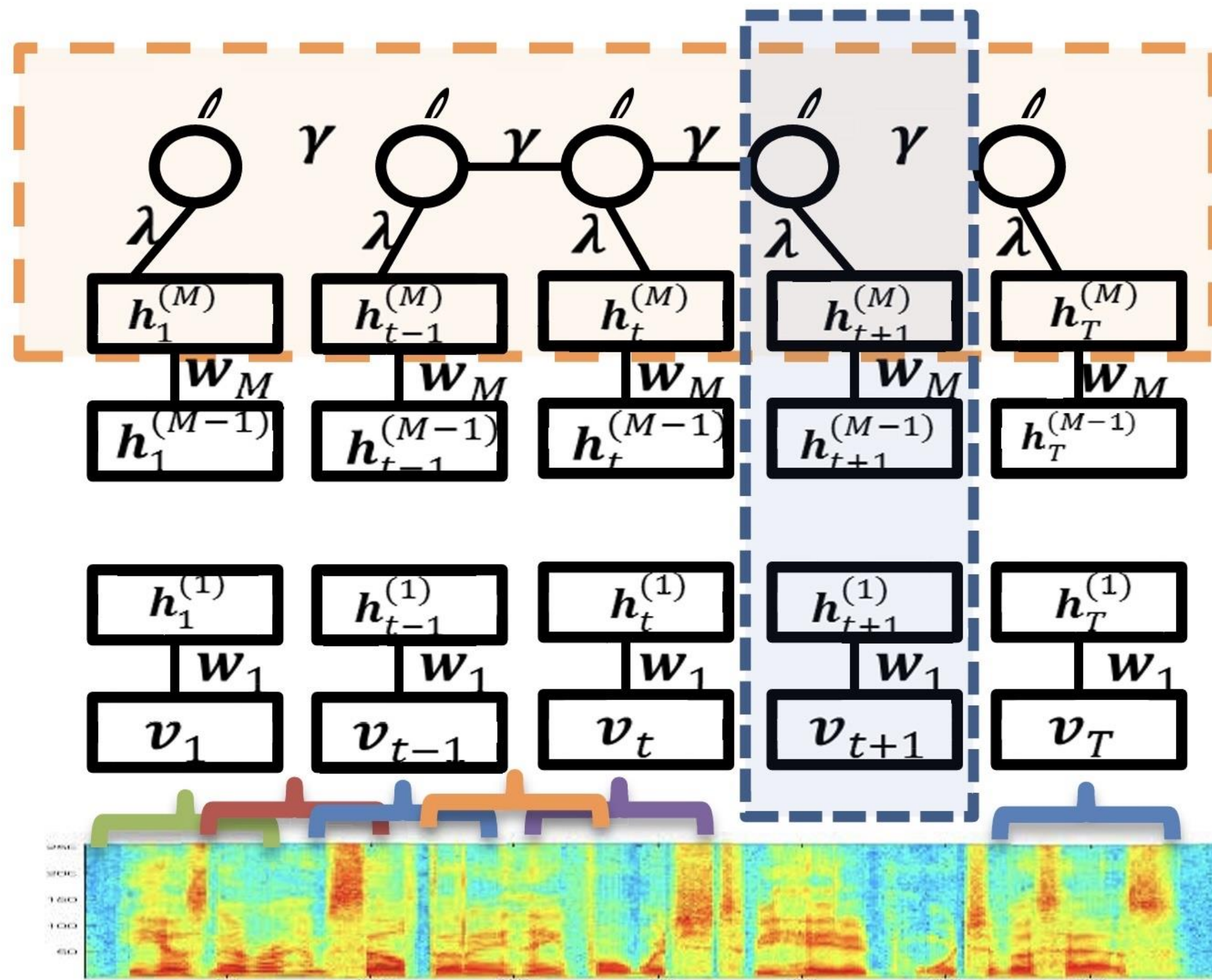
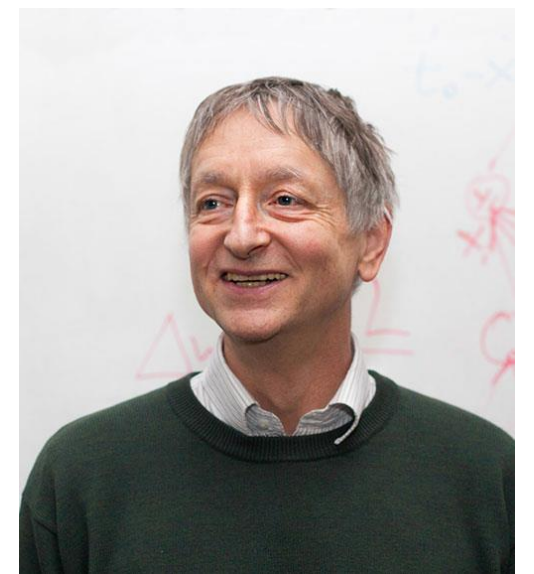


or



Deep Learning Breakthroughs by Hinton

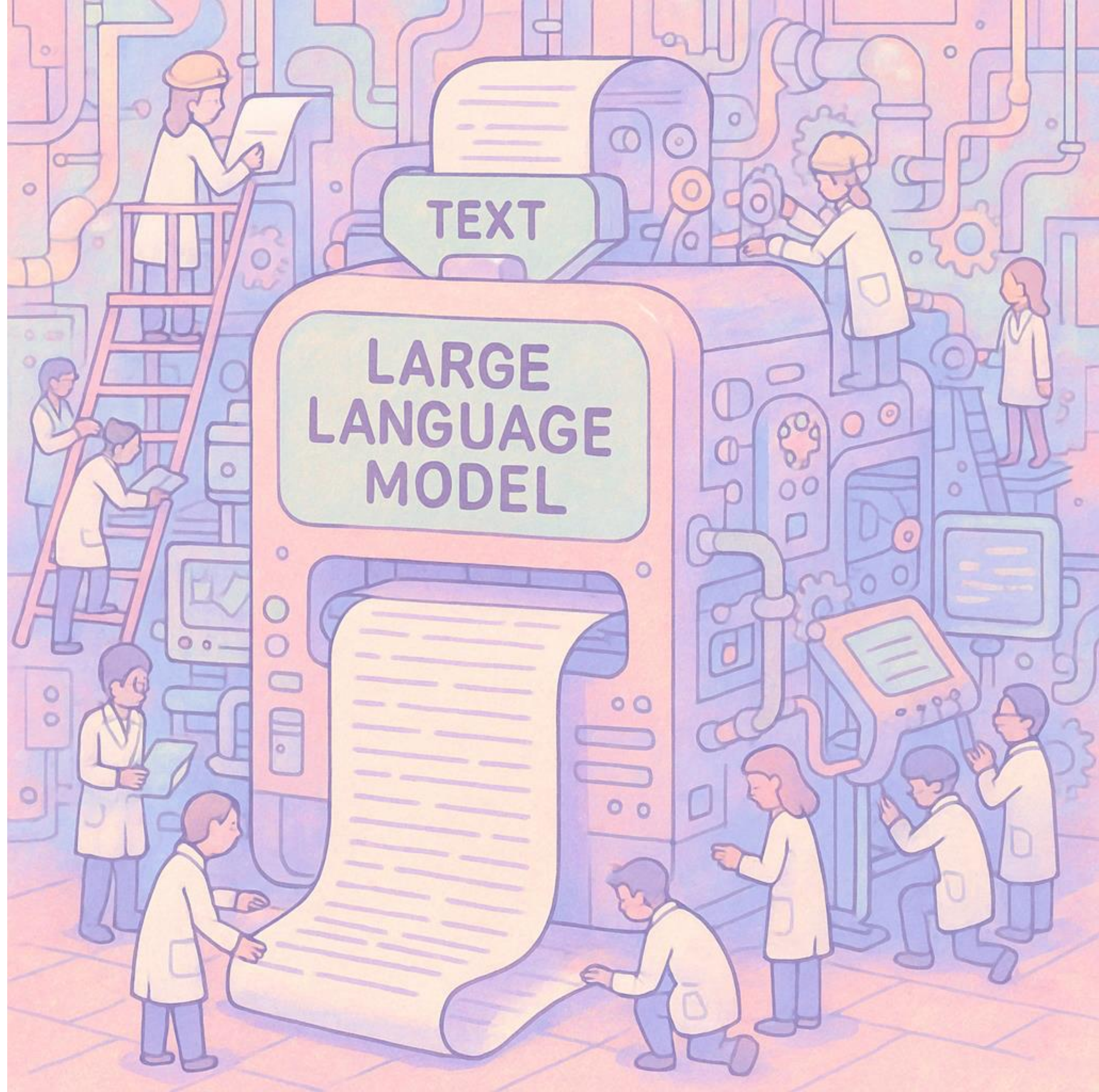
... at the University of Toronto, 2009–2012



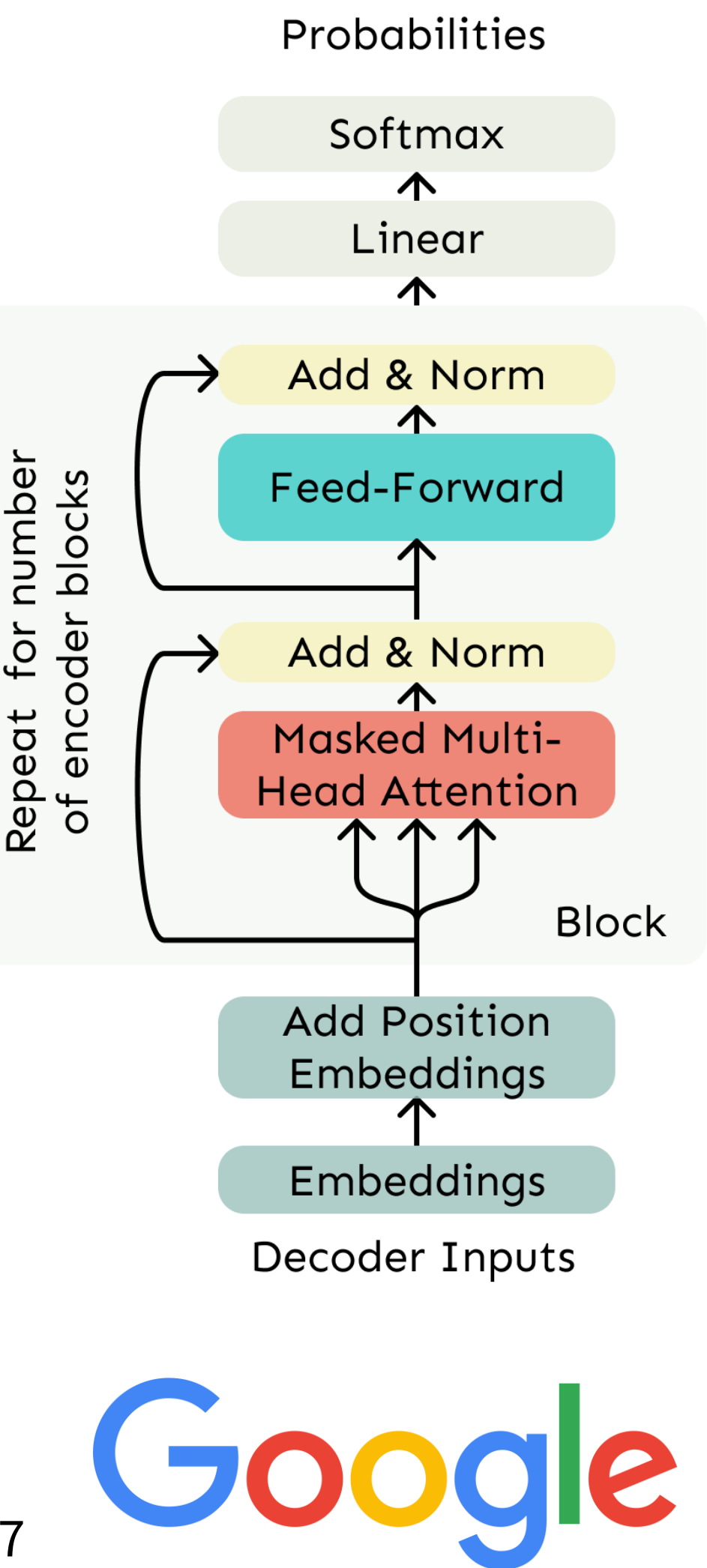
Large Language Models

Source: GPT 4o

An abstract image in pastel manga style of a large language model as a big machine that takes in text and produces fresh writing on an endless scroll and which is maintained by a large number of scientists and engineers.



Modern LLMs Use the Transformer Neural Network Architecture



Attention Is All You Need

Vaswani, Shazeer, Parmar, Uzoreit, Jones, Gomez, Kaiser, and Polosukhin
Google Research

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

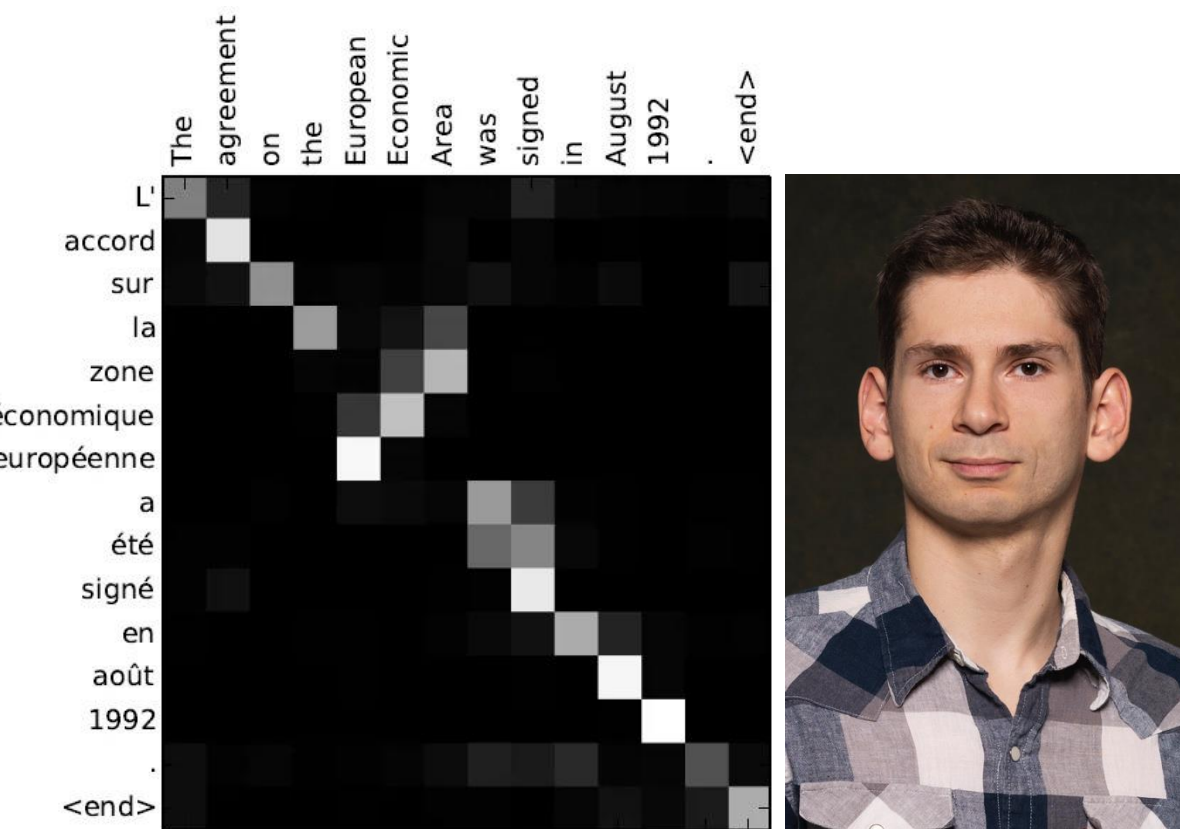
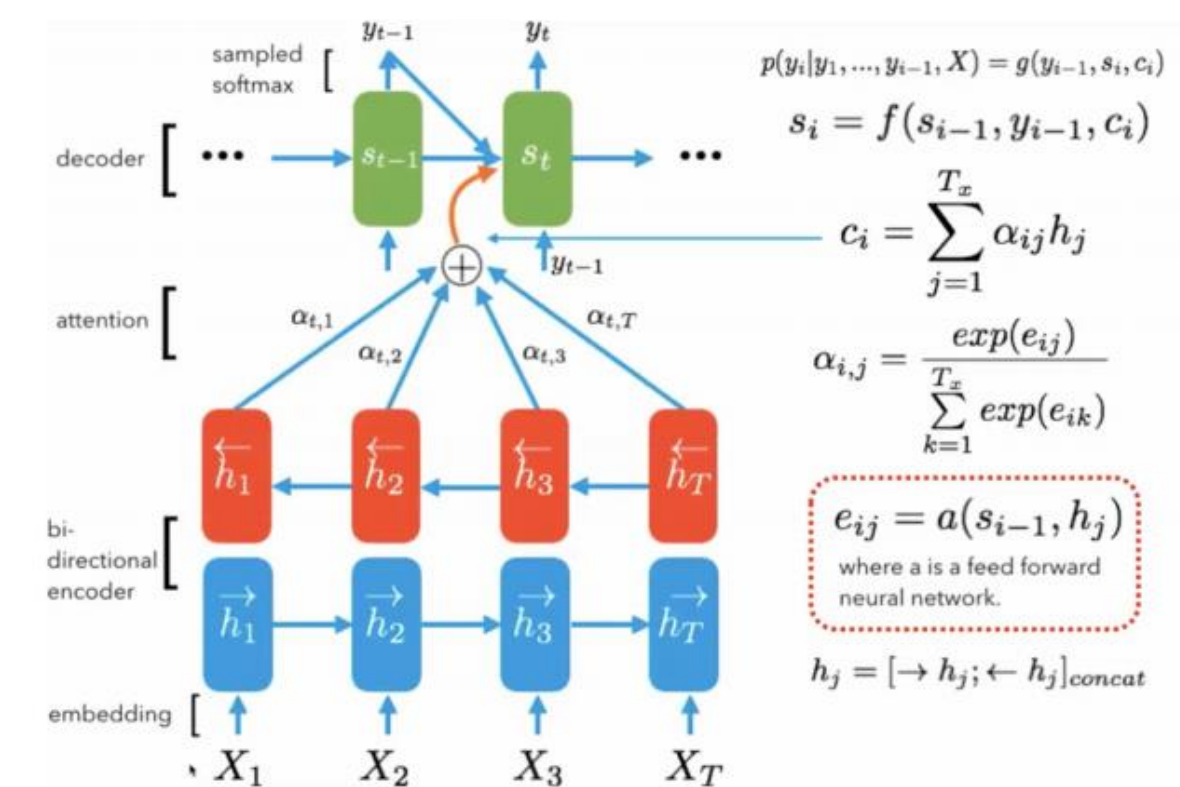
1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

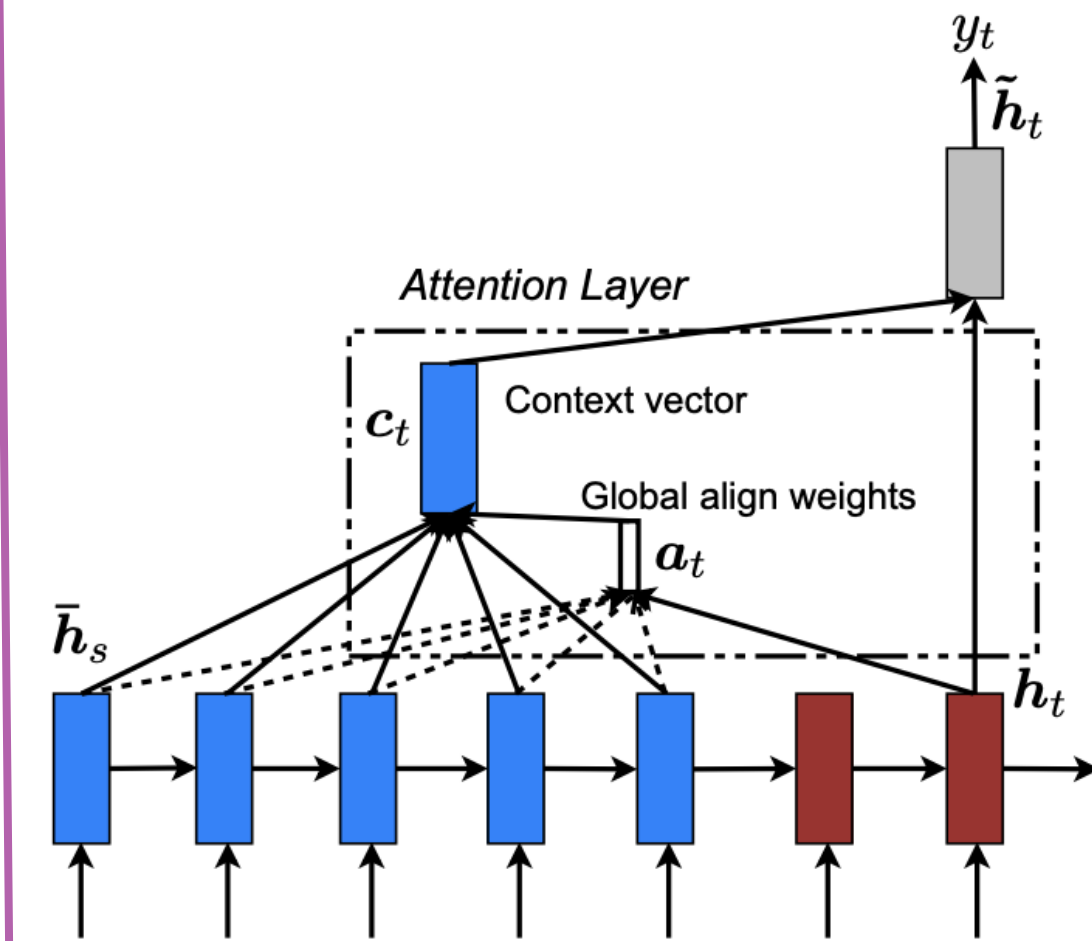
^{*}Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

[†]Work performed while at Google Brain.

[‡]Work performed while at Google Research.

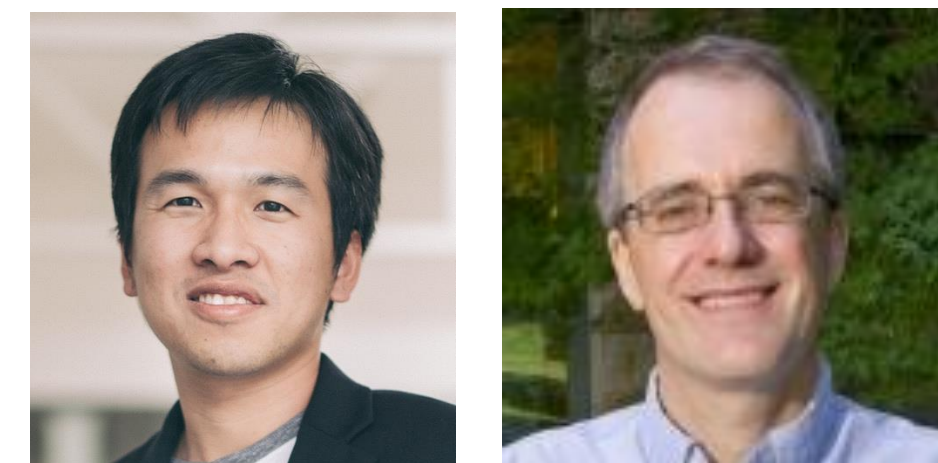


Université de Montréal



Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong Hieu Pham Christopher D. Manning
Computer Science Department, Stanford University, Stanford, CA 94305
{lmthang, hyhieu, manning}@stanford.edu



Stanford

ChatGPT 4o

Write a sonnet praising Wikipedia in which each line begins with the letter W

Wikipedia: A Sonnet of Wonder

With wisdom vast, it spans the world so wide,
Wielding the wealth of knowledge, free to share.
Words weave through time where curious minds confide,
Whispering truths with patience, just and fair.
Wondrous it stands, a beacon shining bright,
Warding off darkness, doubt, and false pretense.
Wrought by the hands of many, day and night,
Witness to learning's boundless recompense.
Who dares to claim the heights of thought alone?
Whence comes this trove but hearts in concert joined?
While scholars build on stones that others've known,
Waking the past, their voices intertwine.
Wisdom and will sustain this noble dream,
Wikipedia—our light, our endless stream.

ACL 1993

* Calculated median survival from Gompertz equation, with the assumption that a is 0.05, and R_0 extrapolated to birth 2.45×10^{-6} .

$$S = e^{(R_0/a)(1-e^{a^*})}$$

References

1. Gavrilov and Gavrilova NS. *The Biology of Life Span, A Quantitative Approach*. Harwood Academic (1991).
2. Brody S. *J. Gen. Physiol.* 6 (1923) 245-257.
3. Jones HR. *Adv. Biol. Med. Physiol.* 4 (1956) 281-337.
4. Hallén A. *Acta Chem. Scand.* 62 (1962) 705-716.
5. " *Biogeront.* 3 (2002) 307-315.
6. " *Biogeront.* 8 (2007) in press.
7. Laurent TC, Kallander J. *J. Chromatogr.* 34 (1964) 317-330.
8. Mirtzai H, van Oudenarden A. *Science* 317 (2007) 483-484.
9. Finkel T, Serrano M, Blasco MA. *Nature* 2007;448:767-774.
10. McFarland G.V., Holliday R. *Expil. Cell Res.* 1994; 212:167-175.
11. Hopkins AH et al. *Mech. Ageing Dev.* 122 (2001) 1431-1445.
12. Verheke et al. *Biogeront.* 3 (2002) 115-126.
13. Olson et al. *Biogeront.* 7 (2006) 221-230.
14. Chondrogianni N et al. *Exp. Gerontol.* 35 (2000) 721-728.
15. Chondrogianni N, Ginos ES. *Biogeront.* 9 (2004) 55-61.

DISTRIBUTIONAL CLUSTERING OF ENGLISH WORDS

Fernando Pereira Naftali Tishby Lillian Lee
AT&T Bell Laboratories Dept. of Computer Science Dept. of Computer Science
600 Mountain Ave. Hebrew University Cornell University
Murray Hill, NJ 07974, USA Jerusalem 91904, Israel Ithaca, NY 14850, USA
pereira@research.att.com tishby@cs.huji.ac.il lllee@cs.cornell.edu

Abstract
We describe and evaluate experimentally a method for clustering words according to their distribution in particular syntactic contexts. Words are represented by the relative frequency distributions of contexts in which they appear, and relative entropy between those distributions is used as the similarity measure for clustering. Clusters are represented by average context distributions derived from the given words according to their probability of cluster membership. In many cases, the clusters can be thought of as encoding coarse sense distinctions. Deterministic rewording is used to find lowest distortion sets of clusters, as the smoothing parameter increases, existing clusters become unstable and subdivide, yielding a hierarchical "soft" clustering of the data. Clusters are used as the basis for class models of word occurrence, and the models evaluated with respect to hold-out test data.

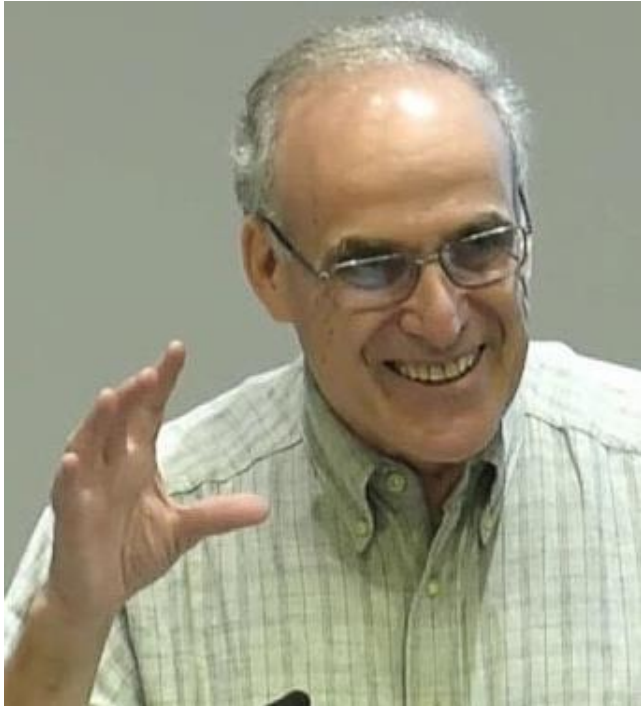
INTRODUCTION
Methods for automatically classifying words according to their contexts of use have both scientific and practical interest. The scientific question arises in connection with theories of natural language (particularly lexical) structure and also in relation to the question of how humans learn both from psychological and computational learning perspectives. From the practical point of view, word classification addresses questions of data sparseness and generalization in statistical language models, particularly models for decoding among alternative analyses proposed by a grammar.

It is well known that a simple tabulation of frequencies of certain words participating in certain configurations, for example of frequencies of pairs of a transitive main verb and its direct object, or of a transitive main verb and the head noun of its direct object, cannot be reliably used for comparing the likelihood of different hypotheses for a given reason. The problem is that for large enough corpora the number of possible joint events is much larger than the number of event occurrences in the corpus, so many events are seen rarely or never, making their frequency counts unreliable estimates of their probabilities.

Hindle (1990) proposed dealing with the



Pereira



Tishby



Lee

Distributional Clustering of English Words

Towards History-based Grammars: Using Richer Models for Probabilistic Parsing*

Eraa Black Fred Jelinek John Lafferty David M. Magerman
Robert Mercer Salim Roukos
IBM T. J. Watson Research Center

Abstract
We describe a generative probabilistic model of natural language, which we call RBG, that takes advantage of detailed linguistic information to resolve ambiguity. RBG incorporates lexical, syntactic, semantic, and structural information from the parse tree into the disambiguation process in a novel way. We use a corpus of bracketed sentences, called a Databank, in combination with decision tree building to tune on the relevant aspects of a parse tree that will determine the correct parse of a sentence. This stands in contrast to the usual approach of further grammar tailoring via the usual linguistic intuition in the hope of generating the correct parse. In head-to-head tests against one of the best existing robust probabilistic parsing models, which we call P-CFG, the RBG model significantly outperforms P-CFG, increasing the parsing accuracy rate from 60% to 75%, a 25% reduction in error.

Introduction
Almost any natural language sentence is ambiguous in structure, reference, or source of meaning. Humans overcome these apparent ambiguities by examining the context of the sentence. But what exactly is context? Frequently, the correct interpretation is apparent from the words or constituents immediately surrounding the phrase in question. This observation begs the following question: How much information about the context of a sentence or phrase is necessary and sufficient to determine its meaning? This question is at the crux of the debate among computational linguists about the application and implementation of statistical methods in natural language understanding.

Previous work on disambiguation and probabilistic parsing has offered partial answers to this question. Hidden Markov models of words and

*Thanks to Philip Resnik and Stanley Chen for their valued input.



Black
Lafferty
Mercer

Jelinek
Magerman
Roukos

Towards History-based Grammars: Using Richer Models for Probabilistic Parsing

Why was no one working on language models?

How were they hoping to get to artificial intelligence?

The History of Language Models

- 1913 Andrey A. Markov explores consonant-vowel transition probabilities in Alexander Pushkin novel, *Eugene Onegin*, developing Markov Models 
- 1948 Claude E. Shannon develops *A Mathematical Theory of Information*; explores character/word-level n -gram models, entropy, text generation 
- 1975 Frederick Jelinek's group at IBM defines and names the modern concept of a (probabilistic) "language model" for next token prediction. Used for spelling correction, speech recognition, machine translation.... 

But this technology wasn't considered a pathway to machines with artificial intelligence: For that we needed models of memory, knowledge representations, planning systems, and reasoning over abstract concepts

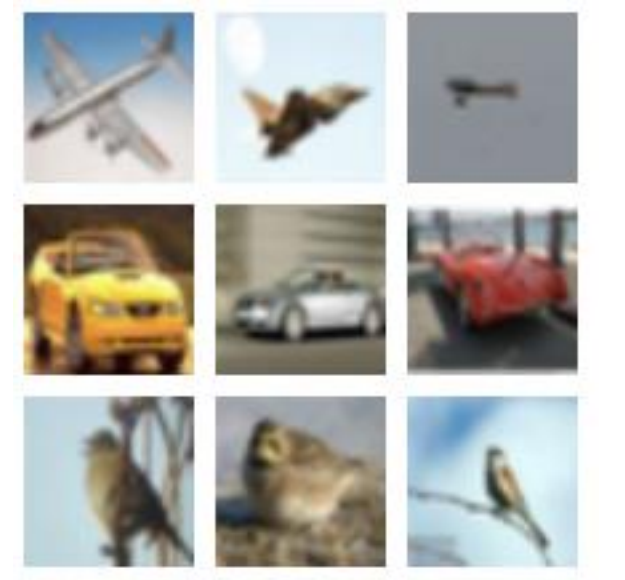
The history of modern (neural) AI/ML that many think of

1950–2004 Experiments on perception and vision,
with minute amounts of data



2005–2009 Small modern-style labeled vision datasets

- 2005–2012 PASCAL VOC 4 – 15,000 images, 20 classes
- 2006 Caltech 101 – ~5000 images, 101 classes, ~300x200
- 2007 Caltech 256 – 30,607 images, 256 classes
- 2009 CIFAR 10/100 – 600 images of 10/100 classes, 32x32



2010–2017 **ImageNet!** 1–14 million images, 1000 classes
Big data! Vision object classification is “solved”



The History of Large Language Models

- 1998 CPAT-Tree-Based Language Models with an Application for Text Verification in Chinese. ROCLing 1998. First use of “LLM” trigram I know of; 200M word corpus
- 2000 A Neural Probabilistic Language Model. Bengio, Ducharme & Vincent NIPS 2000
First neural language model built on 32 million token corpus, 31K vocab
- 2007 Large Language Models in Machine Translation. Brants, Popat, Xu, Och and **Dean**. EMNLP 2007. **2 trillion token corpus** n-gram model of up to 5-grams
- 2018 **GPT** (Radford, Narasimhan, Salimans & Sutskever) and **BERT** (Devlin, Chang, Lee & Toutanova). 3.3 billion token corpus
- 2020– 100+ billion parameter neural language models trained on > 1 trillion tokens: GPT-3, GPT-4, PaLM 2, Llama 3, Nemotron-4, ...

The History of Large Language Models

- 1998 CPAT-Tree-Based Language Models with an Application for Text Verification in Chinese. ROCLing 1998. First use of “LLM” trigram I know of; 200M word corpus
- 2000 A Neural Probabilistic Language Model. Vincent NIPS 2000
First neural language model. 31K vocab
Not enough **compute!**
- 2007 Large Language Models. Dean. EMNLP 2007
Not enough **model flexibility!** Och and 5-grams
- 2018 **GPT** (Radford, Narasimhan, et al.). 3.3 billion parameters. OpenAI, Lee & Toutanova). 3.3 billion parameters
Not enough **data!**
- 2020– 100+ billion parameter neural language models
GPT-3, GPT-4, PaLM 2, Llama 3, etc.
LLMs go brrr!

The Philosophical Turn Leading to Large Language Models



Texts are Knowledge

Christopher Manning
AKBC 2013 Workshop

Gabor Angeli, Jonathan Berant, Bill MacCartney Mihai Surdeanu, Arun Chaganty,
Angel Chang, Kevin Aju Scaria, Mengqiu Wang, Peter Clark, Justin Lewis, Brittany
Harding, Reschke, Julie Tibshirani, Jean Wu, Osbert Bastani, Keith Sillats

Jeremy Zawodny said....

[on a blog sometime in the 2000s]

Back in the late 90s when I was building things that passed for *knowledge management tools* at Marathon Oil, there was all this talk about *knowledge workers*. These were people who'd have vast quantities of knowledge at their fingertips. All they needed was ways to organize, classify, search, and collaborate.

I think we've made it. But the information isn't organized like I had envisioned a few years ago. It's just this big ugly mess known as **The Web**. Lots of pockets of information from mailing lists, weblogs, communities, and company web sites are loosely tied together by hyperlinks. **There's no grand schema or centralized database.**
There's little structure or quality control. No global vocabulary.

But even with all that going against it, it's all indexed and easily searchable thanks largely to Google and the companies that preceded it (Altavista, Yahoo, etc.). Most of the time it actually works. Amazing!



The Surprising Victory of NLP!

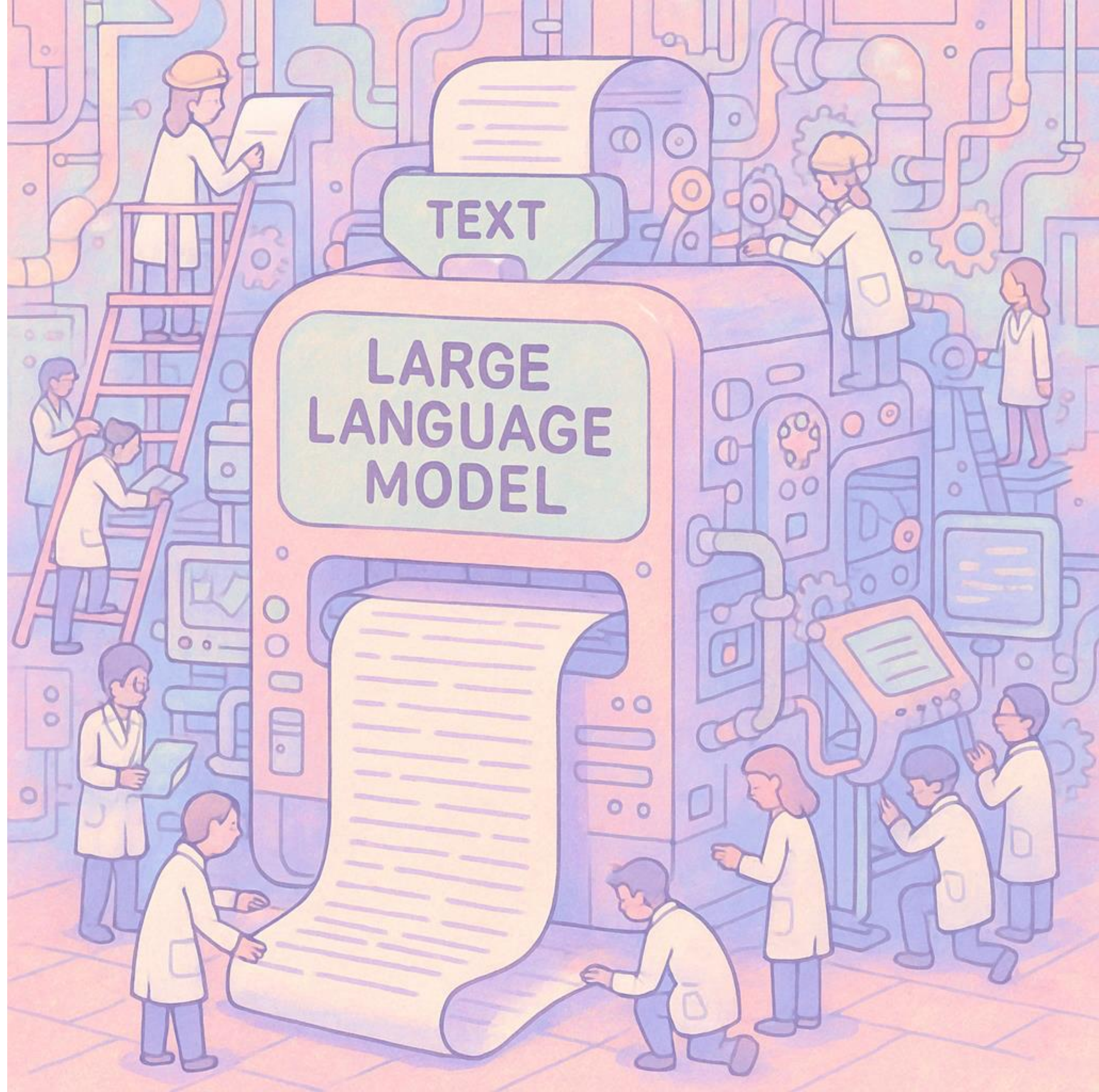
NLP has moved from the margins to the center of ML & AI

For most people now, LLMs **are** AI

A flexible, informal, ubiquitous old knowledge representation has won over the formal languages of recent centuries

Source: GPT 4o

An abstract image in pastel manga style of a large language model as a big machine that takes in text and produces fresh writing on an endless scroll and which is maintained by a large number of scientists and engineers.

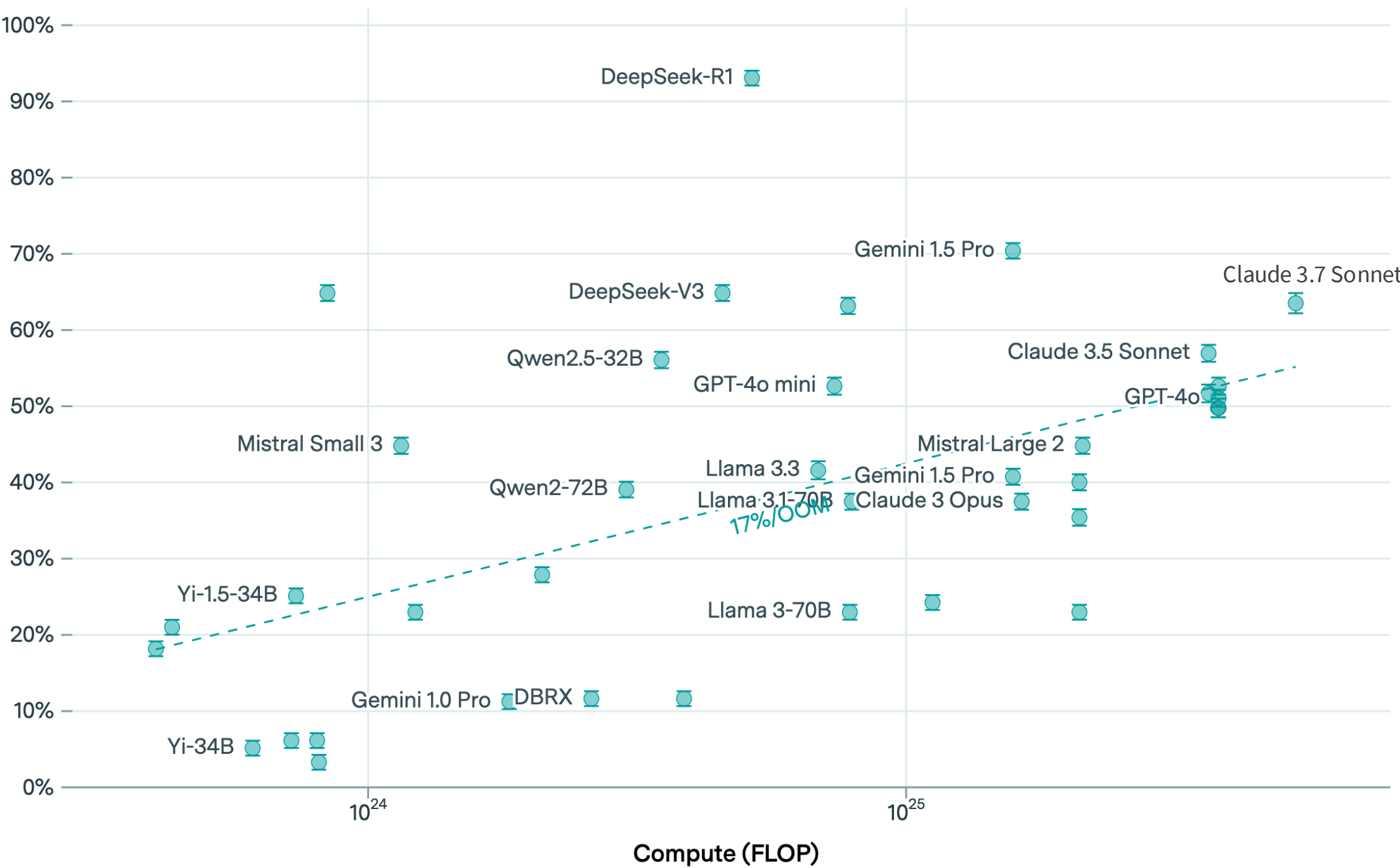


Amazing progress on math, science, and coding

Benchmark accuracy increases with training compute

EPOCH AI

MATH Level 5 accuracy



CC-BY

epoch.ai

MATH Level 5 = Hardest quarter of MATH problems from Art of Problem Solving & similar USAMO-level questions

Example MATH Level 5 question:

You have two circles, one with radius r and the other with radius R . You wish for the difference in the areas of these two circles to be less than or equal to 5π . If $r + R = 10$, what is the maximum difference in the lengths of the radii?

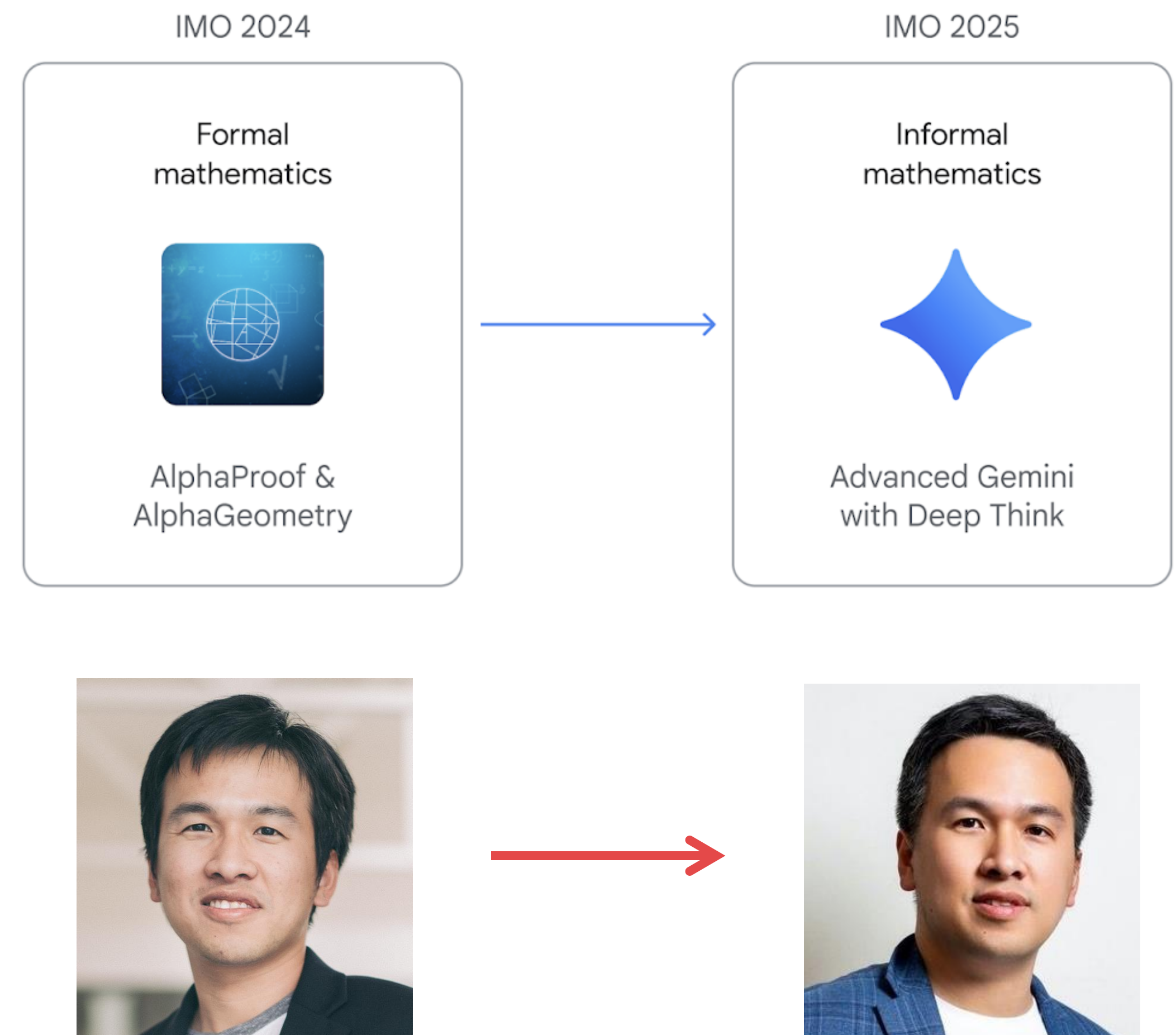
International Math Olympiad 2025

“We can confirm that Google DeepMind has reached the much-desired milestone, earning 35 out of a possible 42 points — a gold medal score.



“Their solutions were astonishing in many respects. IMO graders found them to be clear, precise and most of them easy to follow.”

– IMO President Prof. Dr. Gregor Dolinar

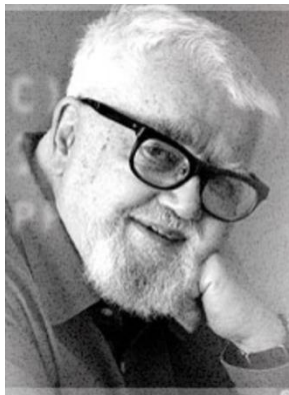


Large Language Models and “Intelligence”

Since ChatGPT, in the popular imagination, large language models have essentially become synonymous with **artificial intelligence**. What is AI?

The **artificial** part is easy:

“The science and engineering of making intelligent machines”
— John McCarthy (1955)



What is **intelligence**?

Is language understanding and generation – or even doing mathematics – distinct from general intelligence?

TICS 2534 No. of Pages 24

ARTICLE IN PRESS

Trends in Cognitive Sciences

CellPress

Feature Review

Dissociating language and thought in large language models

Kyle Mahowald,^{1,5,*} Anna A. Ivanova,^{2,5,*} Idan A. Blank,^{3,*} Nancy Kanwisher,^{4,*} Joshua B. Tenenbaum,^{4,*} and Evelina Fedorenko^{4,*}

Large language models (LLMs) have come closest among all models to date to mastering human language, yet opinions about their linguistic and cognitive capabilities remain split. Here, we evaluate LLMs using a distinction between formal linguistic competence (knowledge of linguistic rules and patterns) and functional linguistic competence (understanding and using language in the world). We ground this distinction in human neuroscience, which has shown that formal and functional competence rely on different neural mechanisms. Although LLMs are surprisingly good at formal competence, their performance on functional competence tasks remains spotty and often requires specialized fine-tuning and/or coupling with external modules. We posit that models that use language in human-like ways would need to master both of these competence types, which, in turn, could require the emergence of separate mechanisms specialized for formal versus functional linguistic competence.

Highlights

Formal linguistic competence (getting the form of language right) and functional linguistic competence (using language to accomplish goals in the world) are distinct cognitive skills.

The human brain contains a network of areas that selectively support language processing (formal linguistic competence), but not other domains like logical or social reasoning (functional linguistic competence).

In the late 2010s, large language models trained on word prediction tasks began achieving unprecedented success in formal linguistic competence, showing impressive performance on linguistic tasks that likely require hierarchy and abstraction.

Consistent performance on tasks requiring functional linguistic competence is harder to achieve for large language models and often involves augmentations beyond next word prediction.

Evidence from cognitive science and neuroscience can illuminate the capabilities and limitations of large language models and pave the way toward better, human-like models of both language and thought.

The language–thought conflation

When we hear a sentence, we typically assume that it was produced by a rational, thinking agent (another person). The sentences that people generate in day-to-day conversations are based on their world knowledge (‘Not all birds can fly.’), their reasoning abilities (‘You’re 15, you can’t go to a bar.’), and their goals (‘Would you give me a ride, please?’). Thus, we often use other people’s statements as a window into their minds.

In 1950, Alan Turing leveraged this tight relationship between language and thought to propose his famous test [1]. The Turing test uses language as an interface to cognition, allowing a human participant to probe the knowledge and reasoning capacities of two conversation partners to determine which of them is a human and which is a machine. Although the utility of the Turing test has since been questioned, it has undoubtedly shaped the way society today thinks of machine intelligence [2].

The popularity of the Turing test, combined with language–thought coupling in everyday life, has led to several common fallacies related to the language–thought relationship. One fallacy is that **an entity (be it a human or a machine) that is good at language must also be good at thinking**. If an entity generates coherent stretches of text, it must possess rich knowledge and reasoning capacities. Let’s call this the **‘good at language -> good at thought’ fallacy**. This fallacy has come to the forefront due to the recent rise of **large language models (LLMs)** (see **Glossary**), including OpenAI’s GPT models, Anthropic’s Claude, and more open alternatives [3] like Meta’s LLaMa models and EleutherAI’s GPT-J. LLMs today can produce text that is difficult to distinguish from human output, outperform humans at some text comprehension tasks [4,5], and show superhuman performance on next-word prediction [6]. As a result, claims have emerged, both in the popular press and in the academic literature, that LLMs are not only a major advance in language processing, but are also showing **‘sparks of artificial general**

¹The University of Texas at Austin, Austin, TX, USA

²Georgia Institute of Technology, Atlanta, GA, USA

³University of California, Los Angeles, CA, USA

⁴Massachusetts Institute of Technology, Cambridge, MA, USA

⁵These authors contributed equally to this work

21

Trends in Cognitive Sciences, Month 2024, Vol. xx, No. xx <https://doi.org/10.1016/j.tics.2024.01.011> 1
© 2024 Elsevier Ltd. All rights reserved.



CALIFORNIA REPUBLIC

CA AB 2885

Section 11546.45.5 of the California Government Code (2024-09-28):

“Artificial intelligence” means an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments.

an engineered or machine-based system that varies in its level of autonomy and that can, for explicit or implicit objectives, infer from the input it receives how to generate outputs that can influence physical or virtual environments



Intelligence vs. knowledge or task effectiveness

More commonly ...

We confuse being knowledgeable with being intelligent

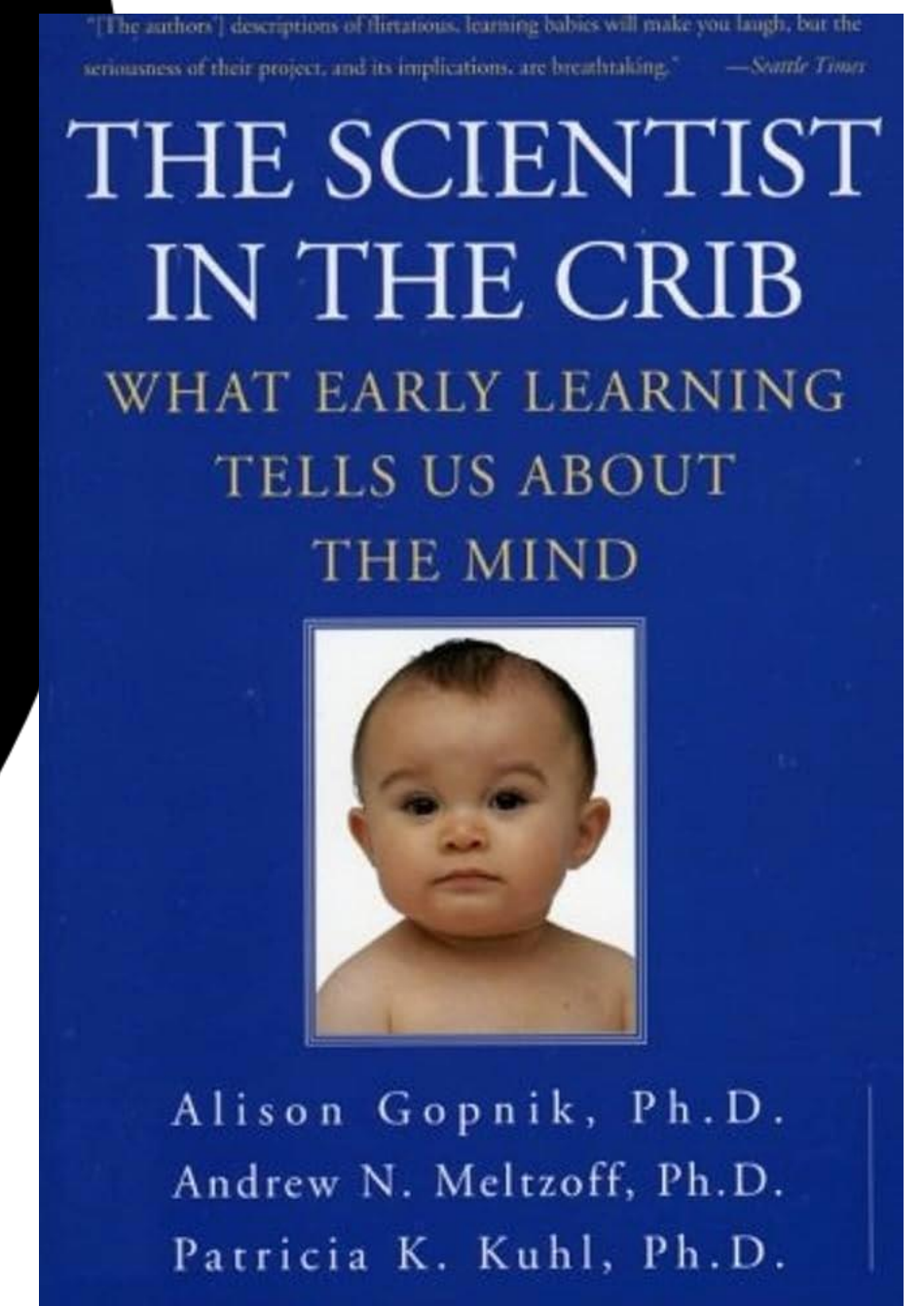
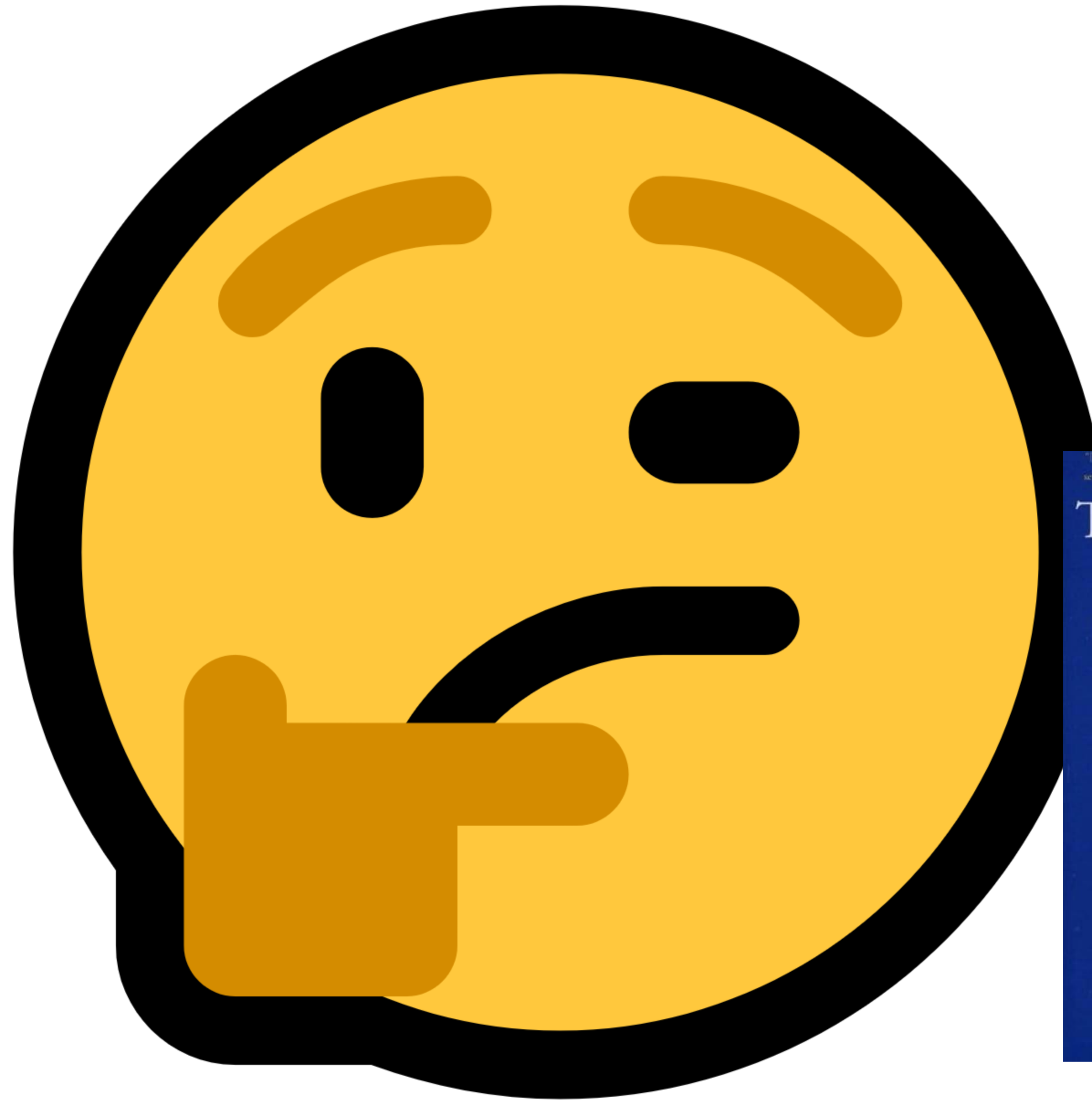
Or the ability to do tasks?

AGI: “ability to perform most (cognitive/ economically valuable) tasks at or above the level of the average person”

Intelligence

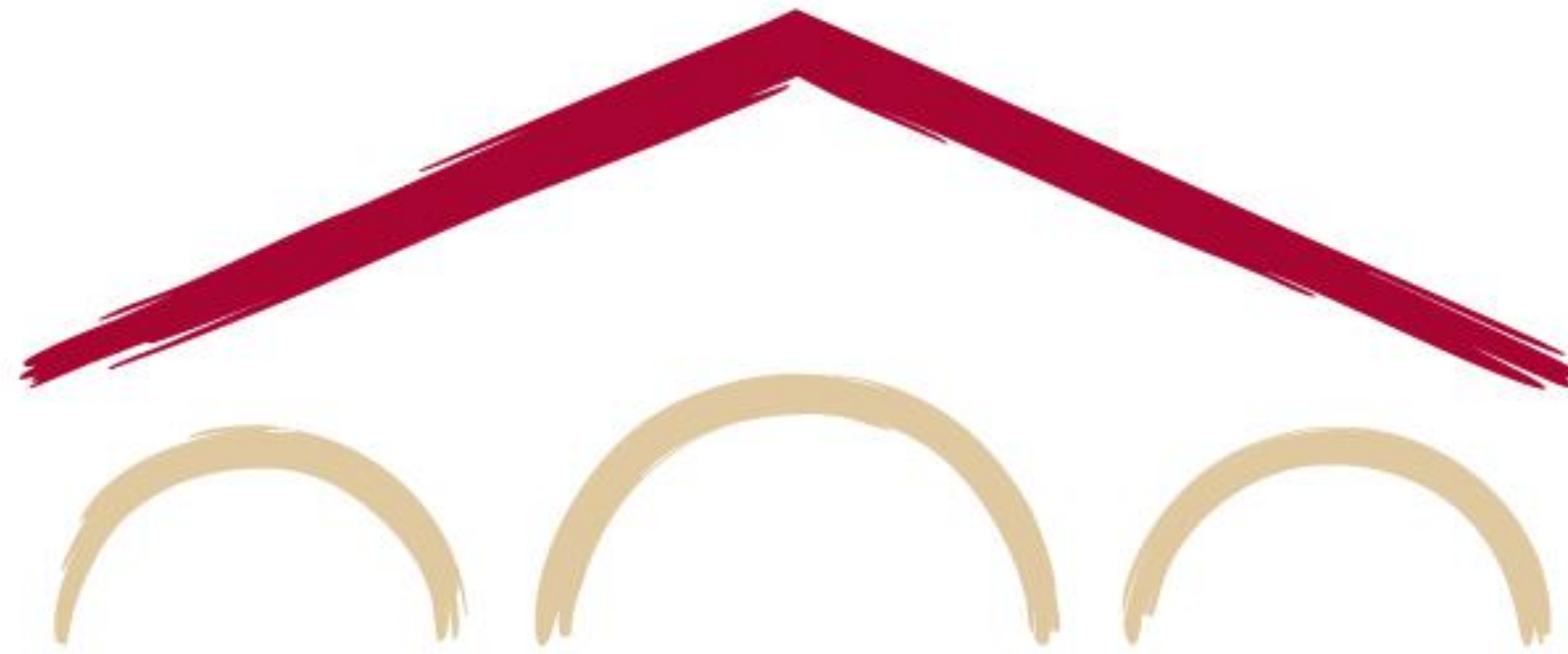
“the ability to learn and perform a range of techniques to solve problems and achieve goals, **appropriate to the context in an uncertain, ever-varying world”**

– Christopher Manning (2020)



MoEUT

Mixture-of-Experts Universal Transformers



Róbert Csordás,
Kazuki Irie, Jürgen Schmidhuber,
Christopher Potts, Christopher Manning

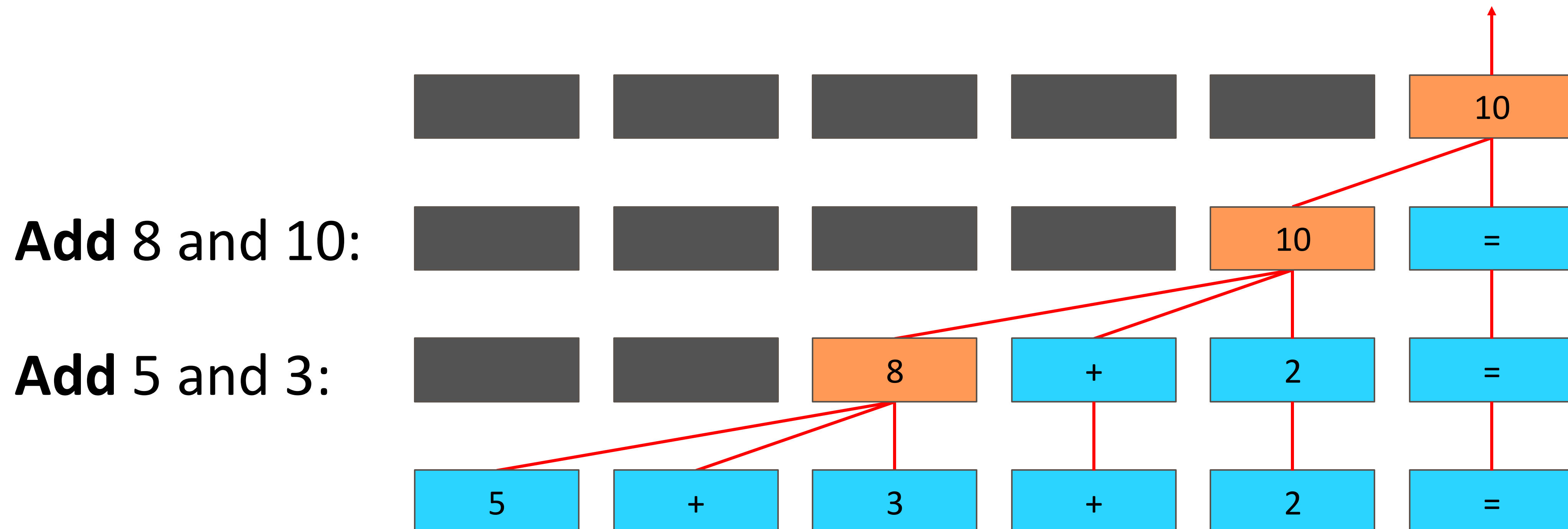
 **GitHub**/robertcsordas/moeut



Motivation: Systematic Generalization

- Transformers can only “see” the results of computations done in earlier layers
- It is often necessary to reuse the same computation in multiple steps

Such a computation cannot be done by one “circuit” with regular Transformers



Universal Transformers: Share weights at each level

[Dehghani et al. 2018]

- ✓ Advantages:
 - Recurrence should result in better systematic generalization
 - Reusing learned “circuits” should result in more parameter efficiency
- ✗ Disadvantages:
 - Layer sharing drastically reduces parameter count
 - Increasing layer sizes is prohibitively expensive

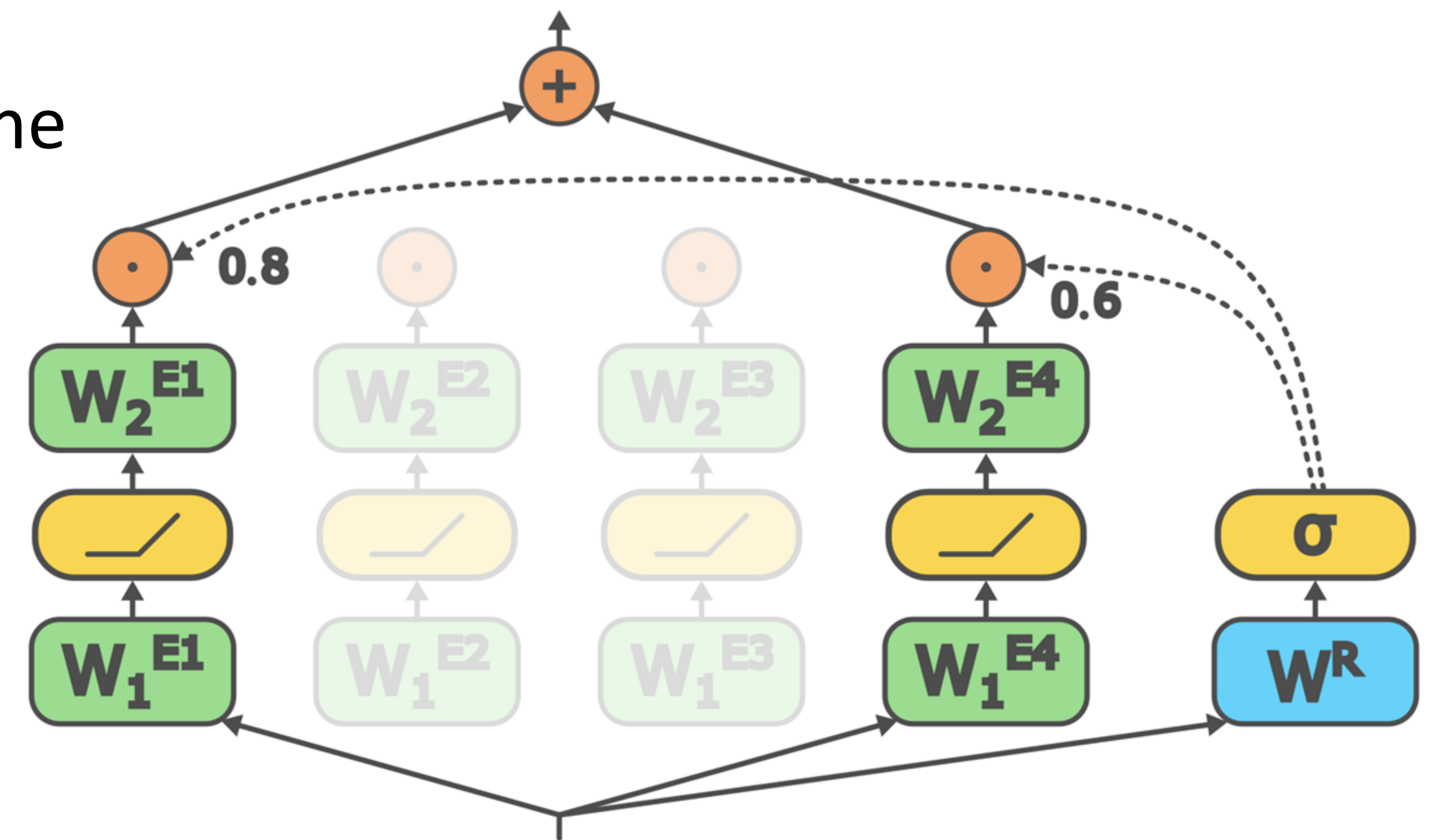
Task	Prior Work	Transformer	Universal Transformer	Rel. Pos. Enc. Transformer	Rel. Pos. Enc. Univers. Transformer
SCAN (length cutoff = 26)	0.00	0.30	0.21	0.72	1.00
CFQ Output length	0.66	0.57	0.77	0.64	0.81
PCFG Systematicity Split	0.72	0.87	0.93	0.89	0.85
COGS	0.35	0.80	0.78	0.81	0.77
Math: add_or_subtract	0.91	0.89	0.94	0.91	0.97

Extension 1: MoE MLP: σ -MoE

- A Mixture of Experts can act as a large layer, without computation overhead.

σ -MoE (Csordás et al. 2023, Approximating Two-Layer Feedforward Networks for Efficient Transformers)

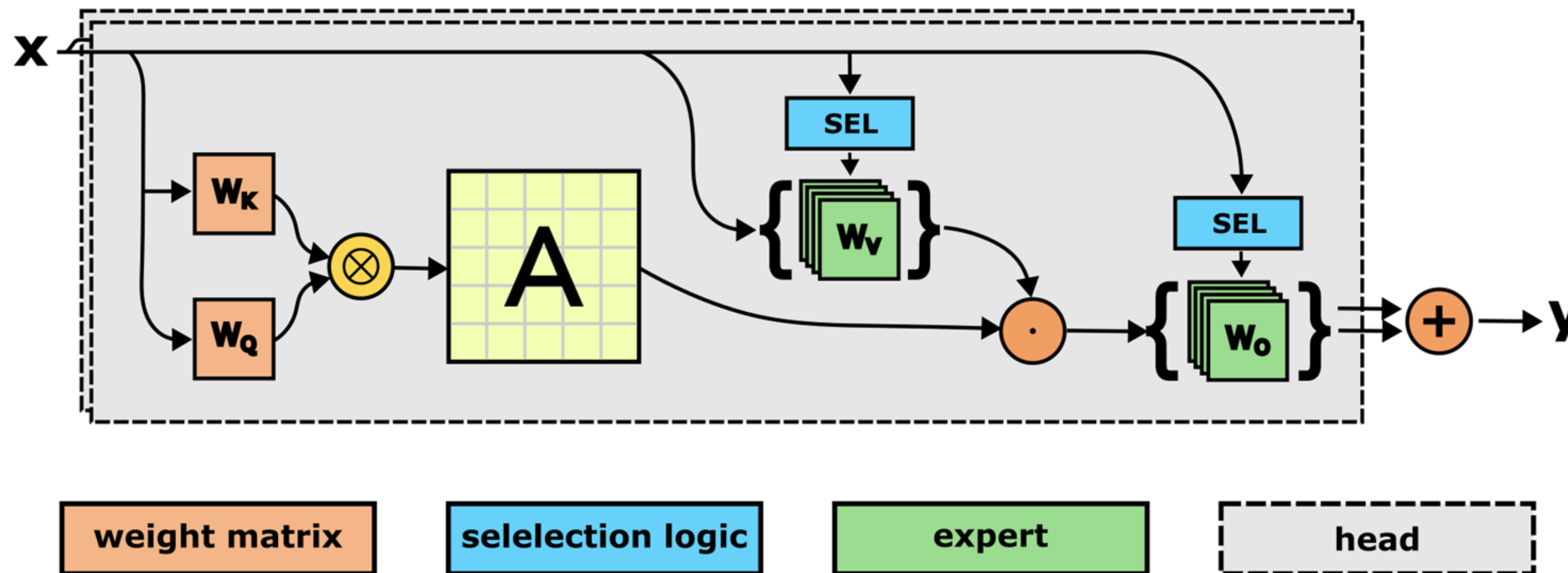
- Many, **tiny** experts
- Each has not enough capacity to work alone
- This helps prevent expert collapse
- Experts have to be composed.
- **Non-competitive** sigmoid activations
- This also helps avoid model collapse
- Goal is to simulate a big MLP
- Mid-MLP activations should not compete
- Weak selection entropy regularization



Extension 2: MoE Attention: SwitchHead

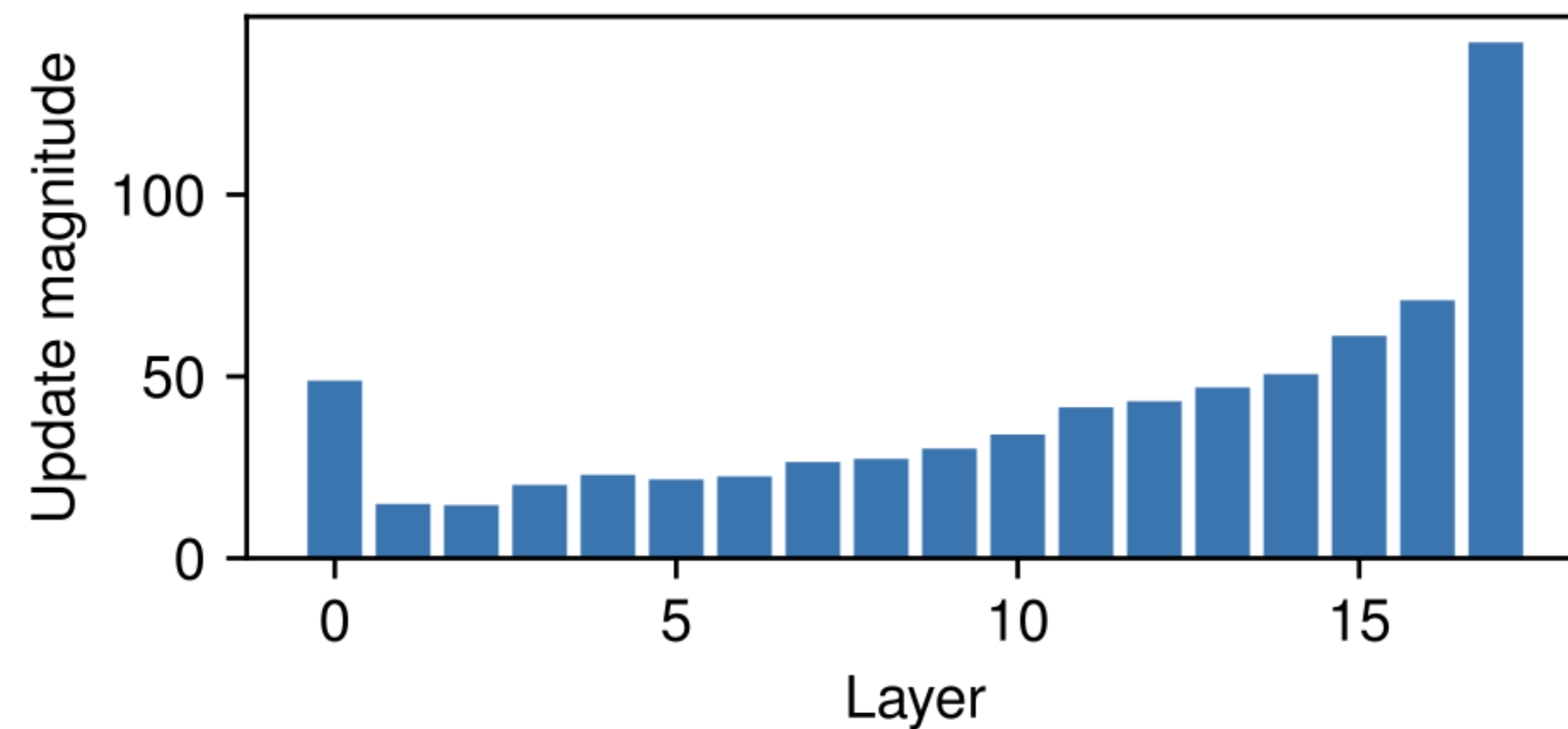
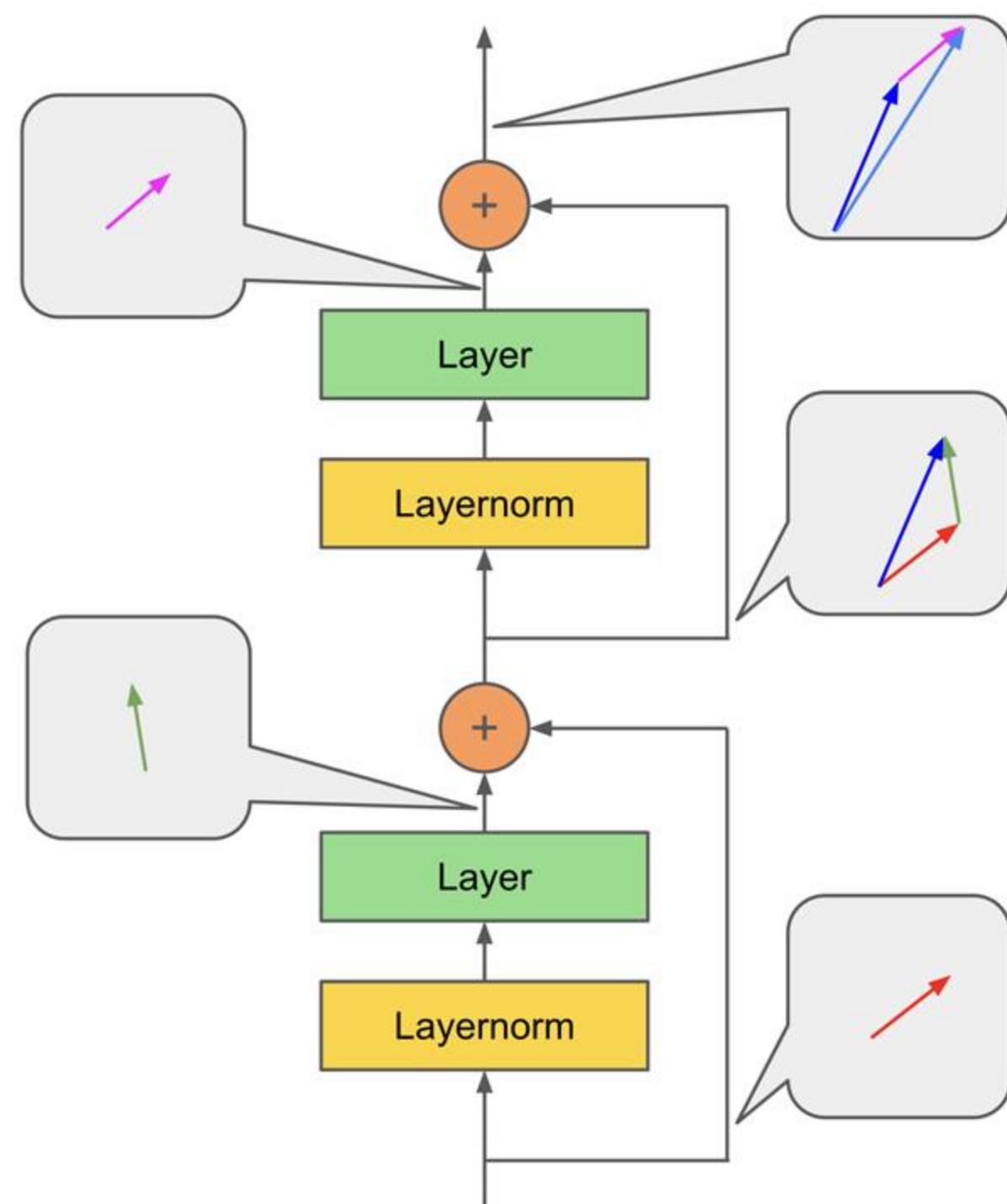
[Csordás et. al, 2024: SwitchHead: Accelerating Transformers with Mixture-of-Experts Attention]

- Difficulty: Naive solution requires selecting N^2 Q/K **pairs**, saving no compute.
- **Expert O and V projections** allows drastically reducing the total number of heads



Extension 3: Residual Norm Growth: The Problem

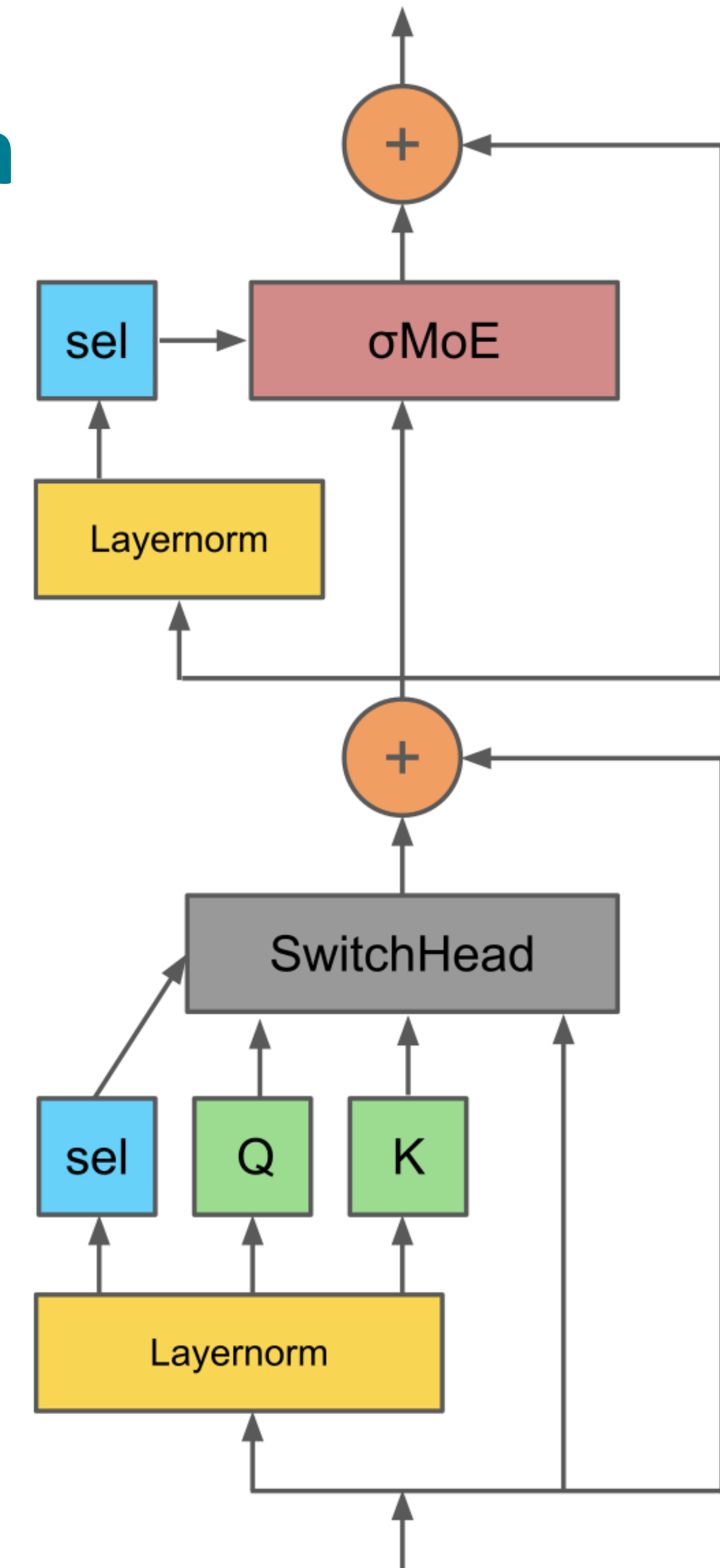
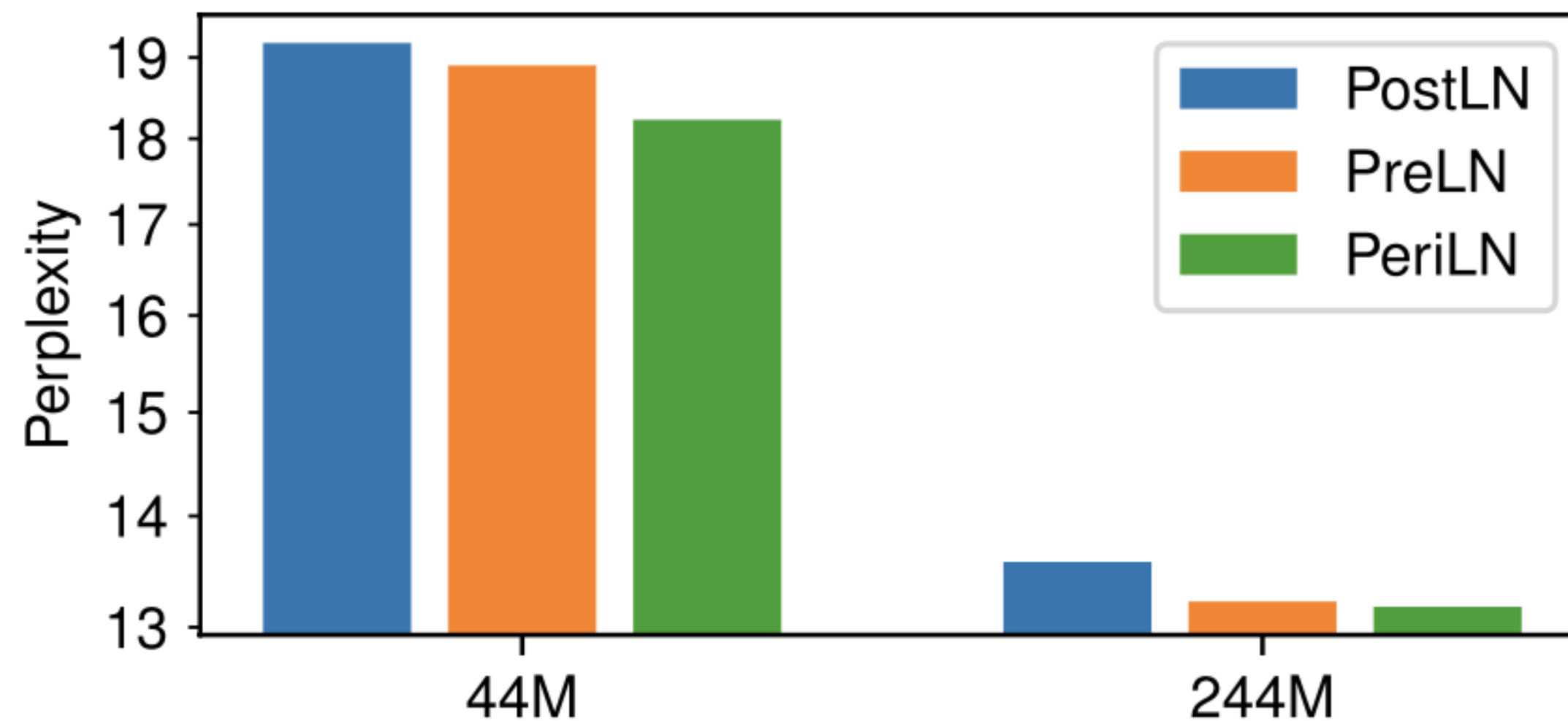
- (The widely used) Pre-layernorm Transformers repeatedly add to the residual
- Thus the norm of the residual grows. Later layers learn bigger updates to be able to influence the residual
- This is not possible for shared-layer models; growth tanks their performance



FFN output norm in a PreLN Transformer

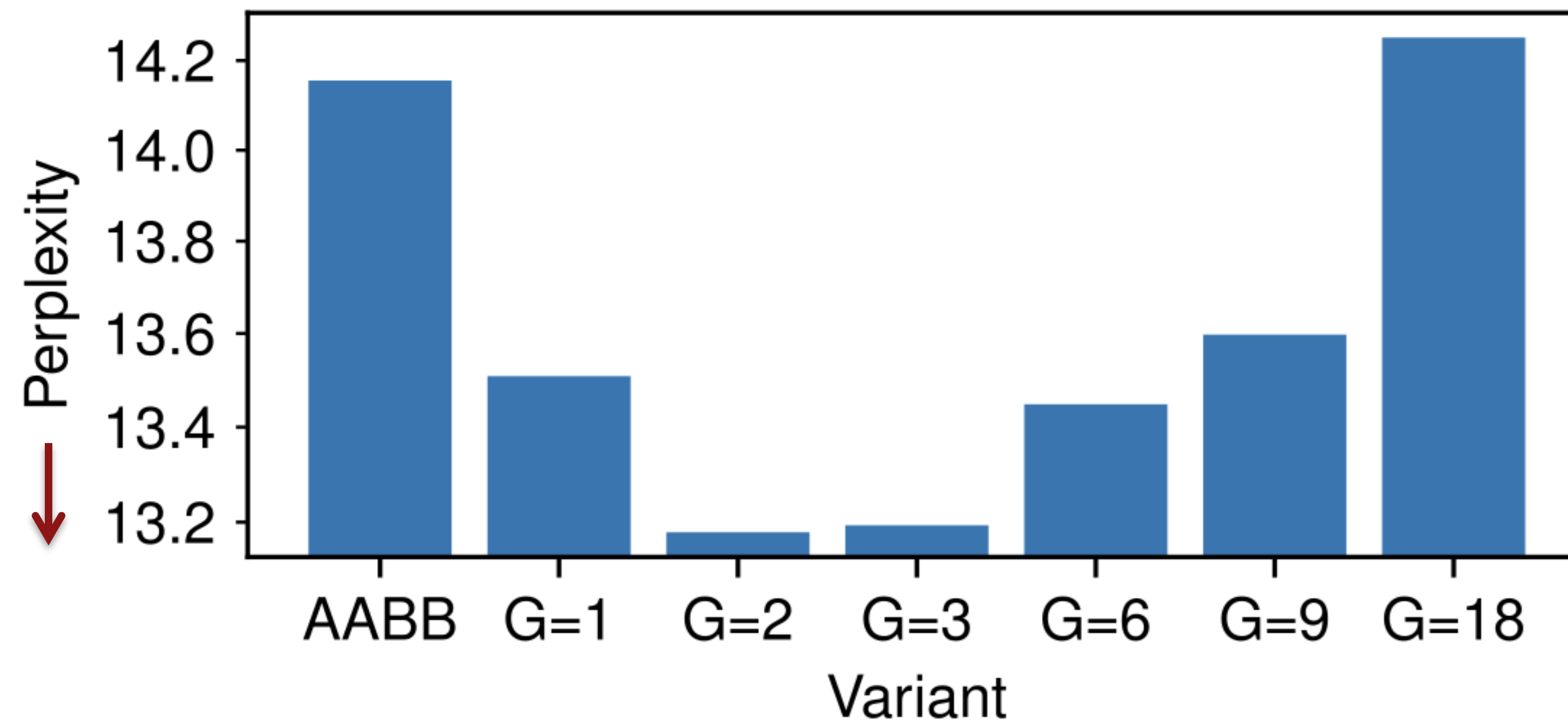
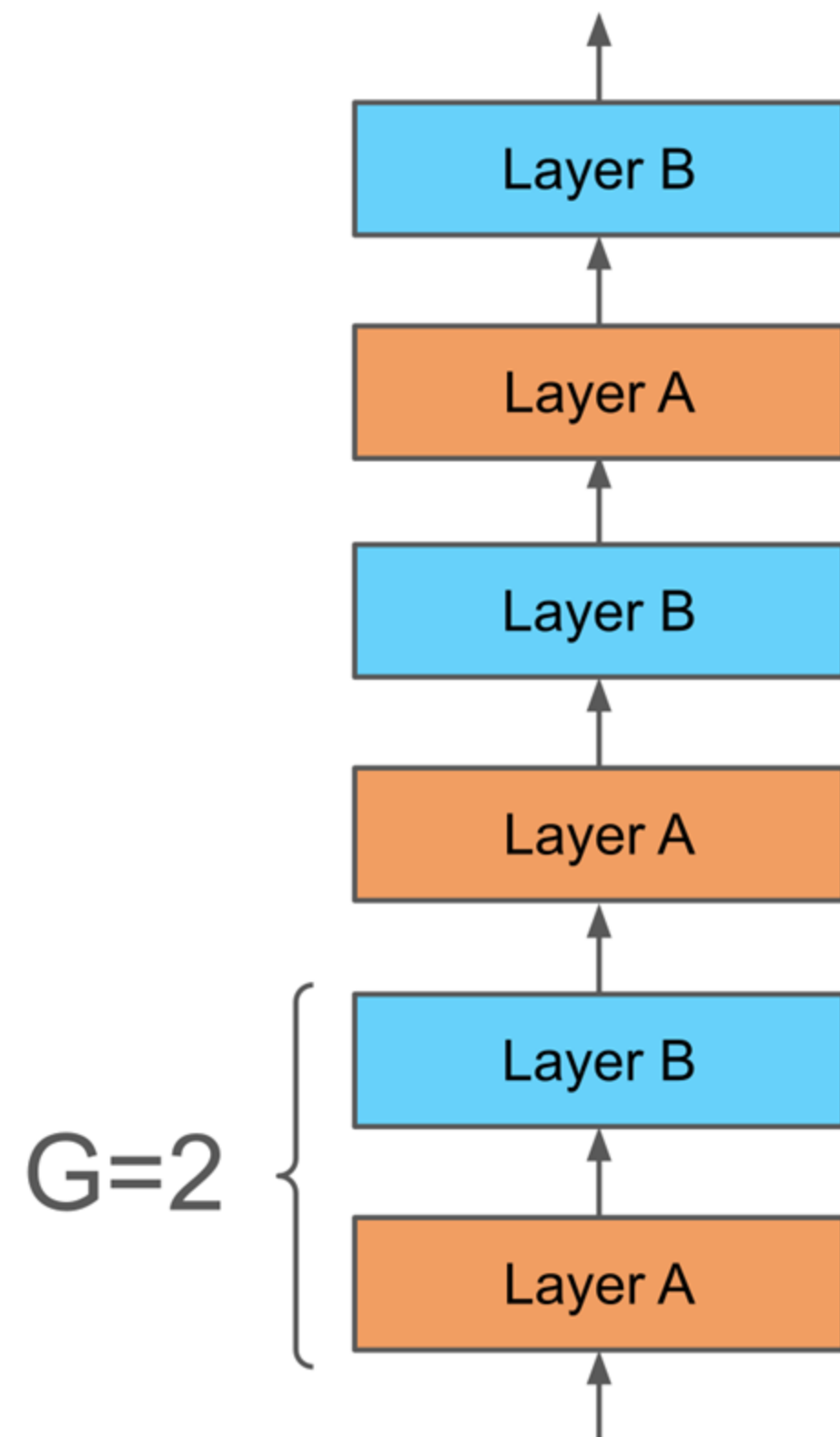
Extension 3: Residual Norm Growth: Solution

- The output norm of all layers should be proportional to their input
- Use **LayerNorm** only before nonlinearities (except ReLU), but not on the main path
- We call this setup **PeriLayerNorm**



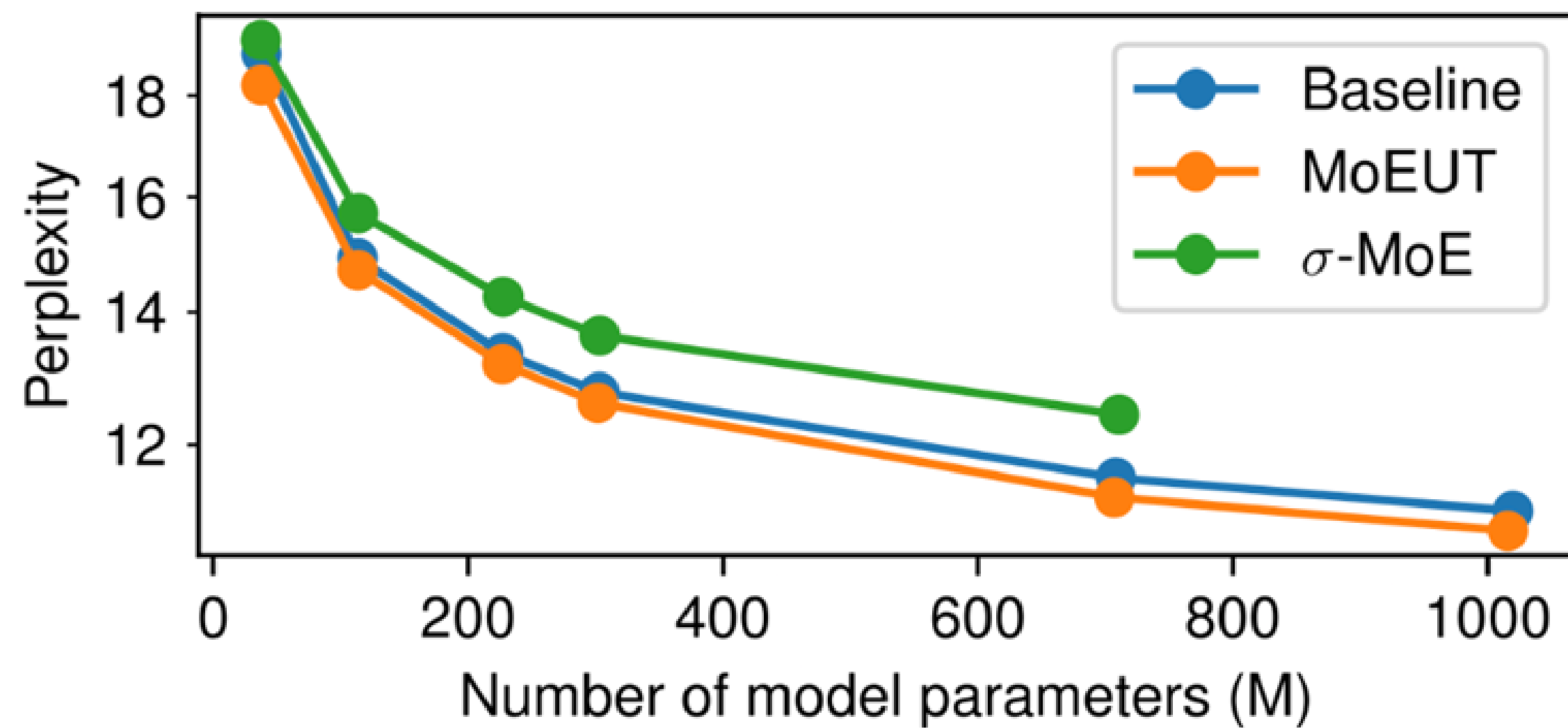
Extension 4: Layer groups

- Motivation: Transformer circuits are often multilayer (e.g., induction heads)
- Repeating a **small group** of layers performs better than a single one
 - Alternating layers (*ABAB*) is better than running them sequentially (*AABB*)

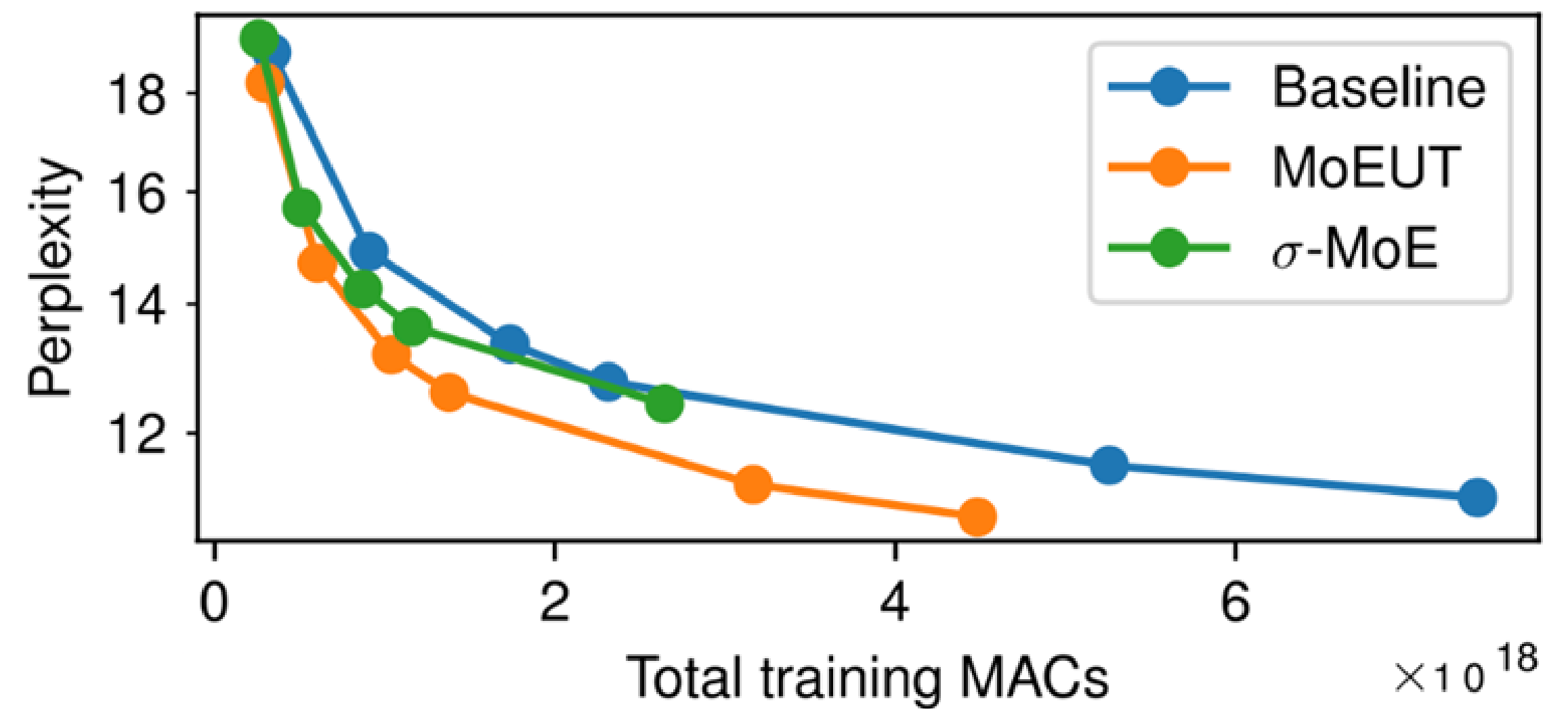


Performance

MoEUT outperforms parameter-matched baselines with half of the MACs



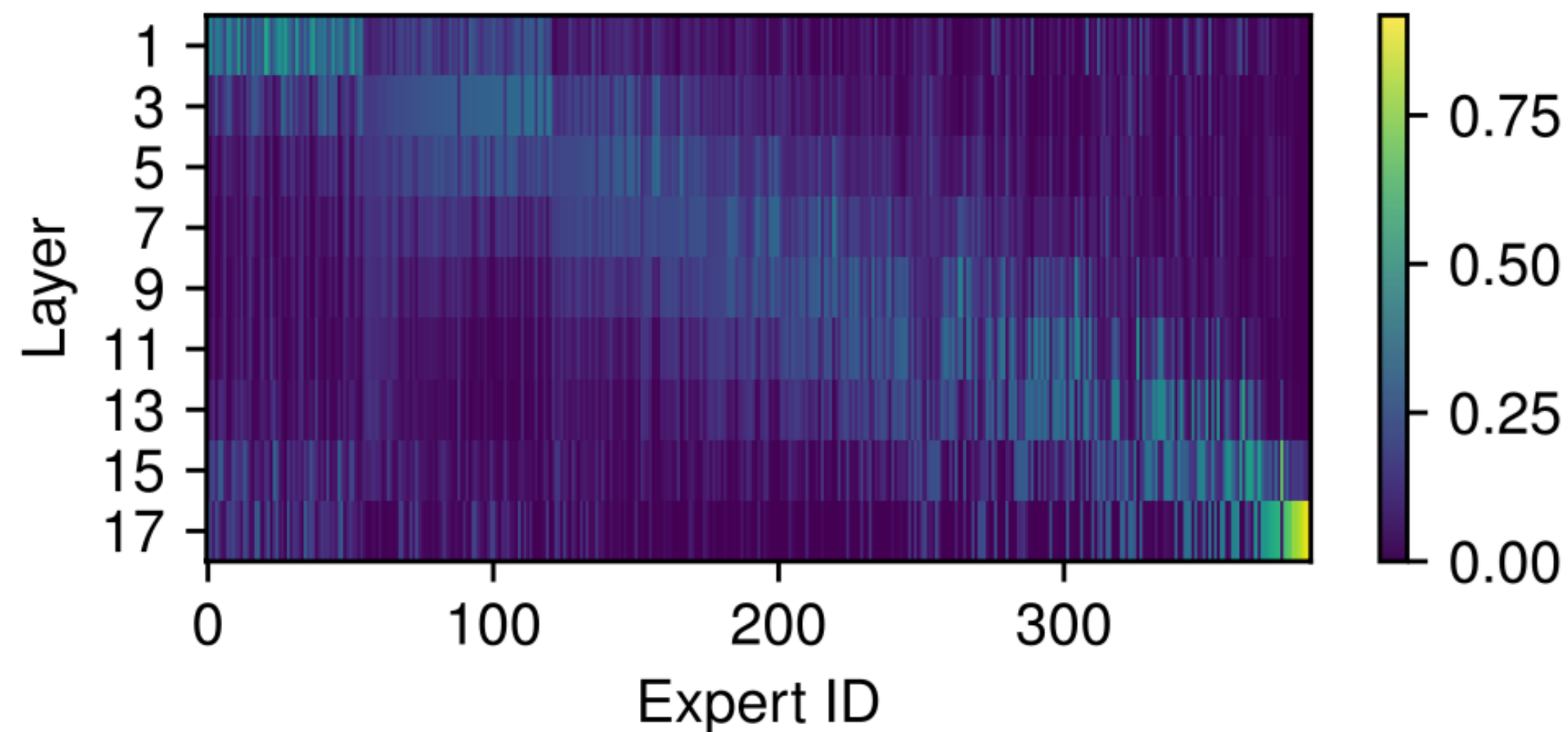
Scaling in the number of parameters.



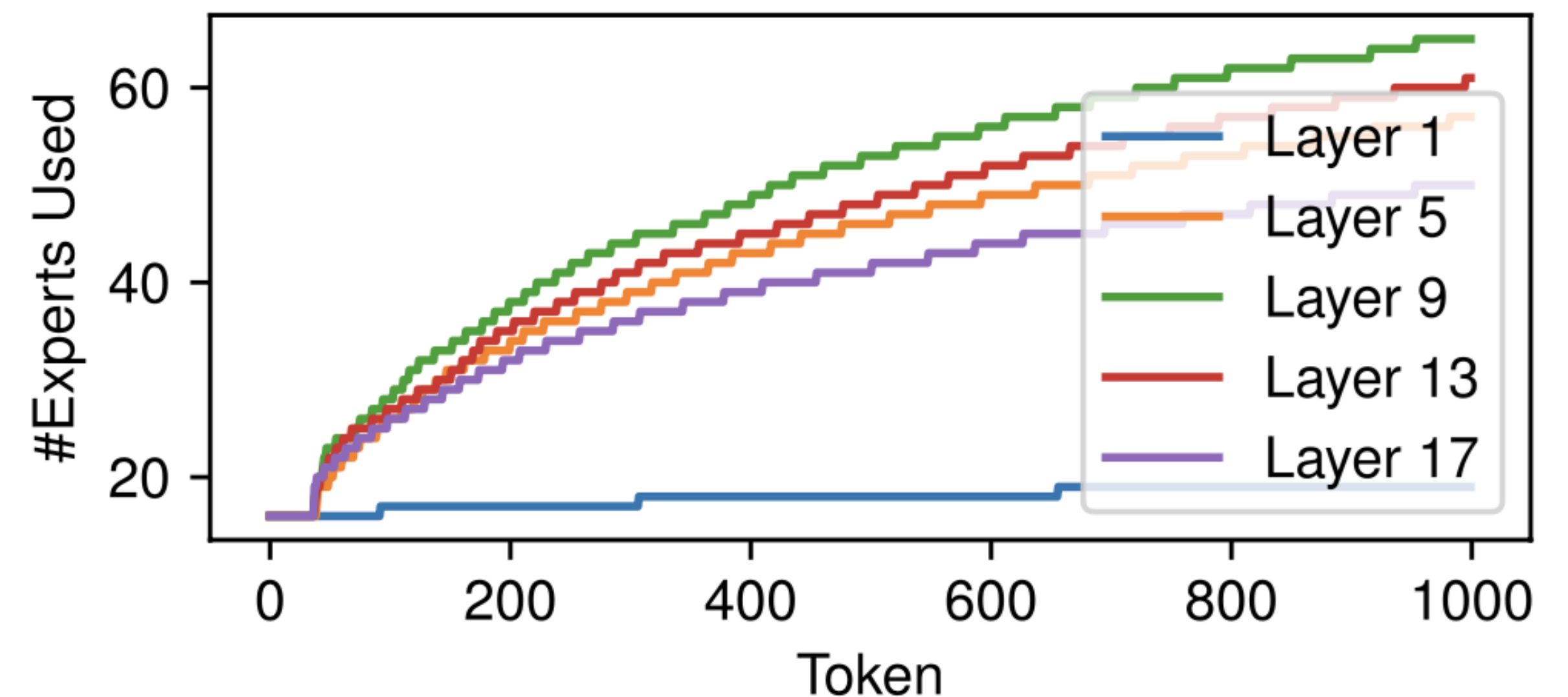
Scaling in the number of MACs used for training.

Analysis

- Experts specialize to be shared or to be layer-specific
- The model takes into account the context for routing: Inner layers use a more diverse set of experts



Distribution of the preferred layers by individual experts

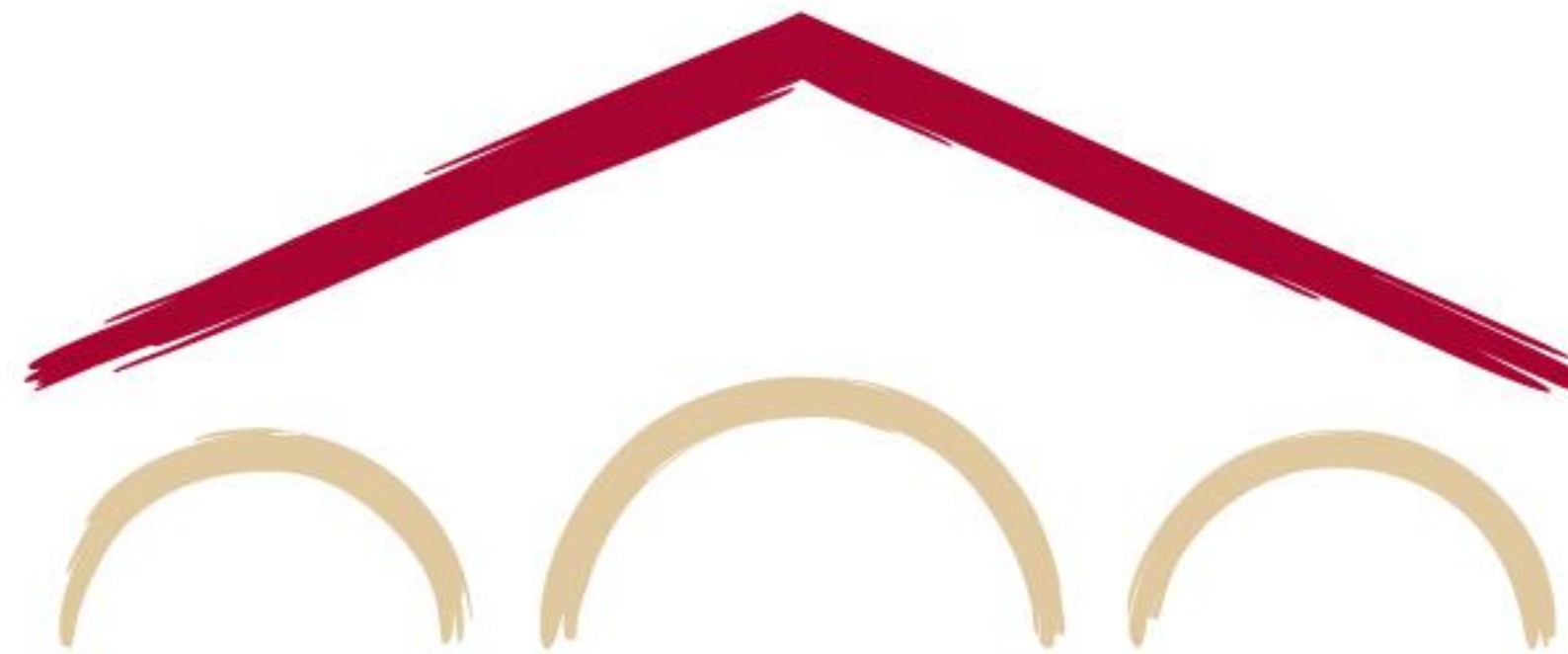


The number of experts used by individual tokens in different layers

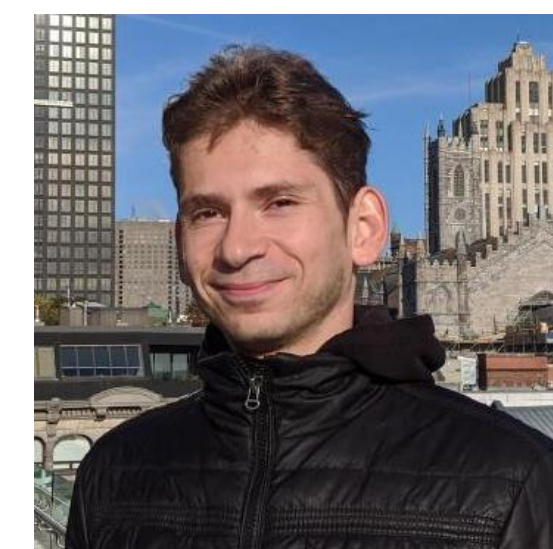
Summary

- Mixture-of-Experts Universal Transformers (MoEUTs or “moots”) can scale in a computationally and memory efficient way
 - We use recent work to apply MoEs for both the feedforward and attention layers
 - We add two innovations:
 - Layer grouping: recurrent stacking of groups of MoE layers
 - A peri-layernorm scheme between pre- and post-layernorm, which applies layer norm only before linear layers that immediately precede sigmoid or softmax activations
- **MoEUTs enable parameter- and resource-efficient UT LMs, for the first time outperforming standard Transformers on medium-scale language modeling tasks**
- Recurrence is essential to the good performance of MoEUTs
- MoEUT models demonstrate good zero-shot performance on downstream tasks:
 - BLiMP and Children’s Book Test, Lambada, HellaSwag, PIQA and ARC-E

Linguistic Agents Learning from Interaction: Learning Grounded Interactions on the Web



Shikhar Murty, Pete Shaw, Mandar Joshi, Kenton Lee,
Dmitry Bahdanau, Christopher D. Manning



An Agent

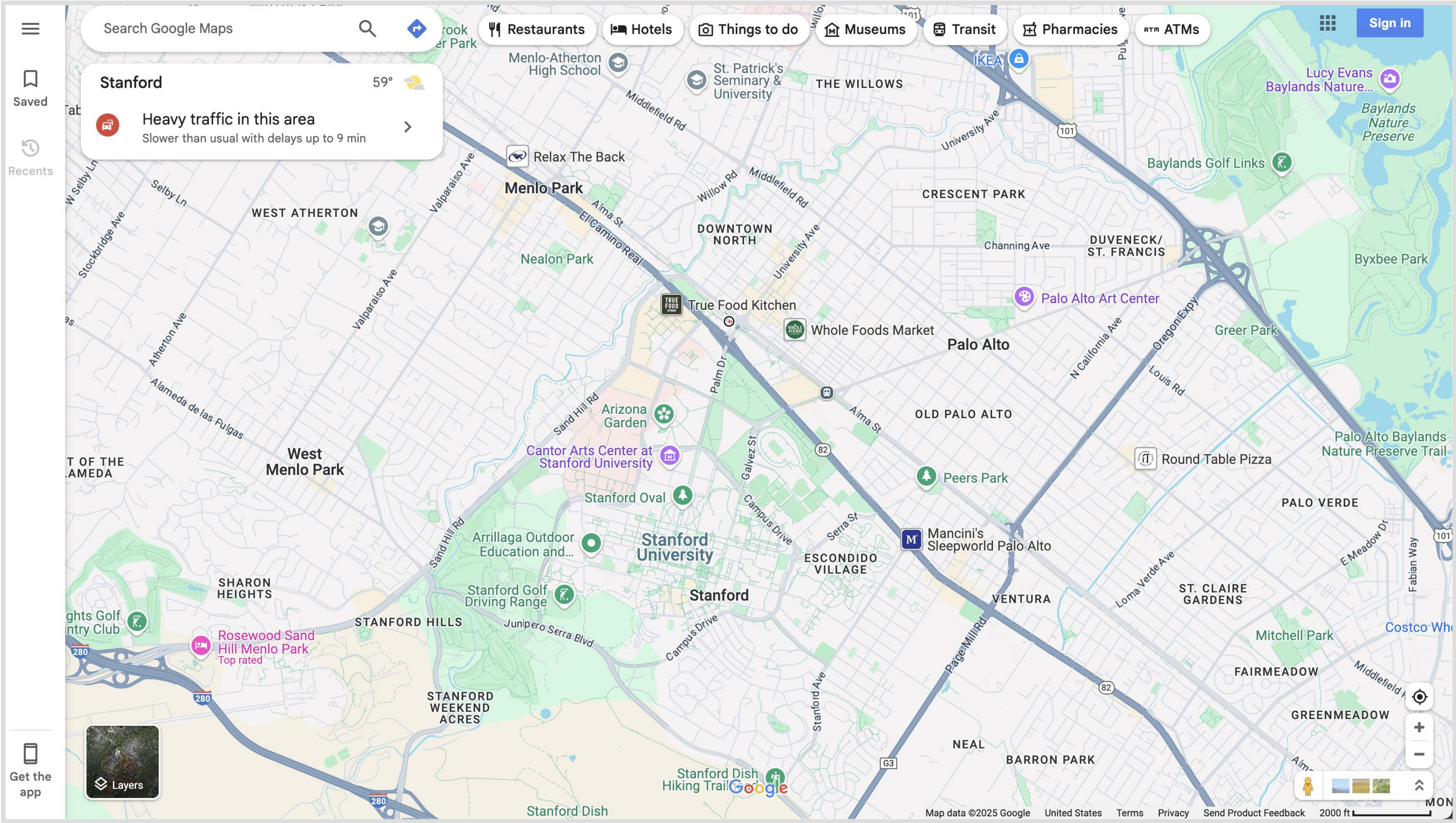
“Anything that perceives its environment (using sensors) and acts upon it (using actuators)” – Russell and Norvig (2002)

A Linguistic Agent

A foundation model used as an agent that understands and operates primarily in the domain of human language

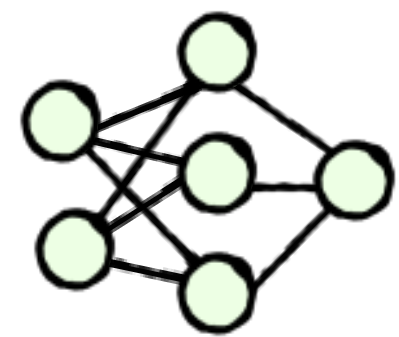


Linguistic Agents

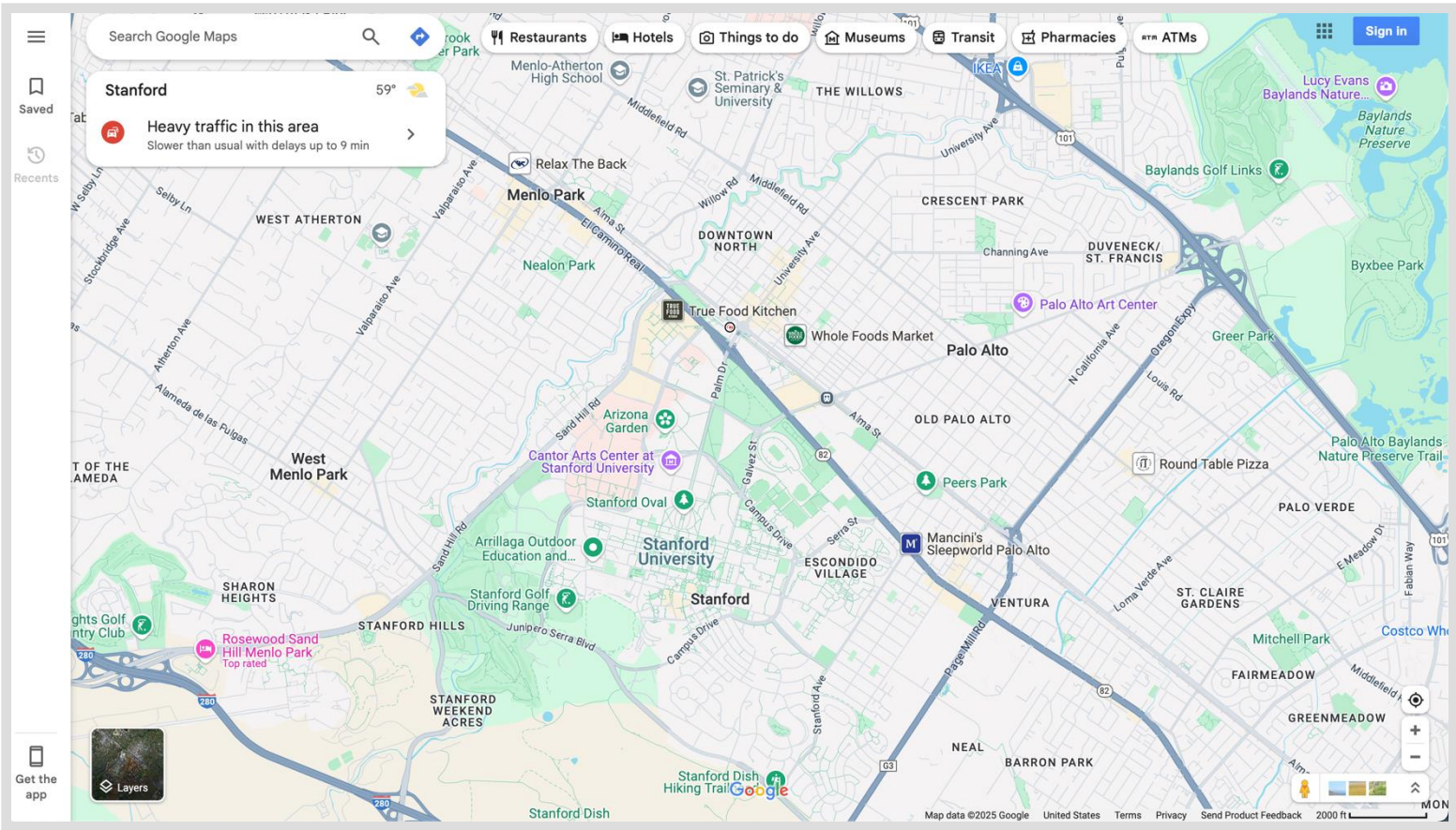
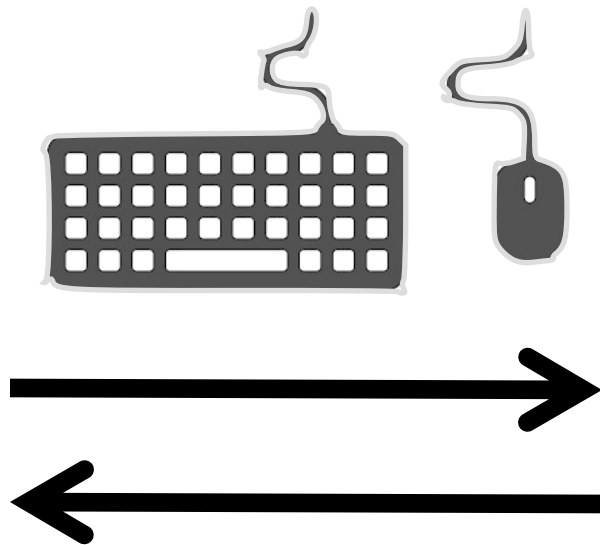


Find the fastest route from a hotel near the Golden Gate Bridge to a café with a view of the bay on Market Street

Linguistic Agents

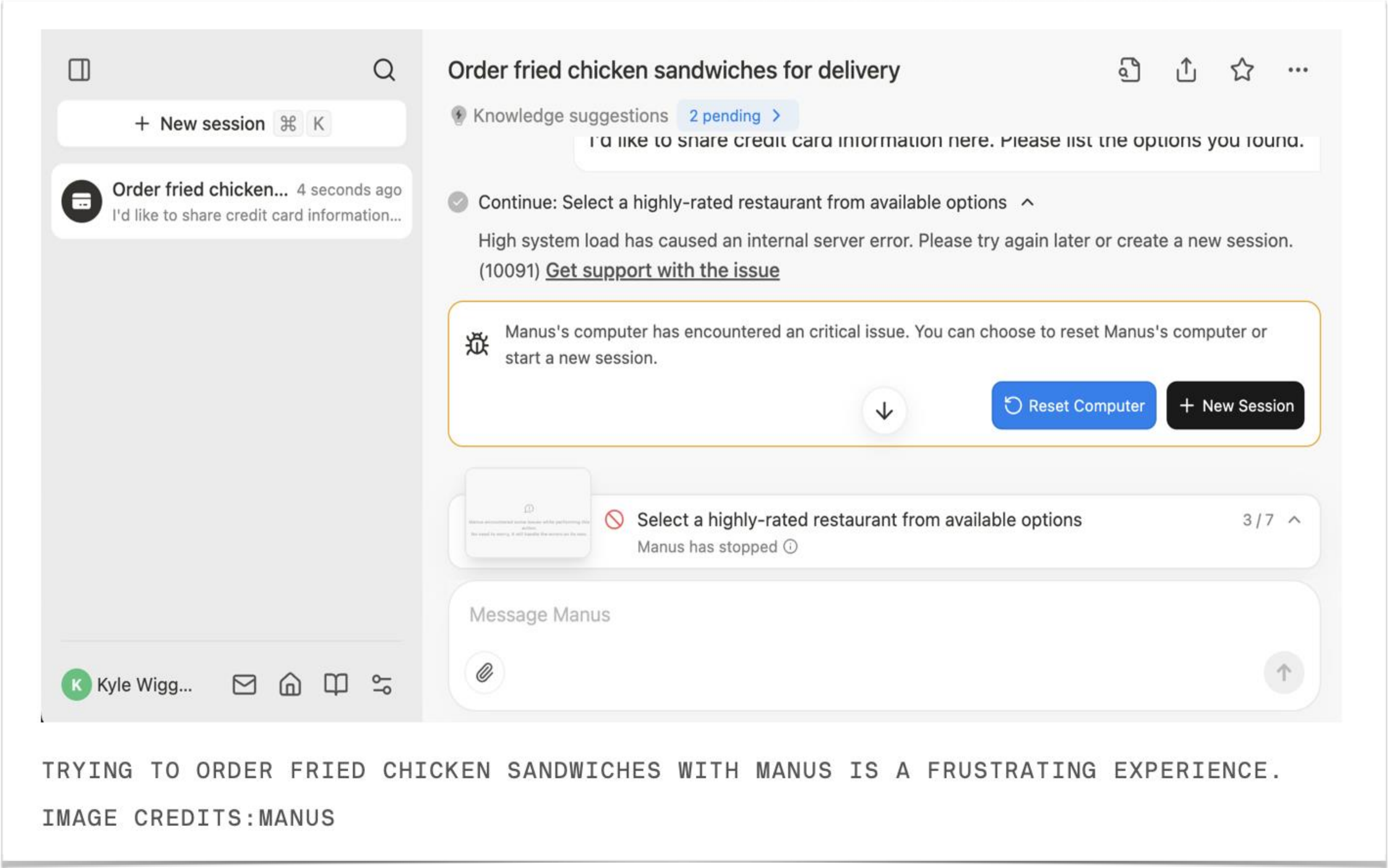


Find the fastest route from a hotel near the Golden Gate Bridge to a café with a view of the bay on Market Street



Type ... on ..., Click on ...,
Choose ... from dropdown, ...

Linguistic Agents: Reality



Linguistic Agents: Reality

stephentoub commented 2 days ago

Member

...

Please fix the style issues causing build failures, and move the new test into an existing test file.

Fix style issues and move balancing group test to existing file

548aed4

Copilot commented 2 days ago · edited

Contributor

Author

...

Please fix the style issues causing build failures, and move the new test into an existing test file.

I've fixed the style issues and moved the test into the existing Regex.Groups.Tests.cs file in commit [548aed4](#). The changes included removing trailing whitespace in Match.cs and properly placing the balancing group test in the RegexGroupTests class.

Warning

Firewall rules blocked me from connecting to one or more addresses

👍 1

👎 1

build-analysis bot mentioned this pull request 2 days ago

Intermittent build failure in AfterSourceBuild: "Could not write state file" #76488

Open

build-analysis bot mentioned this pull request 2 days ago

NuGet failing with Response status code does not indicate success: 503 (Service Unavailable) dotnet/arcade#11723

5 tasks

Open

stephentoub commented yesterday

Member

...

A bunch of regex tests are now failing.

👍 200

🚀 10

👁 8

Fix regex balancing group test failures

fe173fc

Learning from Interaction

To survive, agents must have:
Enough built-in knowledge to survive
The ability to learn

Goal: Intelligent agents that learn from interactions in the world!

Method: Generating interactive trajectories through hierarchical exploration

[1] BAGEL: Bootstrapping Agents by Guiding Exploration with Language

Murty, Manning, Shaw, Joshi, Lee, ICML 2024

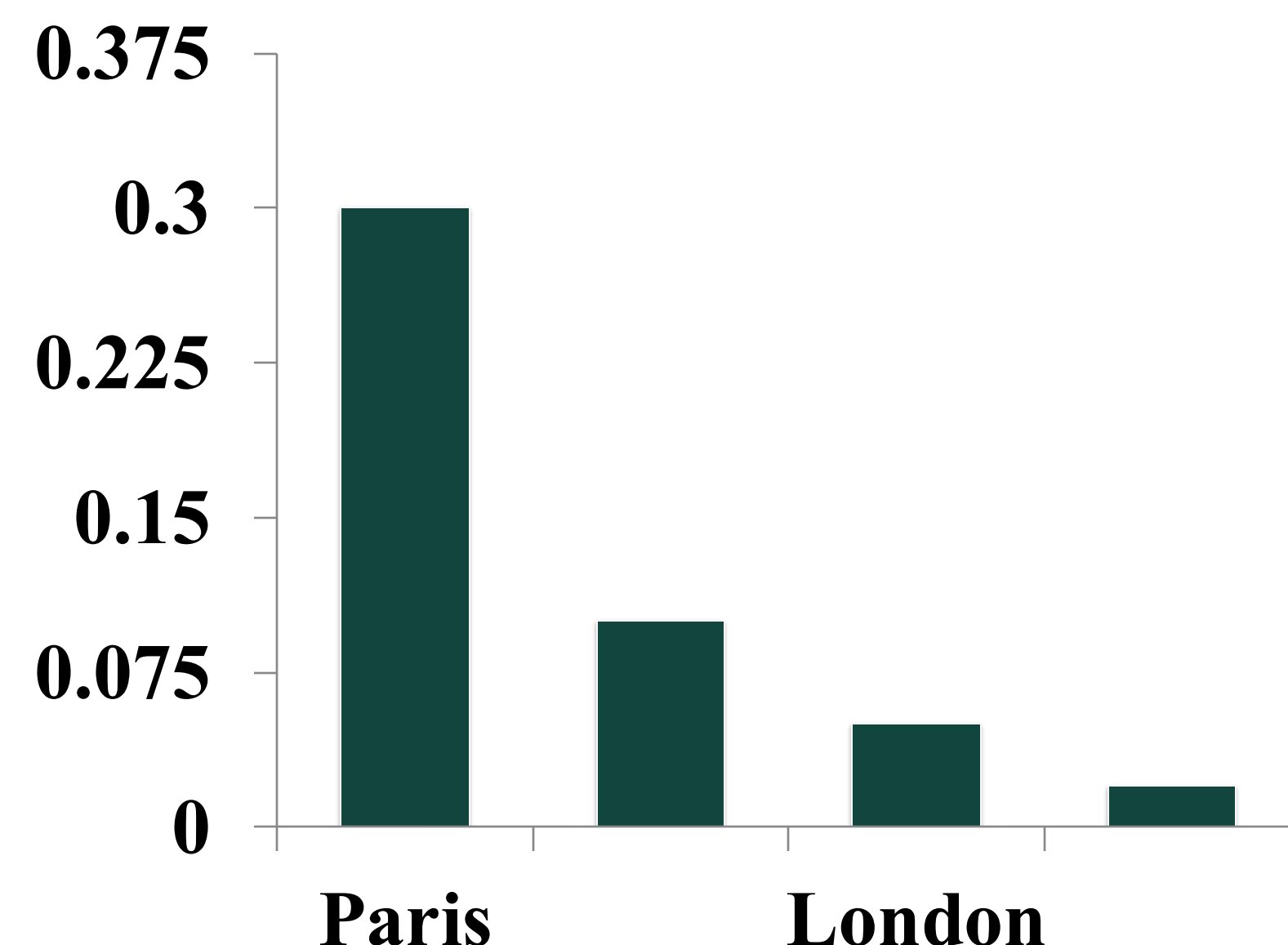
[2] NNetNav: Unsupervised Learning of Browser Agents Through Environment Interaction in the Wild

Murty, Zhu, Bahdanau, Manning, ICLR SSI-FM Workshop 2025



Browser Control with Linguistic Agents

First Attempt: Zero-shot Linguistic Agents



The capital of France is ___

Language Model



Gemini



$$\pi_{\text{LM}}(\cdot \mid g)$$

Language Model
Policy

You are an agent capable of the following actions:

- 1.Type X on Y
- 2.Move mouse to X
- 3.Click on X
- 4.Hover on element X
5. ...

Your objective is to follow user instructions, by mapping them into a sequence of actions.

Instruction: {g}

Here is a history of your interactions:

s1:

s2:

...

After executing these actions, you observe the following HTML state: <AXTree state>

Now, think about your next action:

Thought: [model-pred]

Now, take an action:

Action: [model-pred]

Action
space

Instruction

History

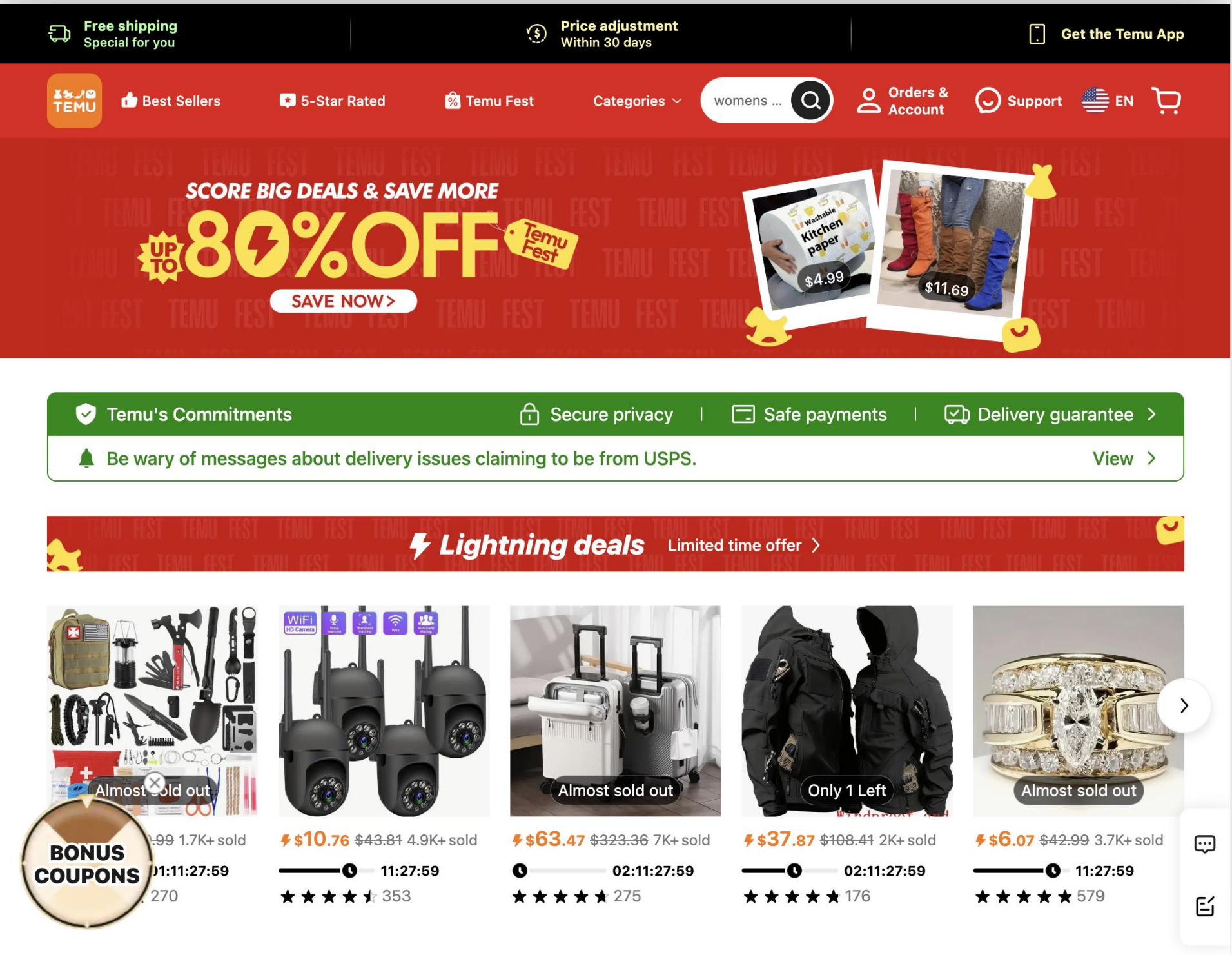
Observation

Action
string

Browser Control with Linguistic Agents

First Attempt: Zero-shot Linguistic Agents

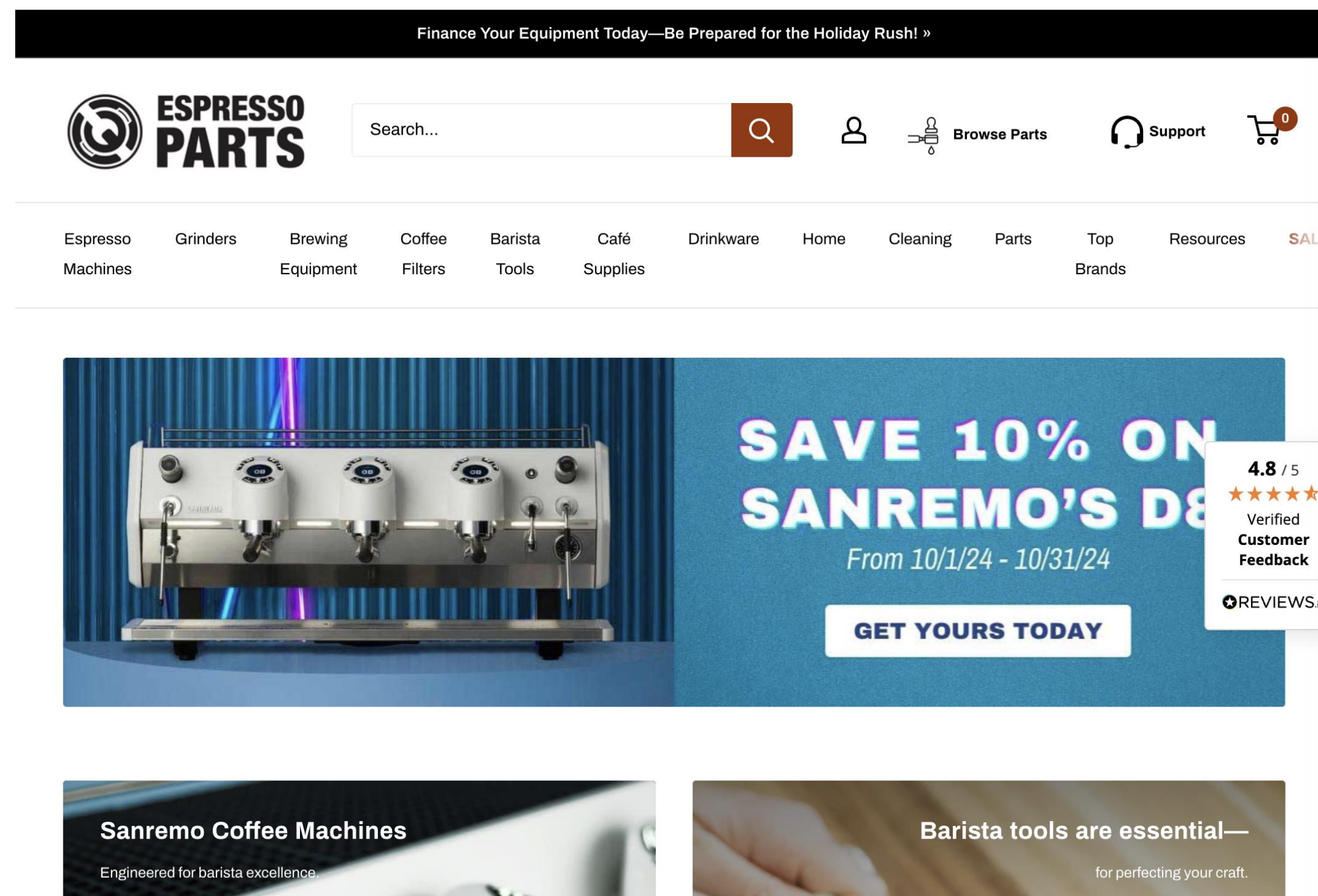
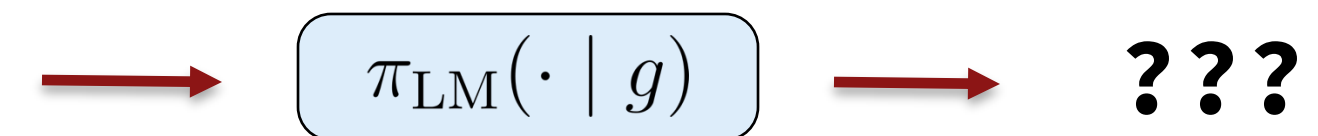
Cancel shampoo subscription from the list of “subscribe and save” items $\longrightarrow \pi_{\text{LM}}(\cdot \mid g) \longrightarrow ???$



Browser Control with Linguistic Agents

First Attempt: Zero-shot Linguistic Agents

Order monthly delivery of descaler

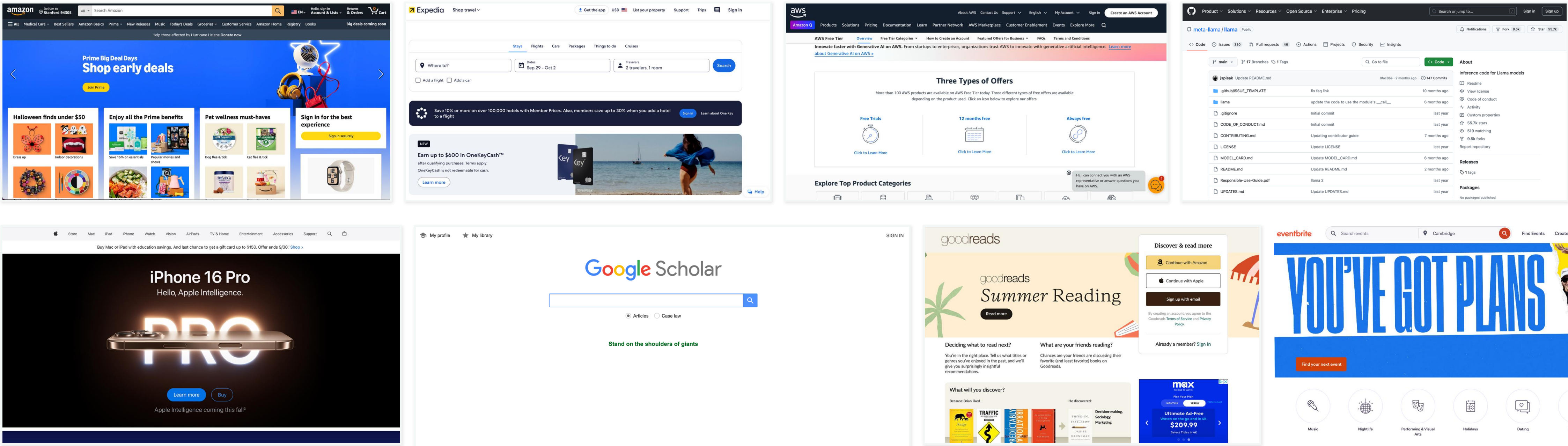


LMs cannot know how to navigate to the right section of a website without prior knowledge

How can we provide this prior knowledge about environments?

Browser Control with Linguistic Agents

Second Attempt: Behavioral Cloning with Human demonstrations

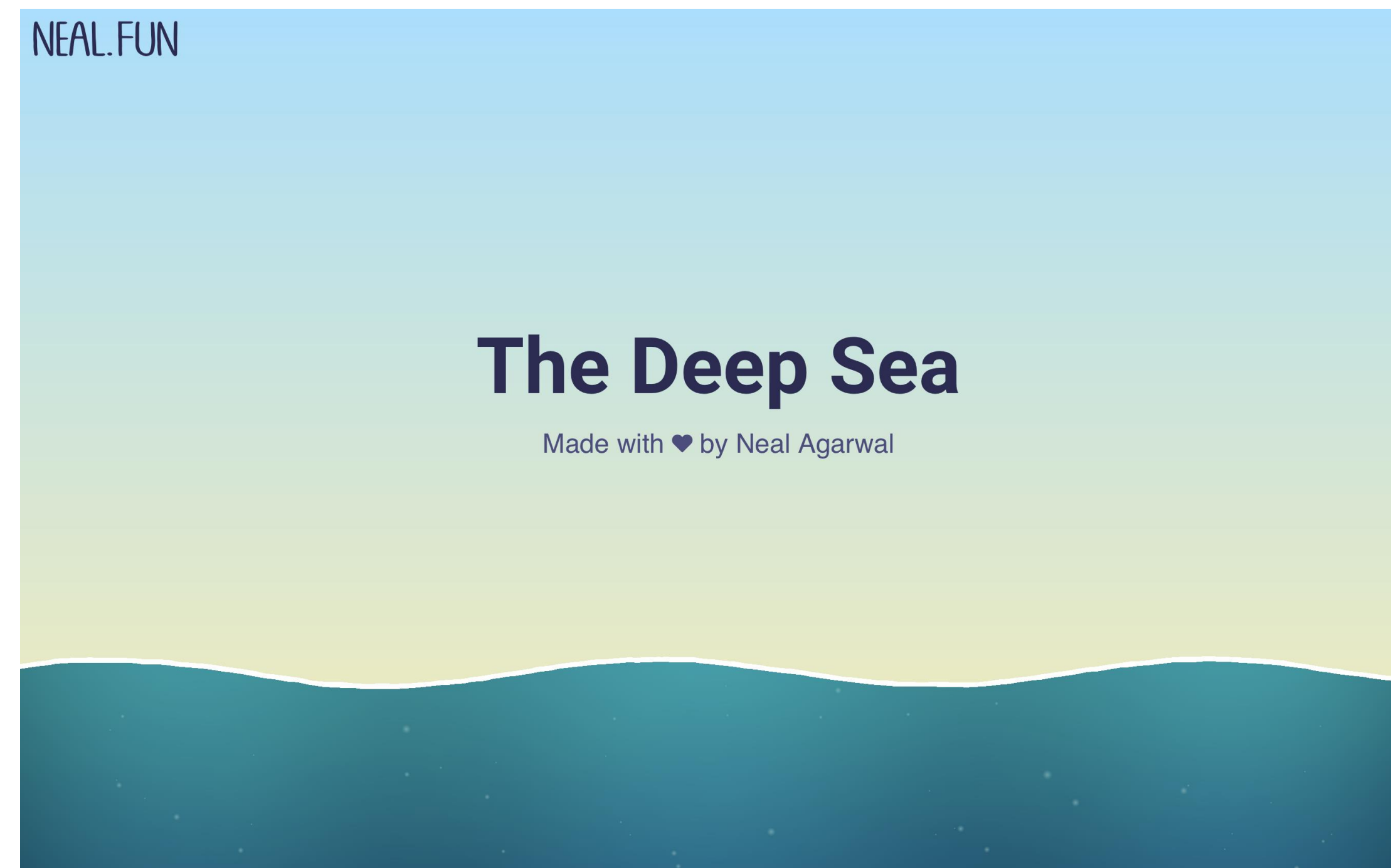


Collecting human demonstrations that cover most use-cases across websites is expensive
Humans learn about new web-interfaces simply by interacting with them.
Can we do the same with models?

Our work: Third Attempt, NNetNav, Discovering Goals in Web UI

Main Challenge: Given a new URL, it's not even clear what kinds of goals exist!

<https://neal.fun/deep-sea/>



[NNetNav: Unsupervised Learning of Browser Agents Through Environment Interaction in the Wild
Murty, Zhu, Bahdanau, Manning, ICLR SSI-FM Workshop 2025

Our work: Third Attempt, NNetNav, Discovering Goals in Web UI

Main Challenge: Given a new URL, it's not even clear what kinds of goals exist!

<https://neal.fun/deep-sea/>



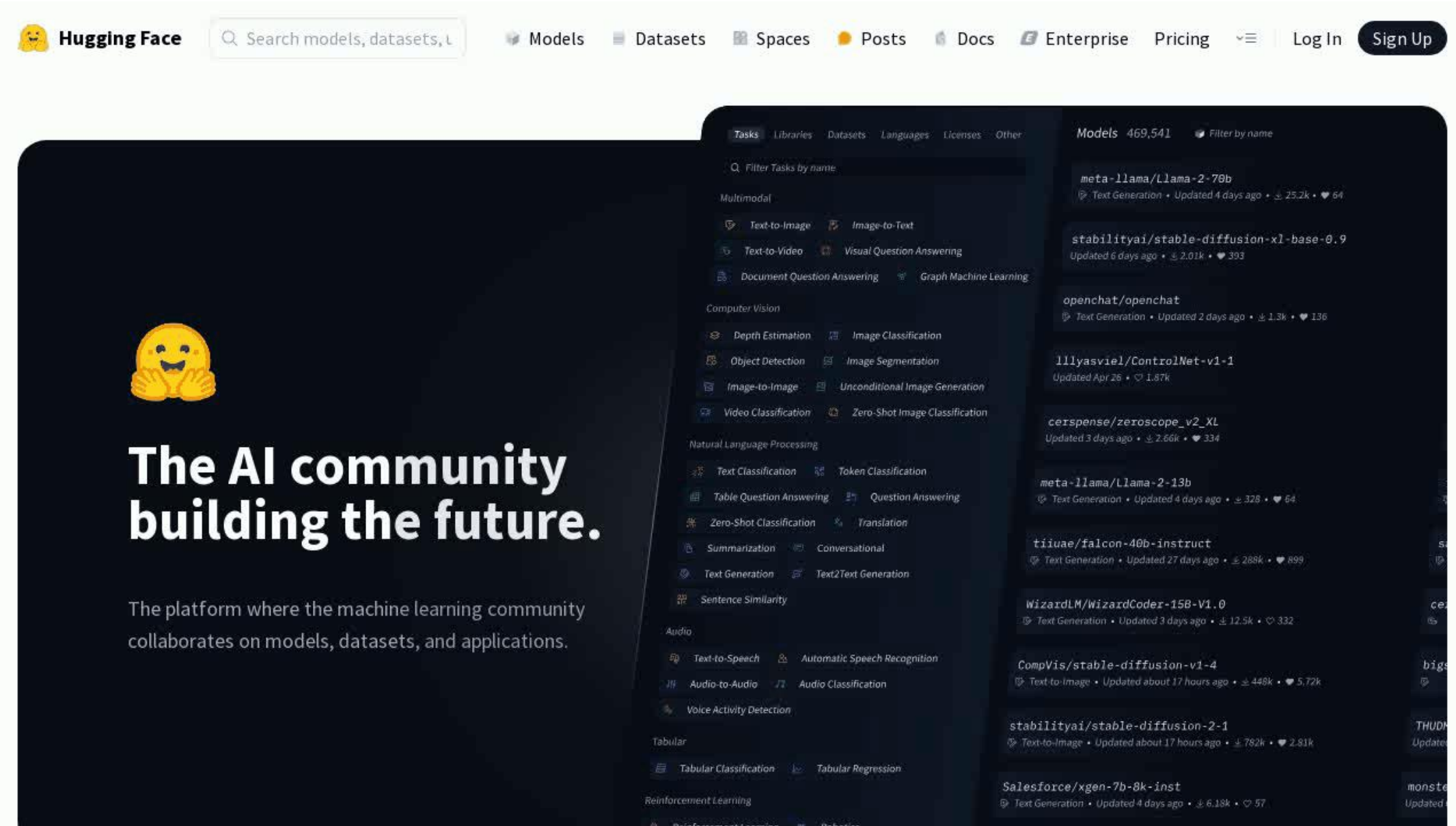
We can learn about goals on unknown digital inferences only by interaction

NNetNav: Discovering Goals on Web UI

An Exploration Policy interacts with a website to generate trajectories

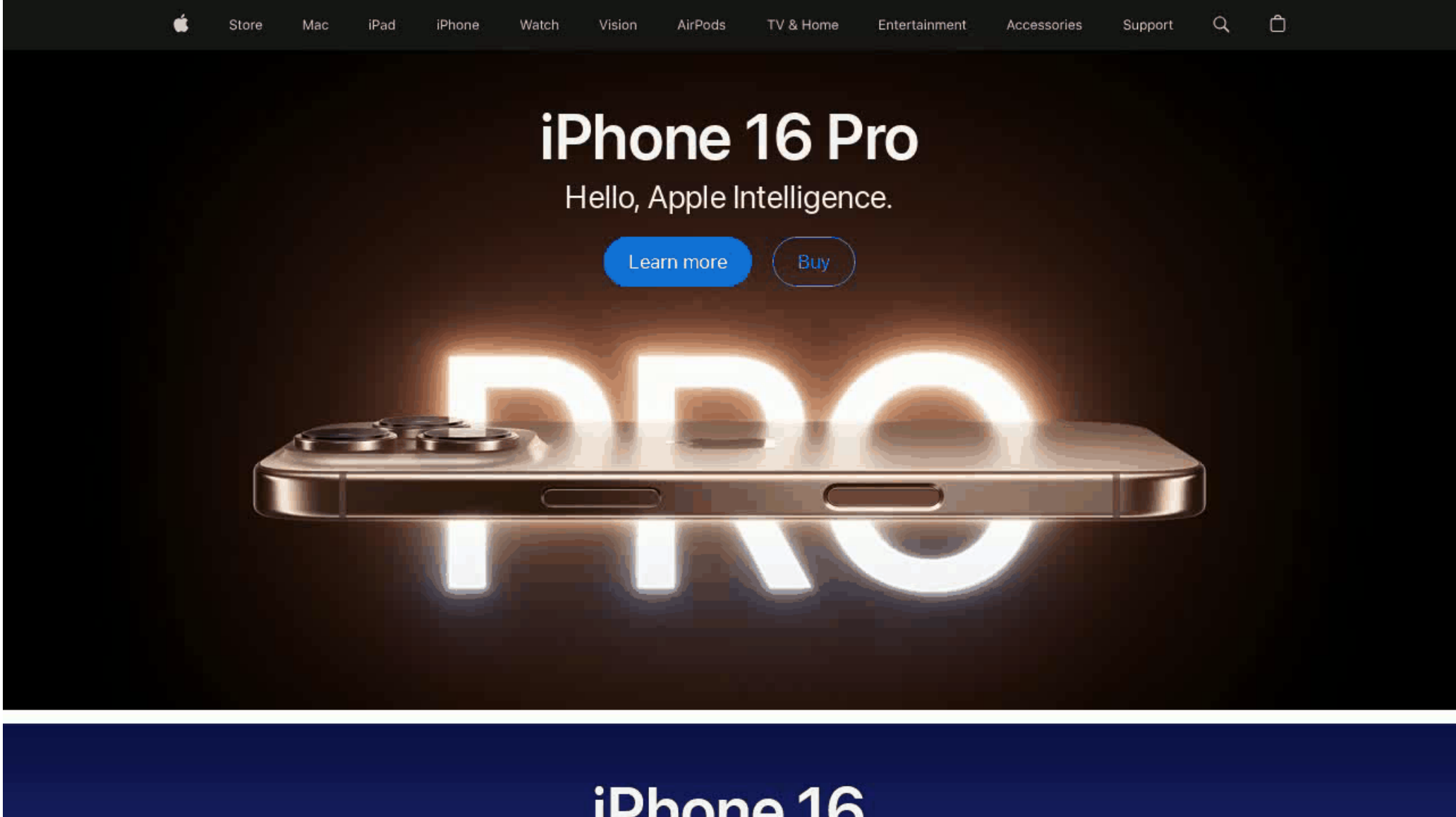
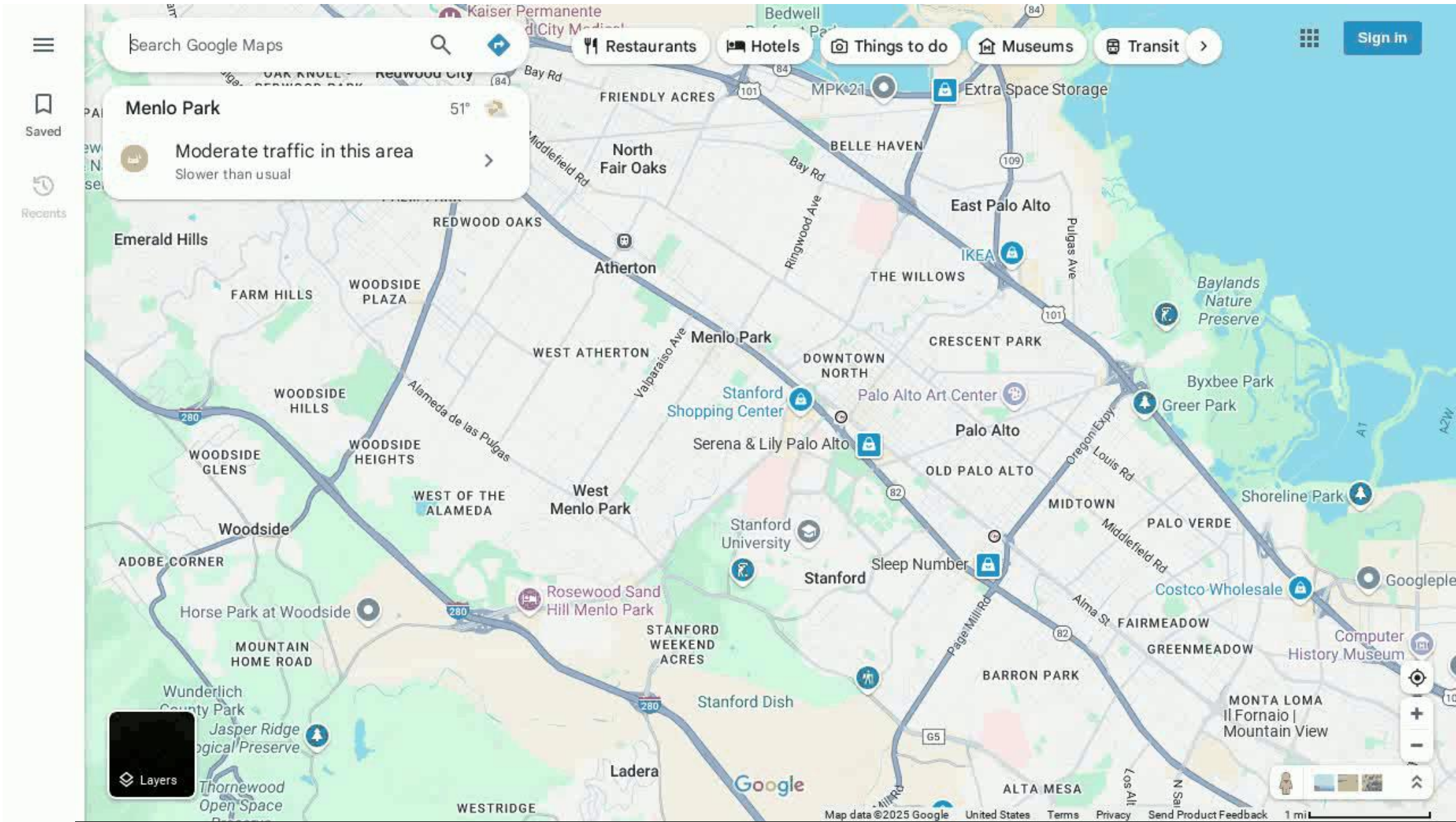
NNetNav: Discovering Goals on Web UI

An Exploration Policy interacts with a website to generate trajectories



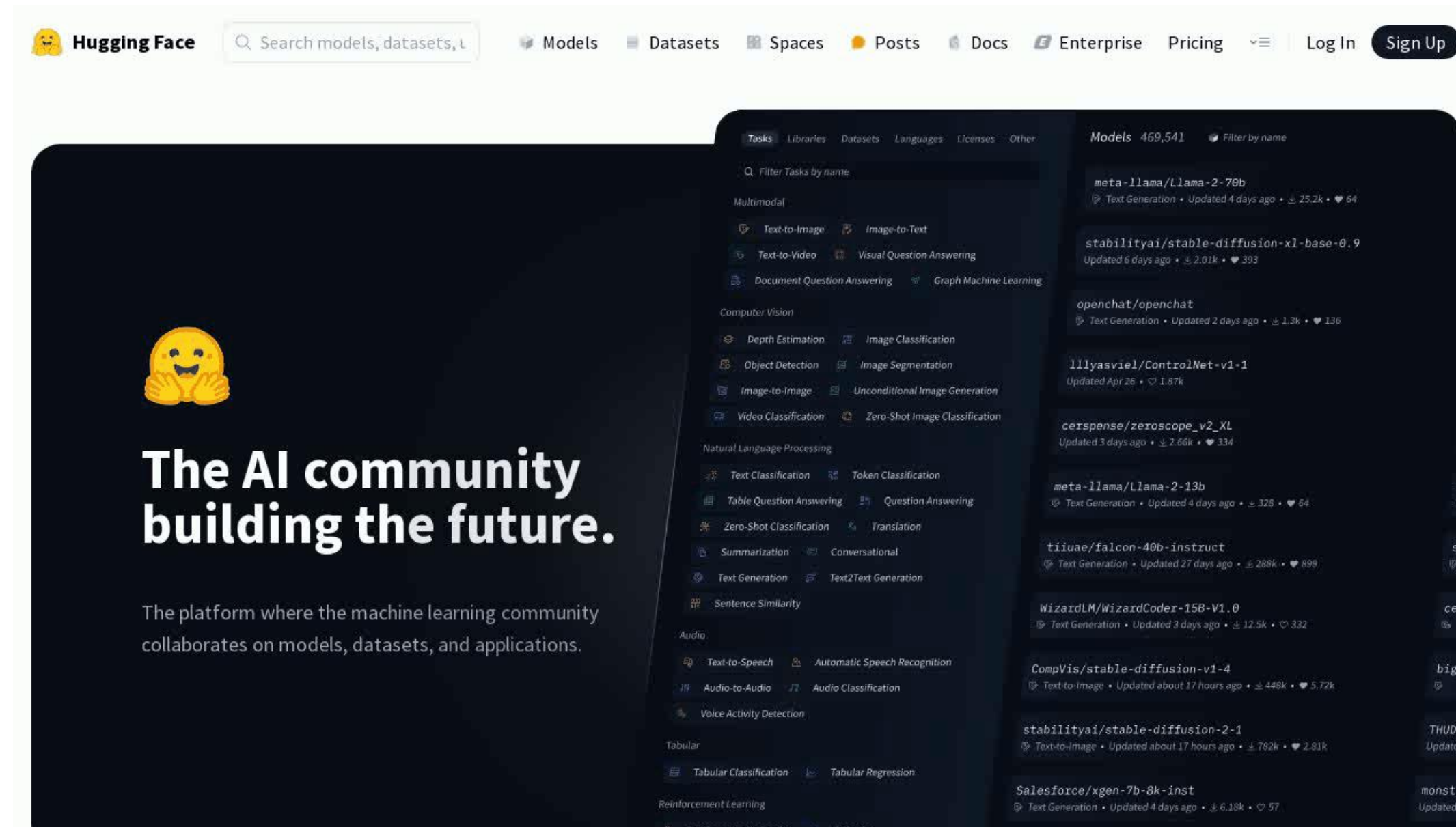
NEWS

LIVE New Orleans attacker named as US citizen as FBI says IS flag, possible explosives found



NNetNav: Discovering Goals on Web UI

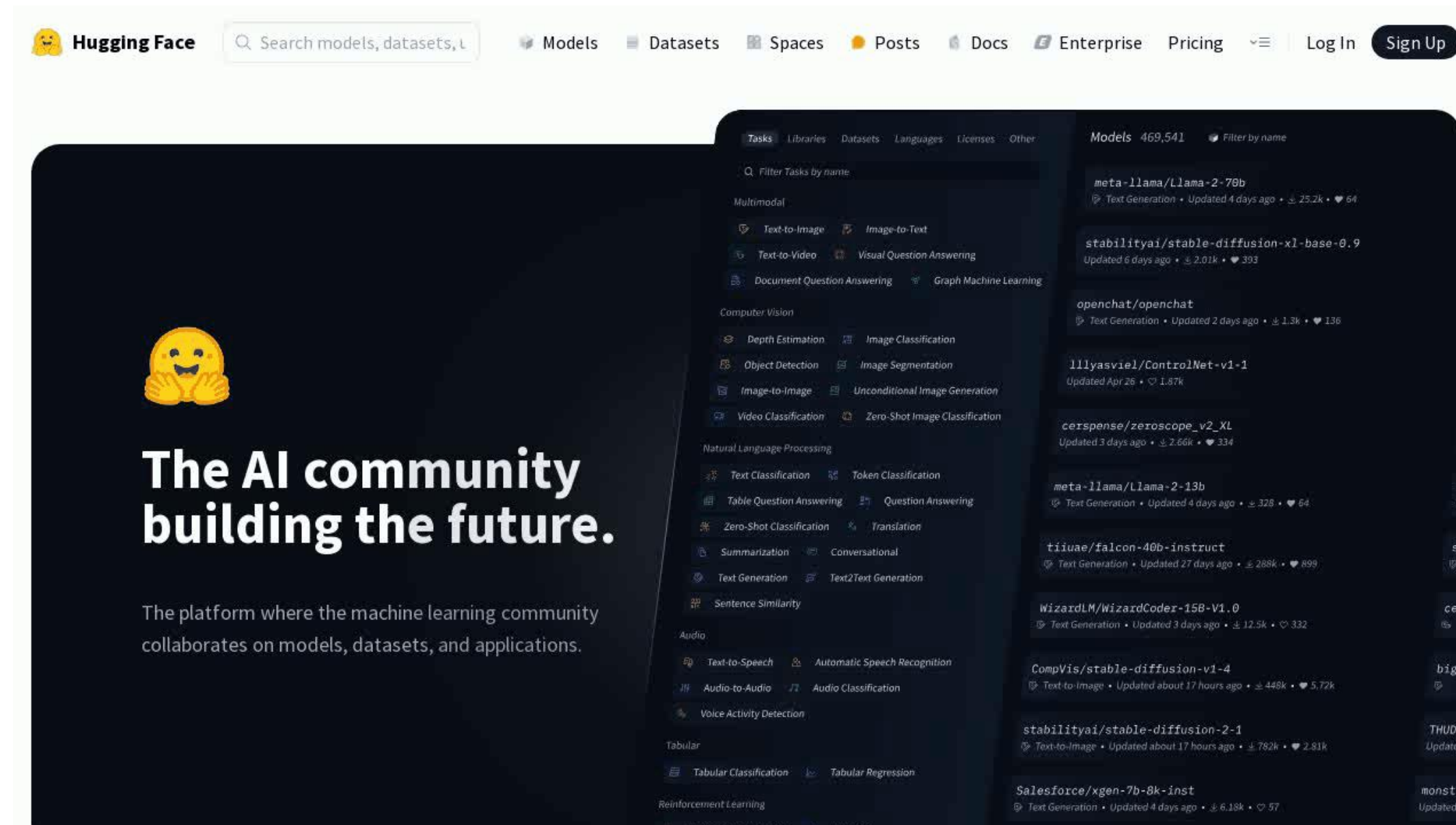
We retroactively relabel interactions, *using natural language to guide and shape exploration*



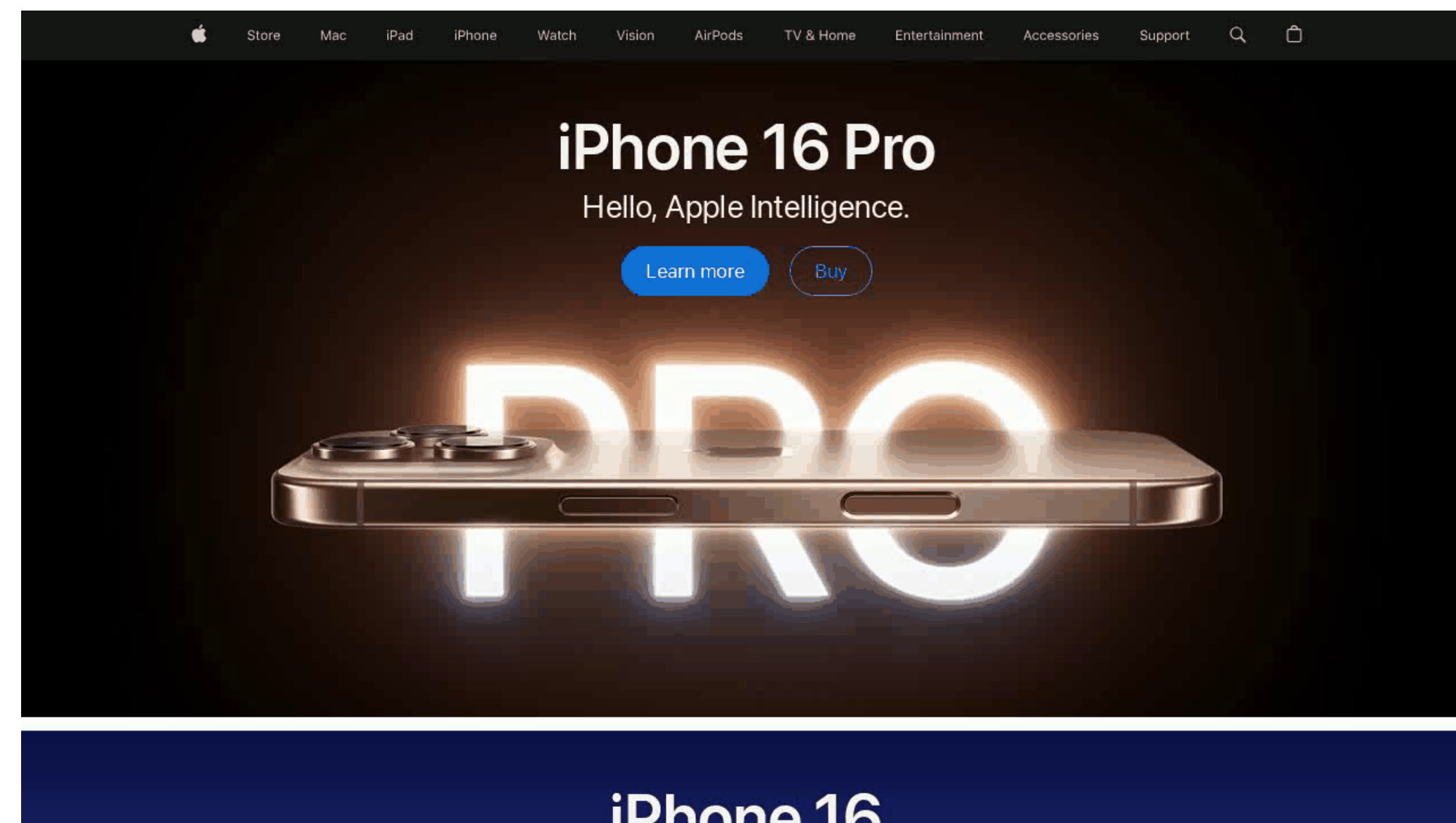
- Find a BERT model for natural language processing tasks
- Find the occurrences of the term "tesla" in a dataset related to "natural language processing".

NNetNav: Discovering Goals on Web UI

We retroactively relabel interactions, *using natural language to guide and shape exploration*



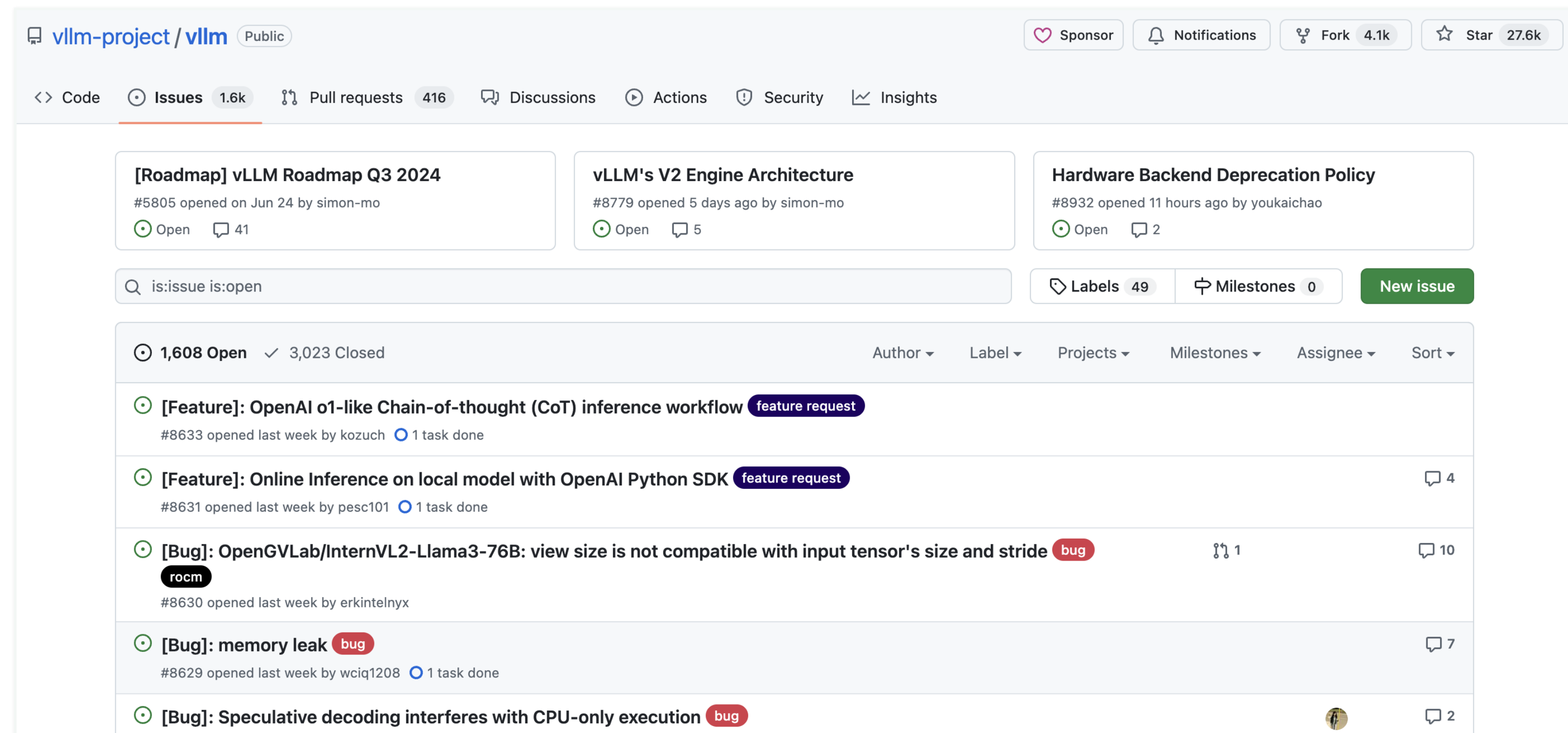
- Find a BERT model for natural language processing tasks
- Find the occurrences of the term "tesla" in a dataset related to "natural language processing".



- Research the Apple Watch Ultra 2.
- Find out the battery life of the Apple Watch Ultra 2.
- Find the battery life of the Apple Watch Ultra 2 and learn about Apple's shipping options.

Using natural language to guide and shape exploration

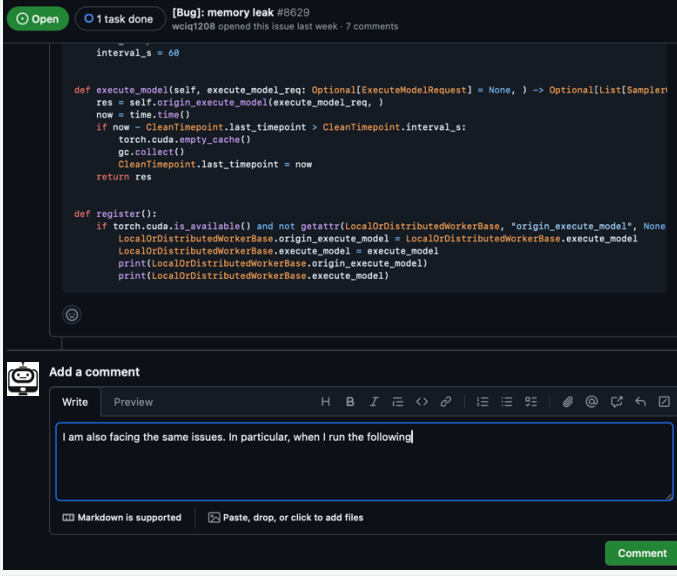
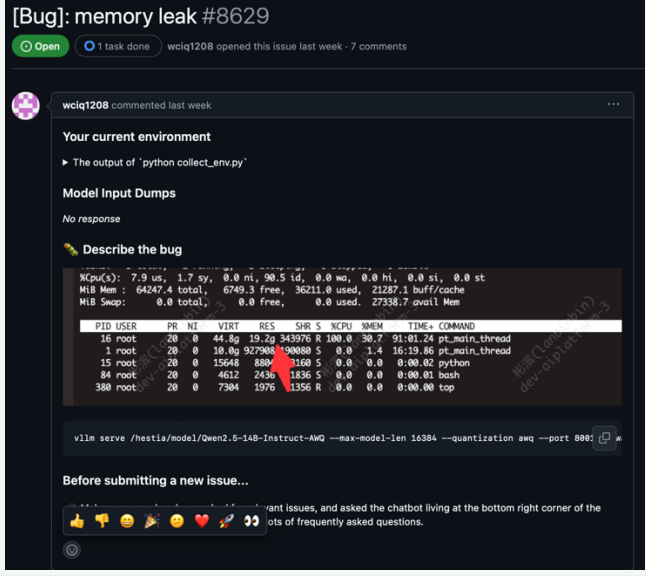
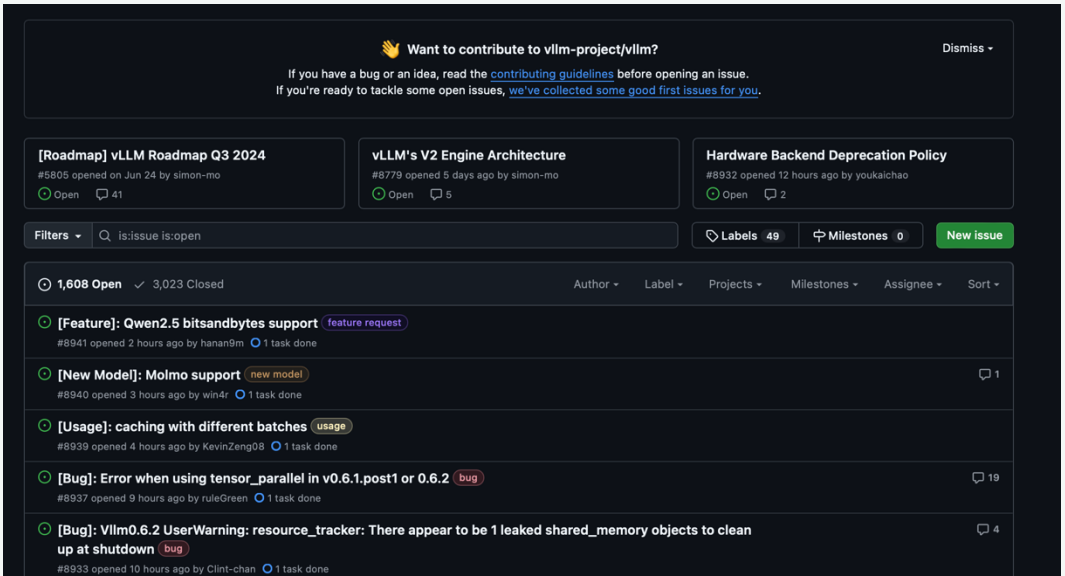
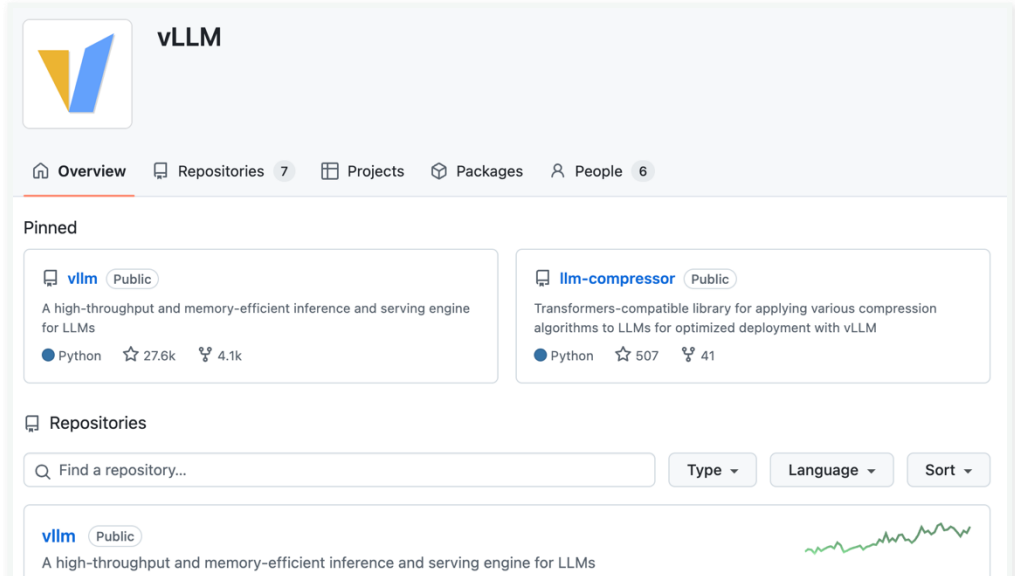
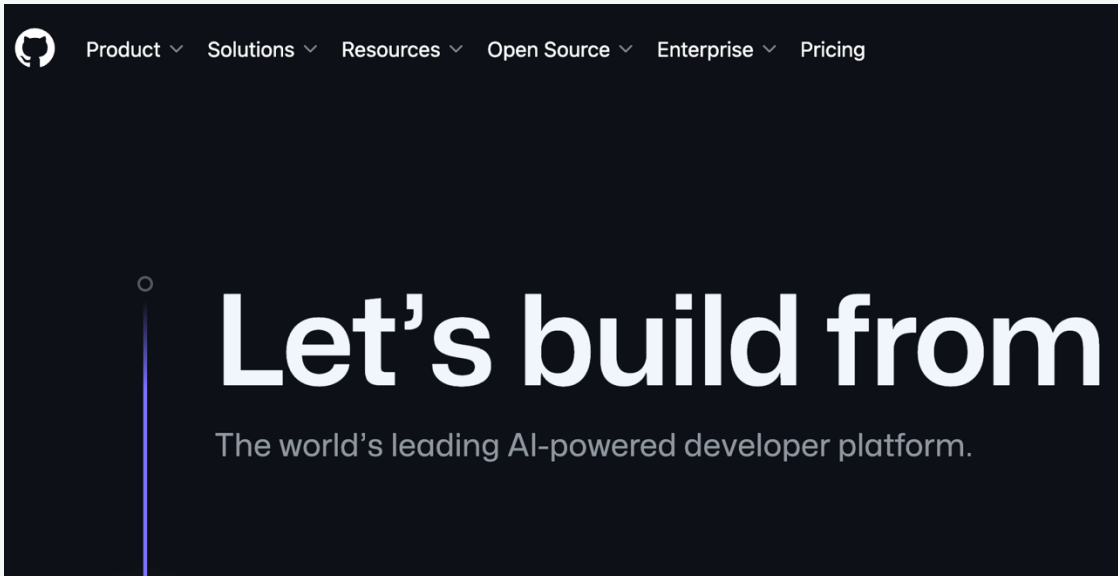
Meaningful interactions on real-websites are long and often hierarchical



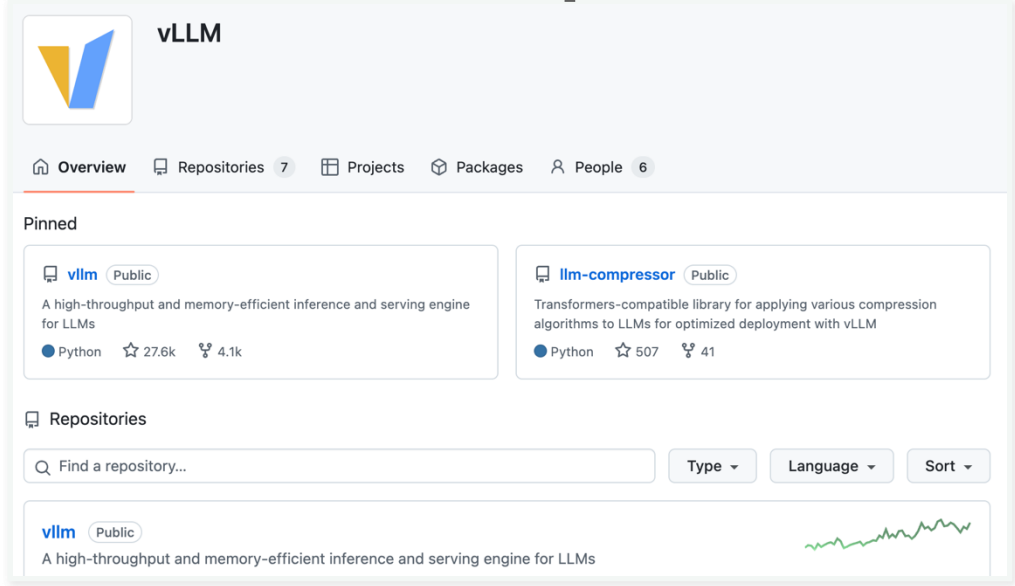
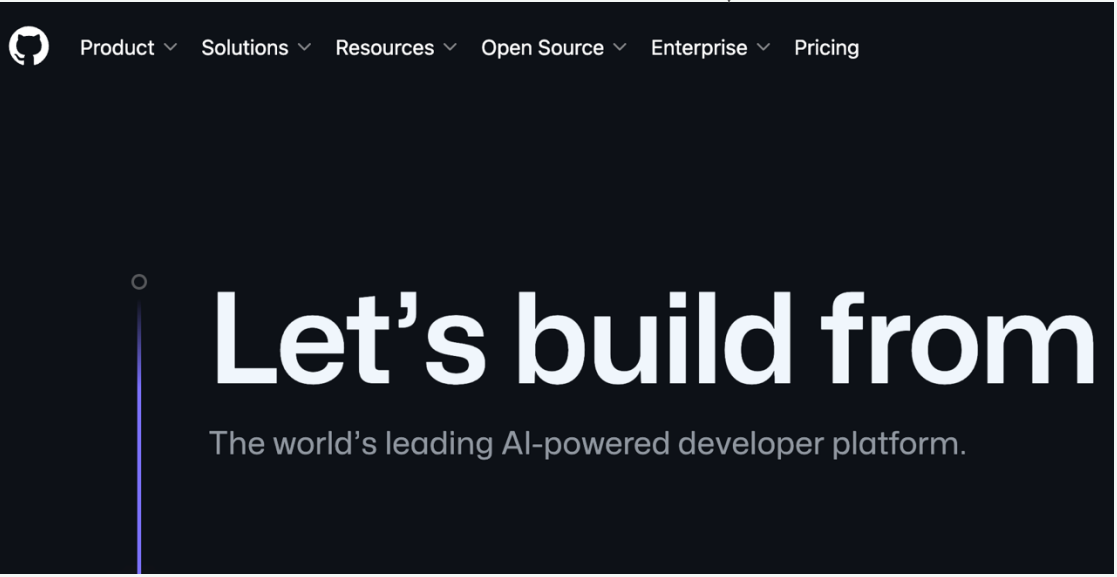
Navigate to the issue titled memory leak for the vllm project, like the comment by user wciq1208 and then respond to the post saying “I am also”

How can we search efficiently over the space of possibly exponential meaningful interactions?
Use Linguistic priors

Navigate to the issue titled memory leak for the vllm project, like the comment by user wciq1208 and then respond to the post saying “I am also”



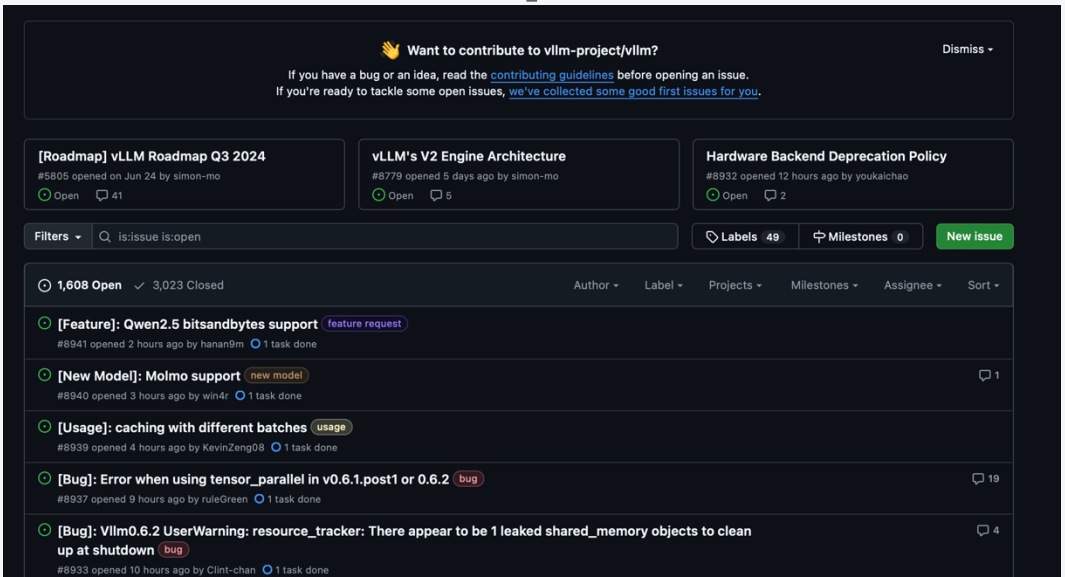
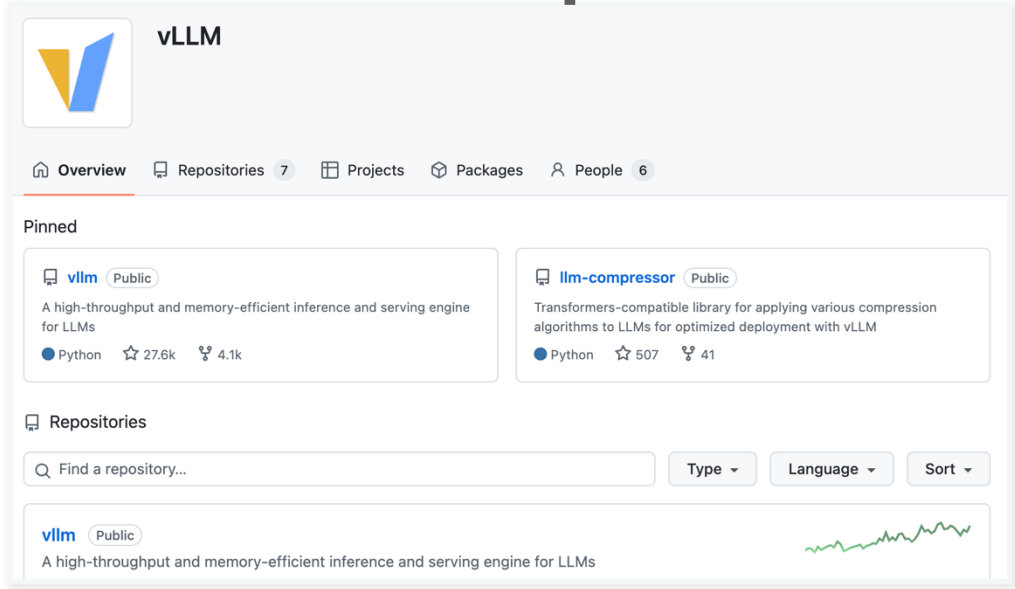
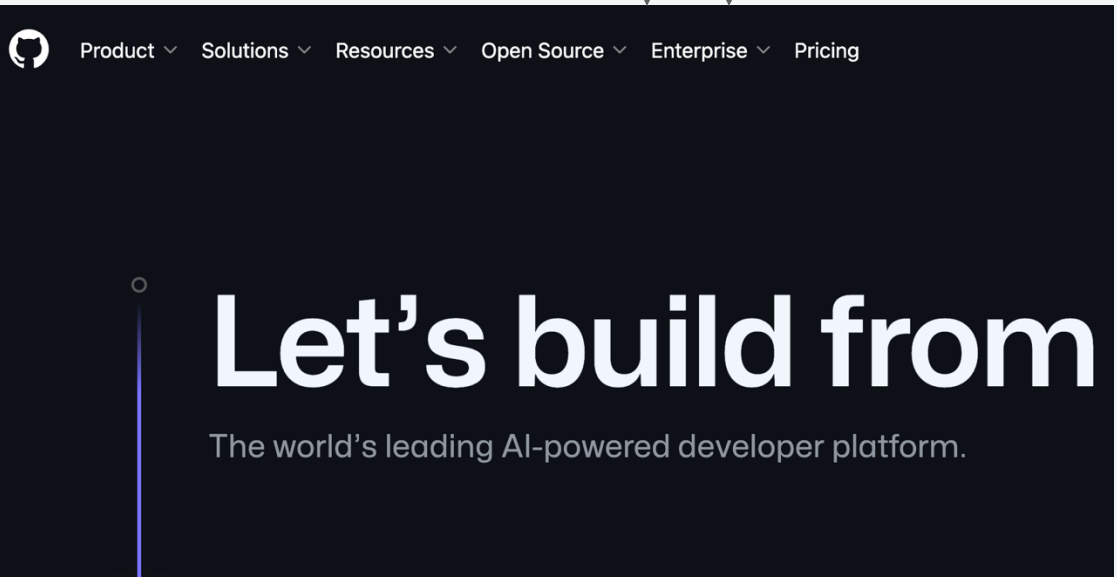
Navigate to the vllm repository on github.com



Navigate to the issue titled memory leak for the vllm project, like the comment by user wciq1208 and then respond to the post saying “I am also”

Open the issues page for the vllm repository on github.com

Navigate to the vllm repository on github.com

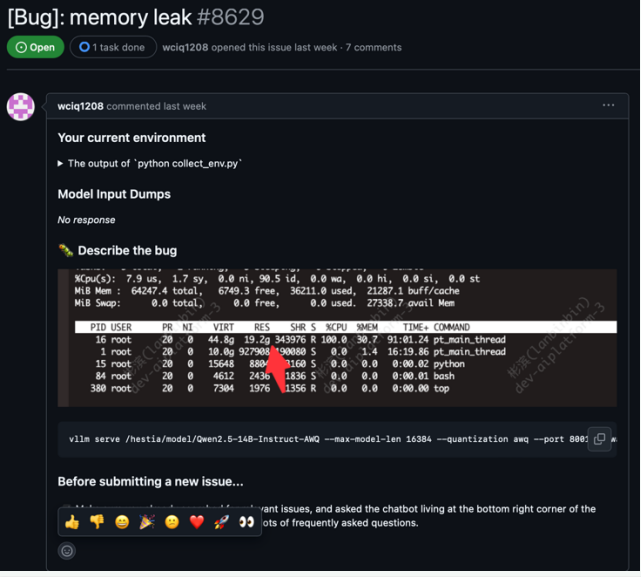
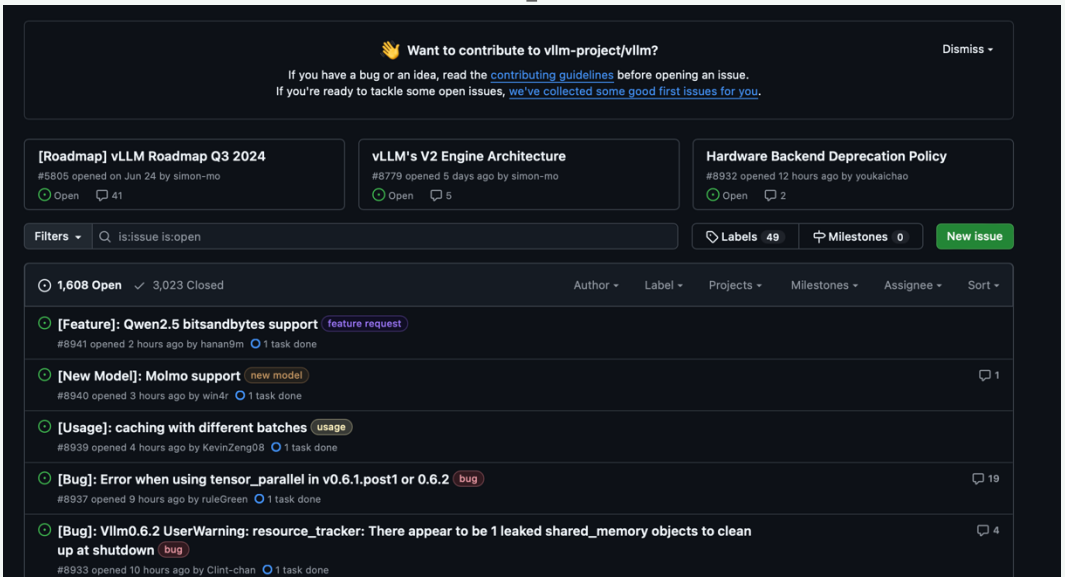
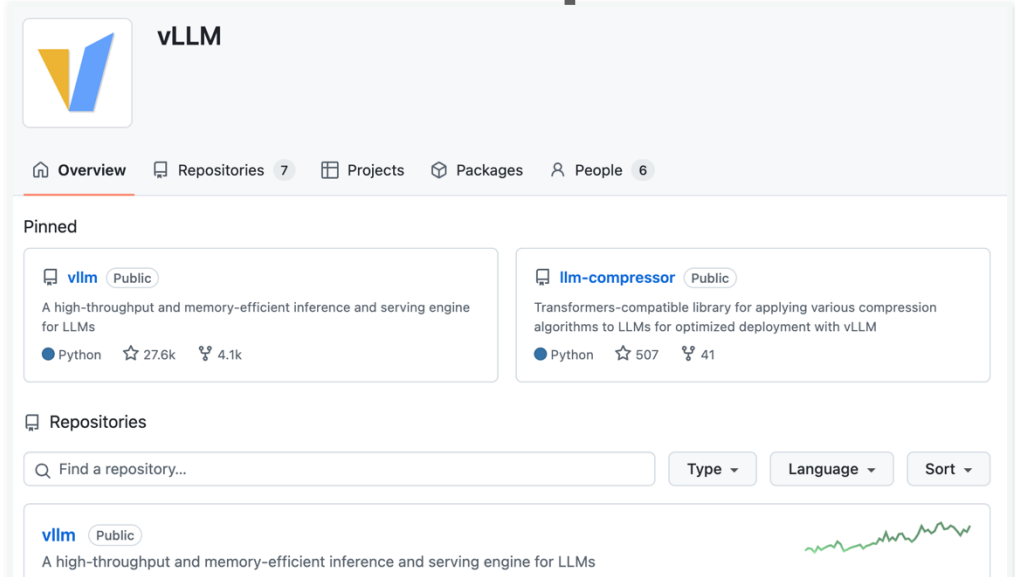
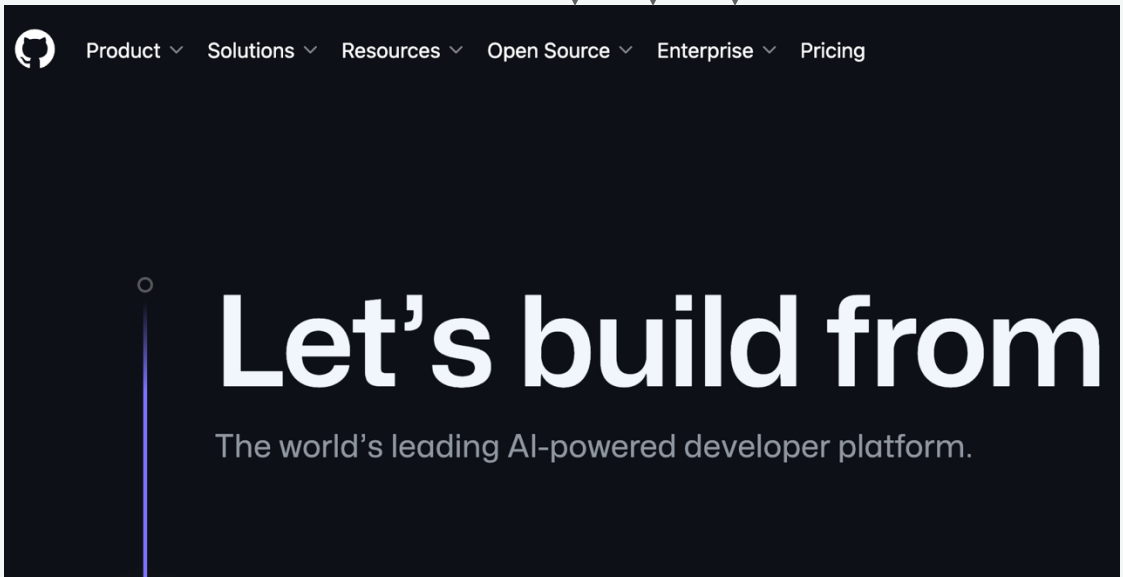


Navigate to the issue titled memory leak for the vllm project, like the comment by user wciq1208 and then respond to the post saying “I am also
....”

Navigate to issue titled memory leak for the vllm repository on github.com

Open the issues page for the vllm repository on github.com

Navigate to the vllm repository on github.com



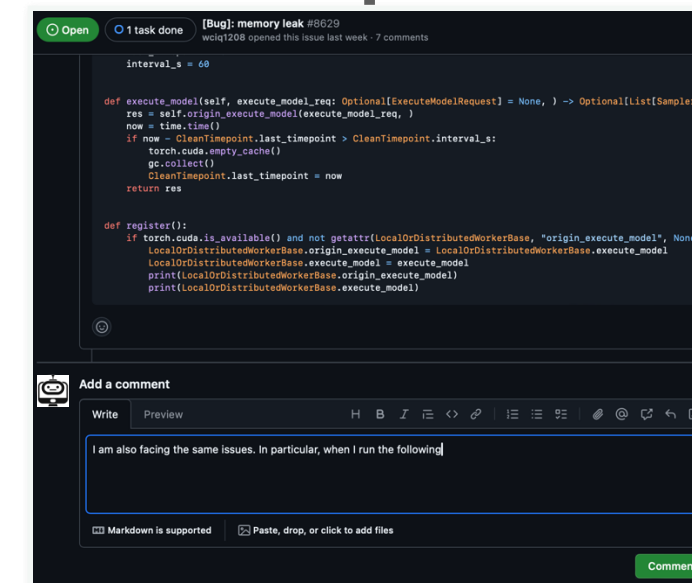
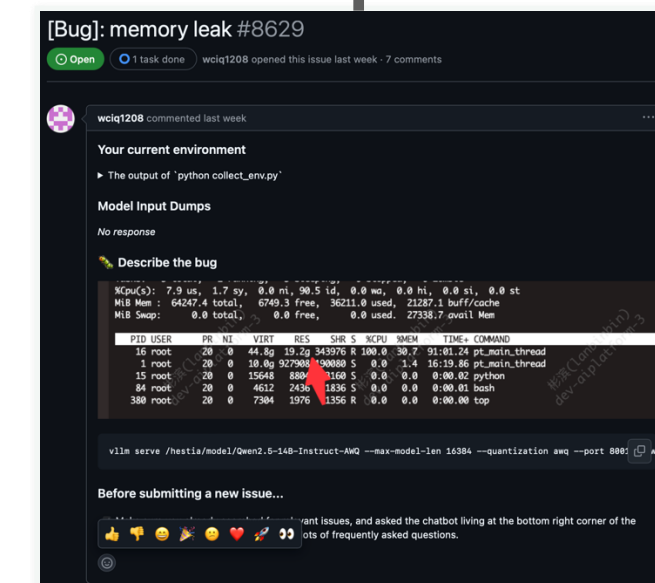
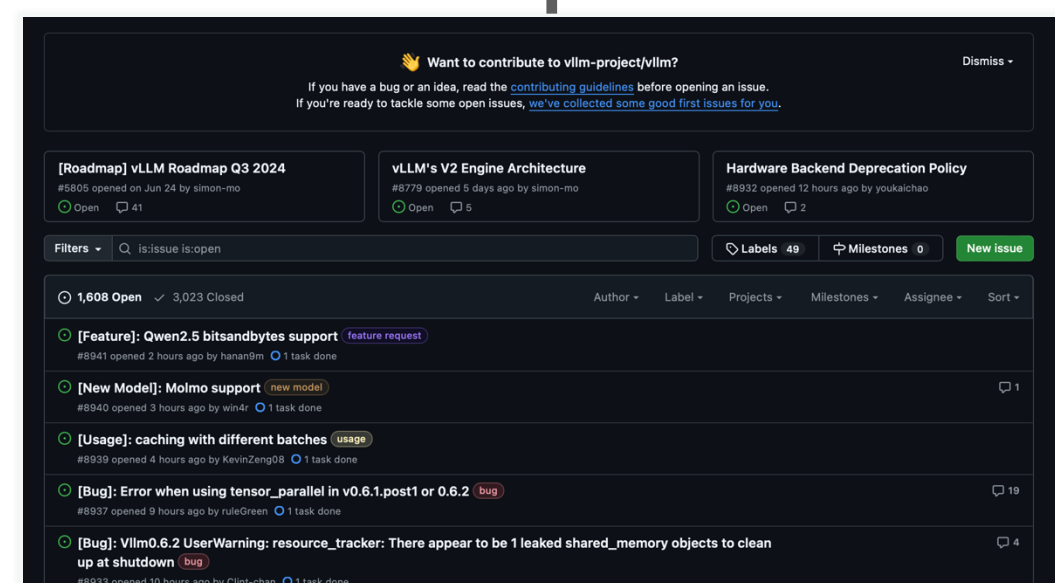
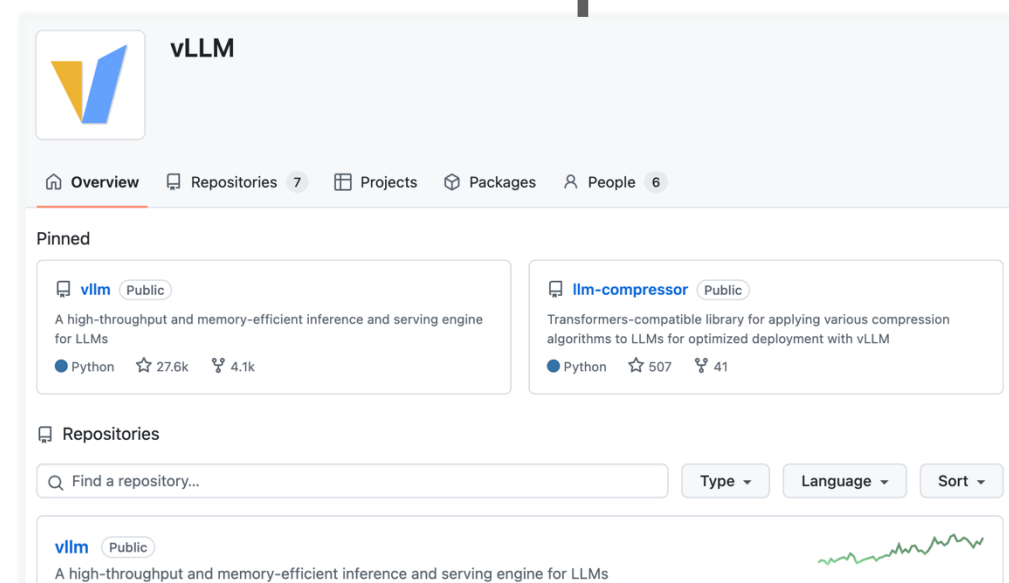
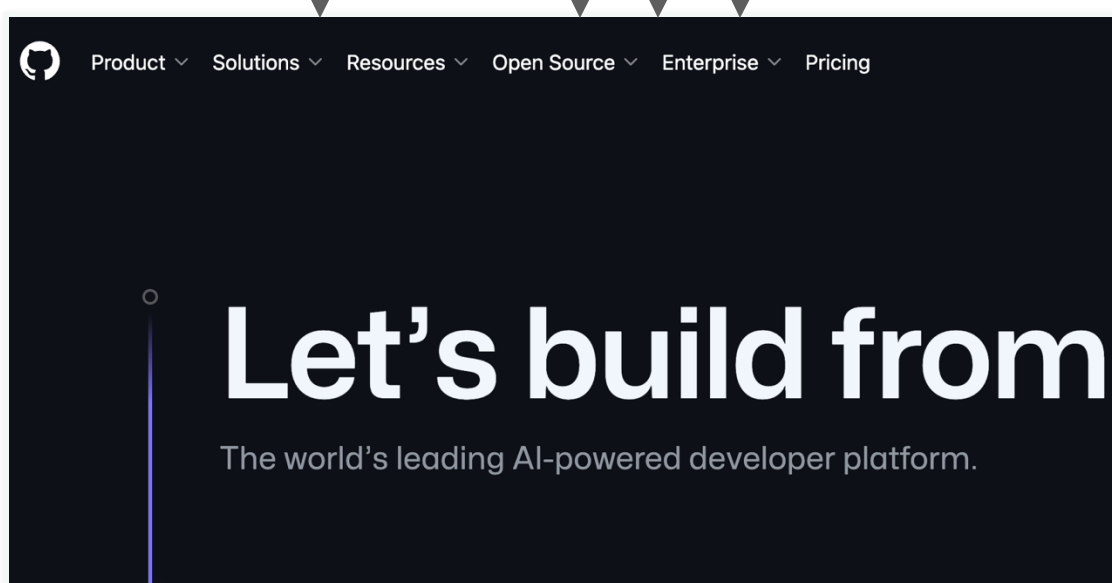
Navigate to the issue titled memory leak for the vllm project, like the comment by user wciq1208 and then respond to the post saying “I am also”

Navigate to the issue titled memory leak for the vllm project, like the comment by user wciq1208 and then respond to the post saying “I am also”

Navigate to issue titled memory leak for the vllm repository on github.com

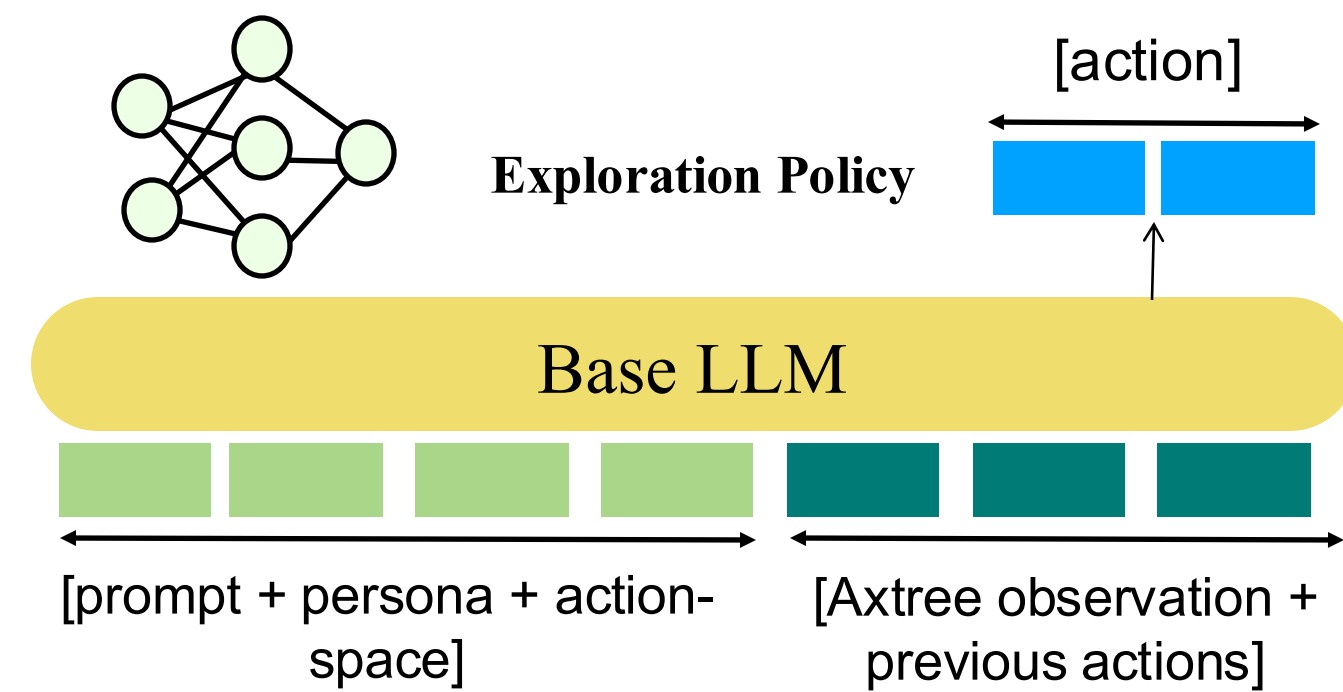
Open the issues page for the vllm repository on github.com

Navigate to the vllm repository on github.com



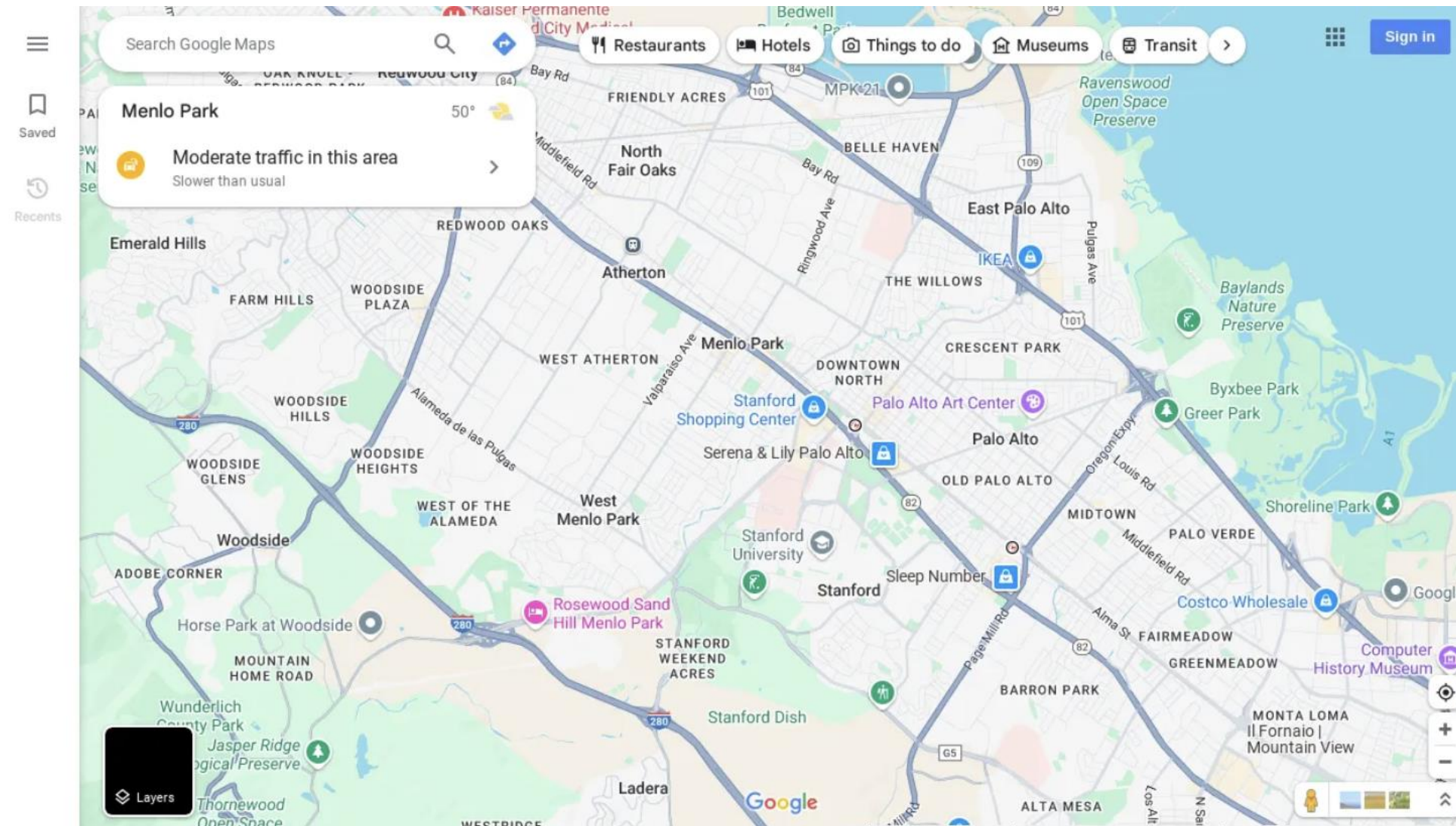
Complex trajectories can be segmented such that prefixes correspond to sub-tasks. We can use this as a pruning function for structuring exploration!

NNetNav: Main Components

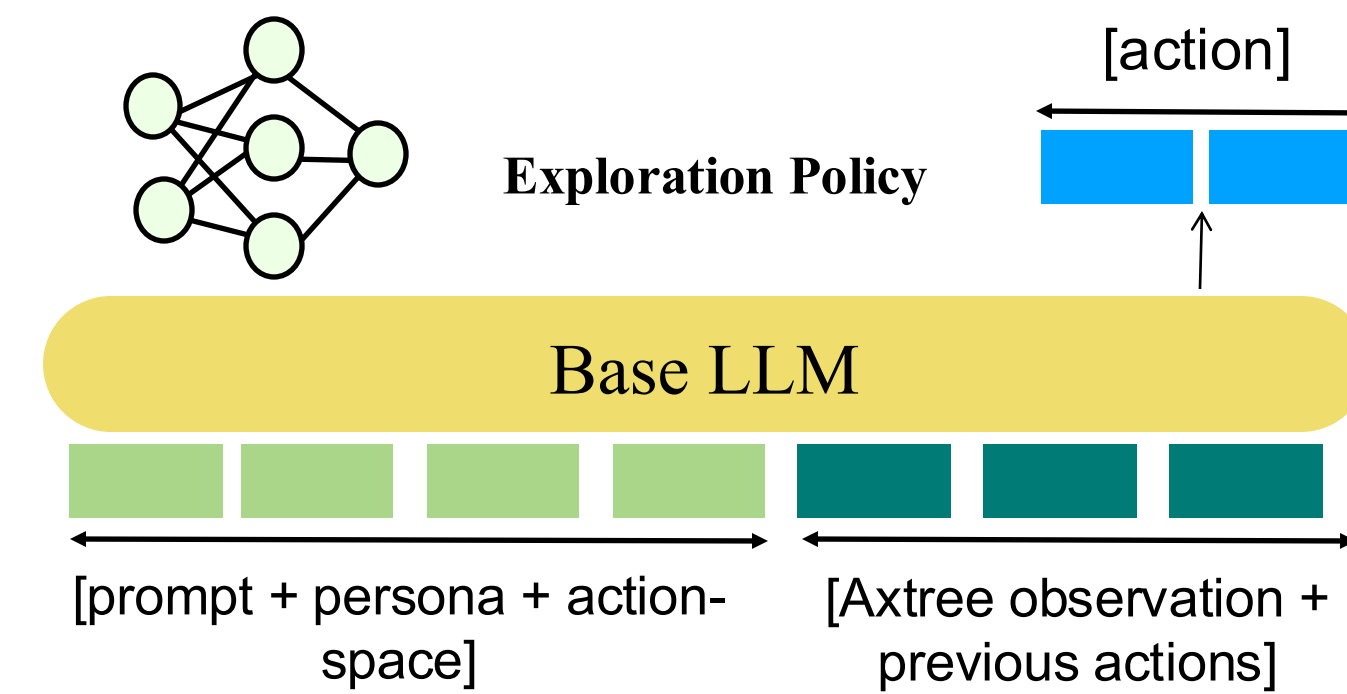


Base LLM: LLama-3.1-70B-Instruct

NNetNav: Main Components

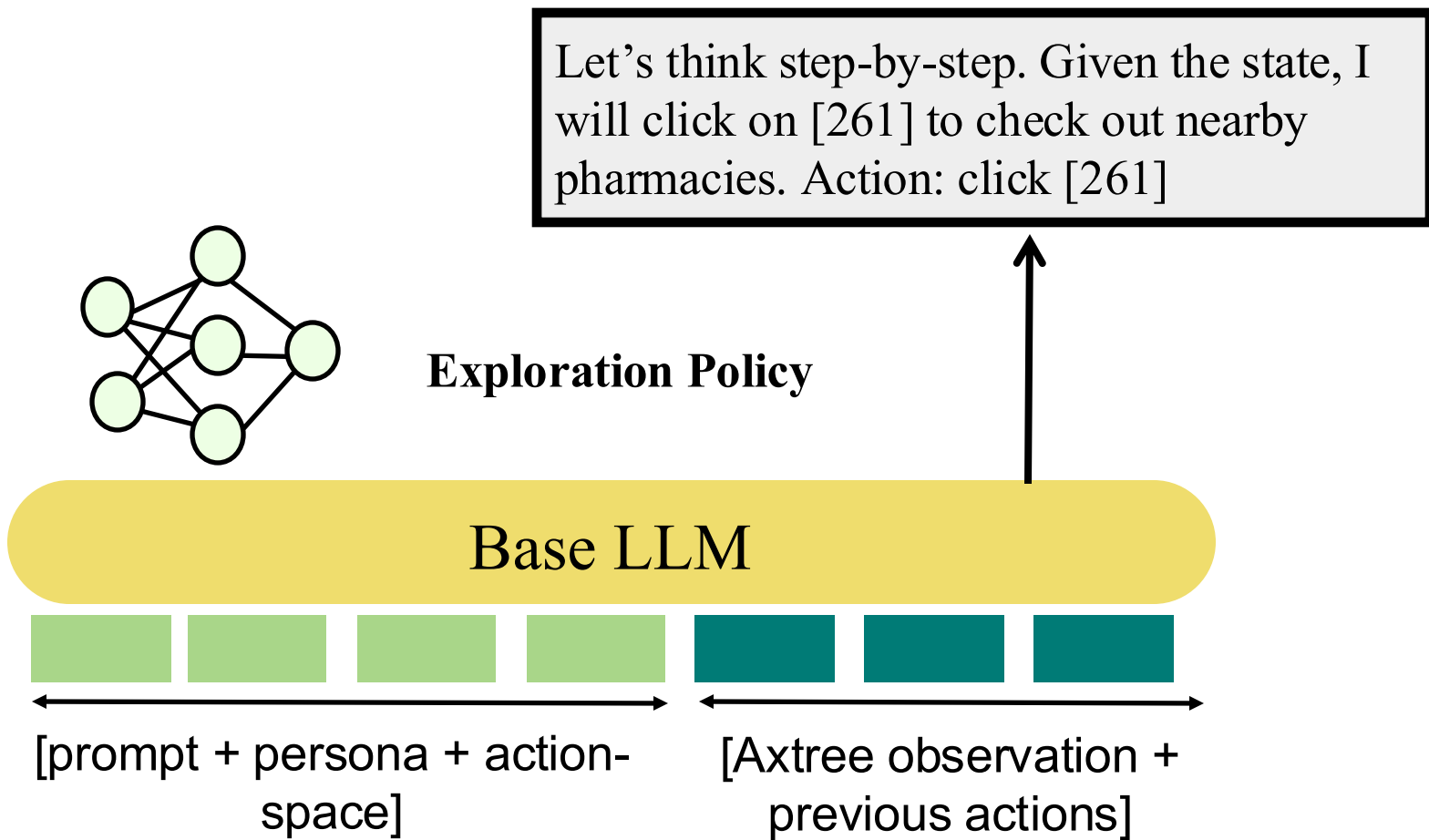


Actual browser observation



NNetNav: Main Components

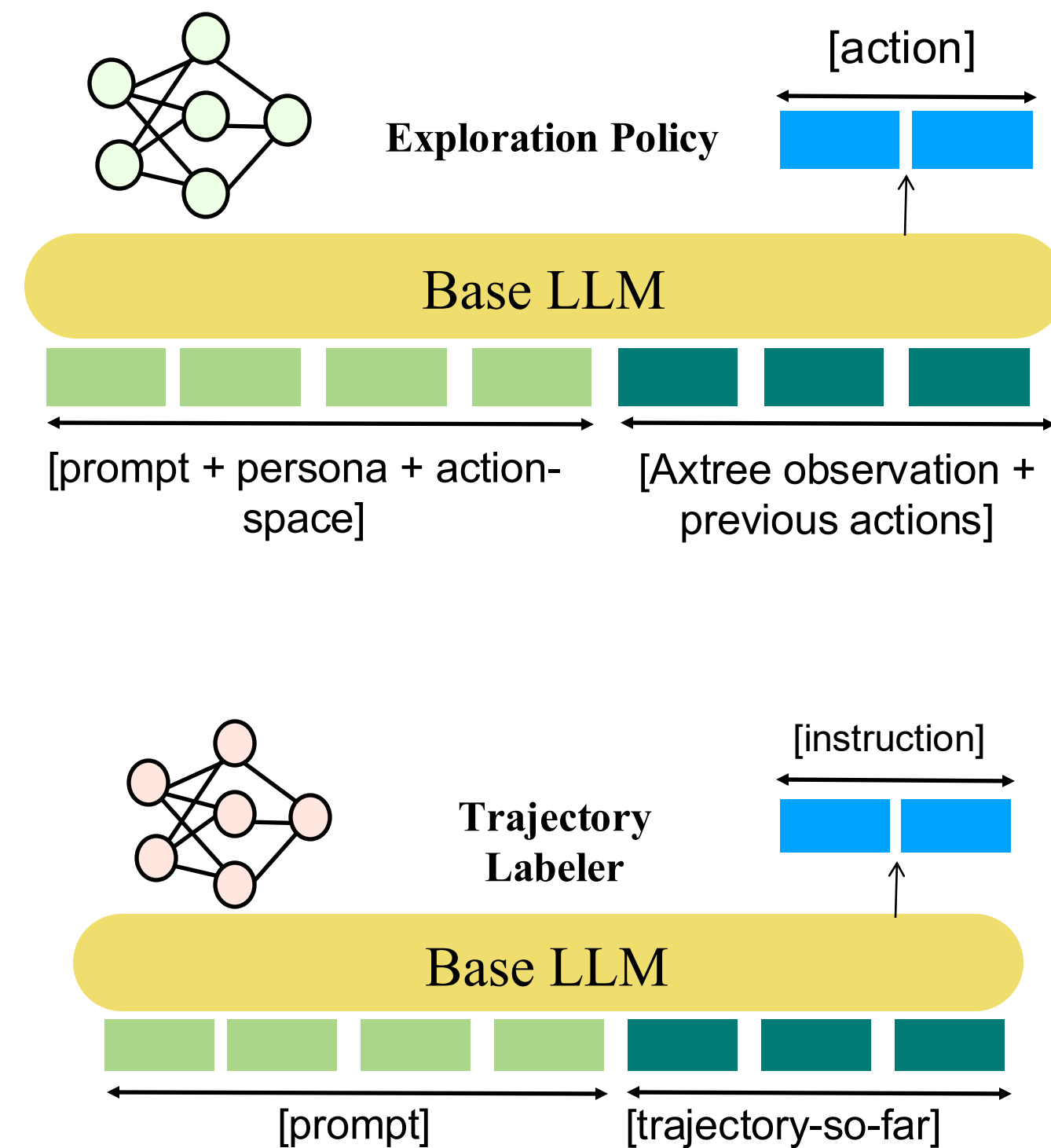
```
RootWebArea 'Google Maps', focused, url='https://www.google.com/maps'
  [120] application 'Map · Use arrow keys to pan the map. · Get details about a place by pressing its corresponding number key.', visible
    [123] Canvas ", visible
  [134] search ", visible
    [138] LabelText ", clickable, visible
      StaticText 'Search Google Maps'
    [139] combobox 'Search Google Maps', clickable, visible, focused, hasPopup='grid', expanded=False
    [141] button 'Search', visible
    [175] button 'Directions', visible
  [194] region 'Menlo Park', visible
    [197] heading 'Menlo Park', visible
    [198] group 'Menlo Park weather', visible
      StaticText '50°'
      image 'Partly cloudy', url='https://ssl.gstatic.com/onebox/weather/128/partly_cloudy.png'
    [199] region 'Traffic', visible
      [200] button 'Moderate traffic in this area Slower than usual', visible
        StaticText 'Moderate traffic in this area'
        StaticText 'Slower than usual'
  [220] region 'Available search options for this area', visible, roledescription='carousel'
    [226] button 'Restaurants', visible
      StaticText 'Restaurants'
    [233] button 'Hotels', visible
      StaticText 'Hotels'
    [240] button 'Things to do', visible
      StaticText 'Things to do'
    [247] button 'Museums', visible
      StaticText 'Museums'
    [254] button 'Transit', visible
      StaticText 'Transit'
    [261] button 'Pharmacies'
      StaticText 'Pharmacies'
    [268] button 'ATMs'
      StaticText 'ATMs'
    button 'Next page'
  [296] list ", visible
    [297] listitem ", visible
      [298] button 'Menu', visible
    [301] listitem ", visible
      [302] button 'Saved', visible
        StaticText 'Saved'
    [308] listitem ", visible
      [309] button 'Recents', visible, disabled=True
        StaticText 'Recents'
  [347] button 'Google apps', clickable, visible, expanded=False
  [350] link 'Sign in', clickable, visible,
    url='https://accounts.google.com/ServiceLogin?hl=en&passive=true&continue=https://www.google.com/maps&service=local&ec=GAZAacQ'
  [386] button 'Show Your Location', visible, pressed='false'
  [400] button 'Zoom in', visible
  [412] button 'Zoom out', visible
  [421] button 'Show Street View coverage', visible
  [423] button 'Show imagery', visible
    [425] LabelText ", visible
  [469] LabelText ", visible
    StaticText 'Layers'
  [476] button 'Layers', visible, expanded=False
```



Base LLM: LLama-3.1-70B-Instruct

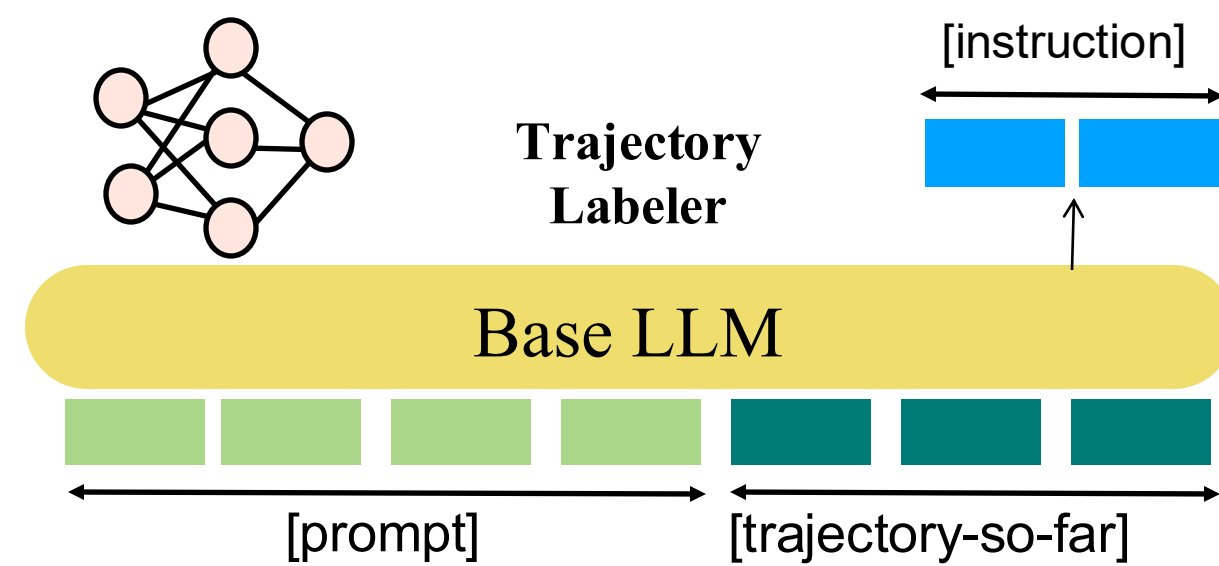
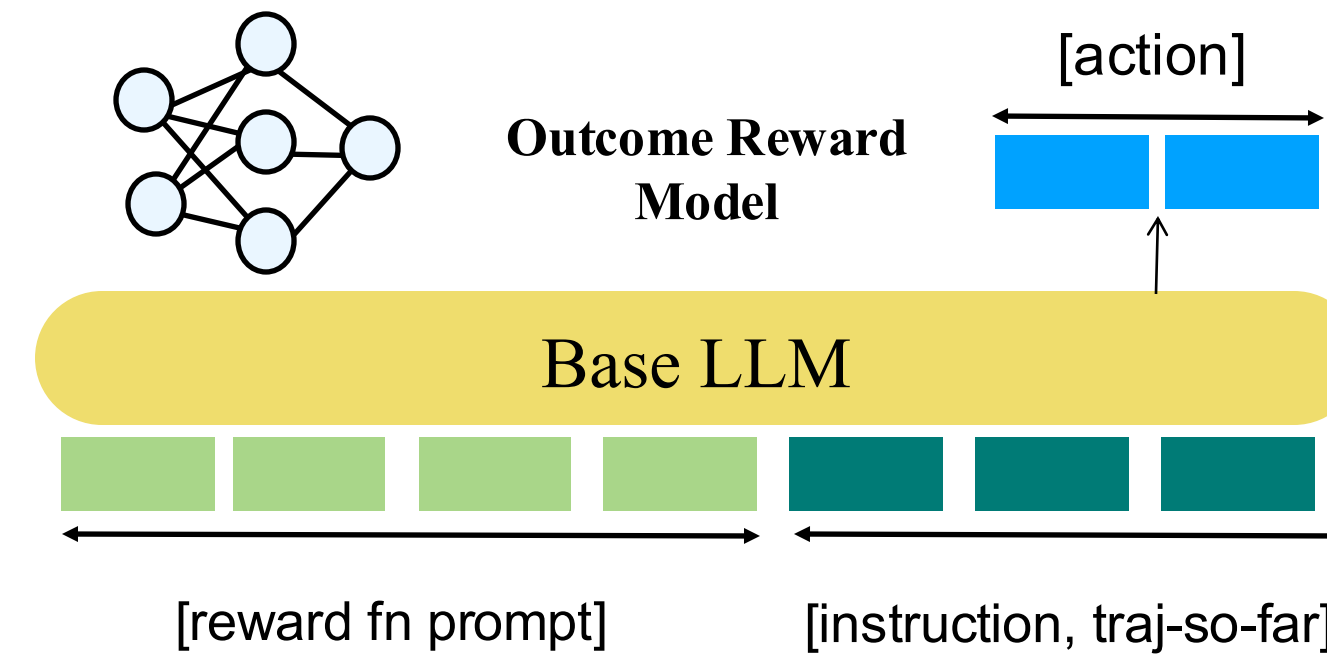
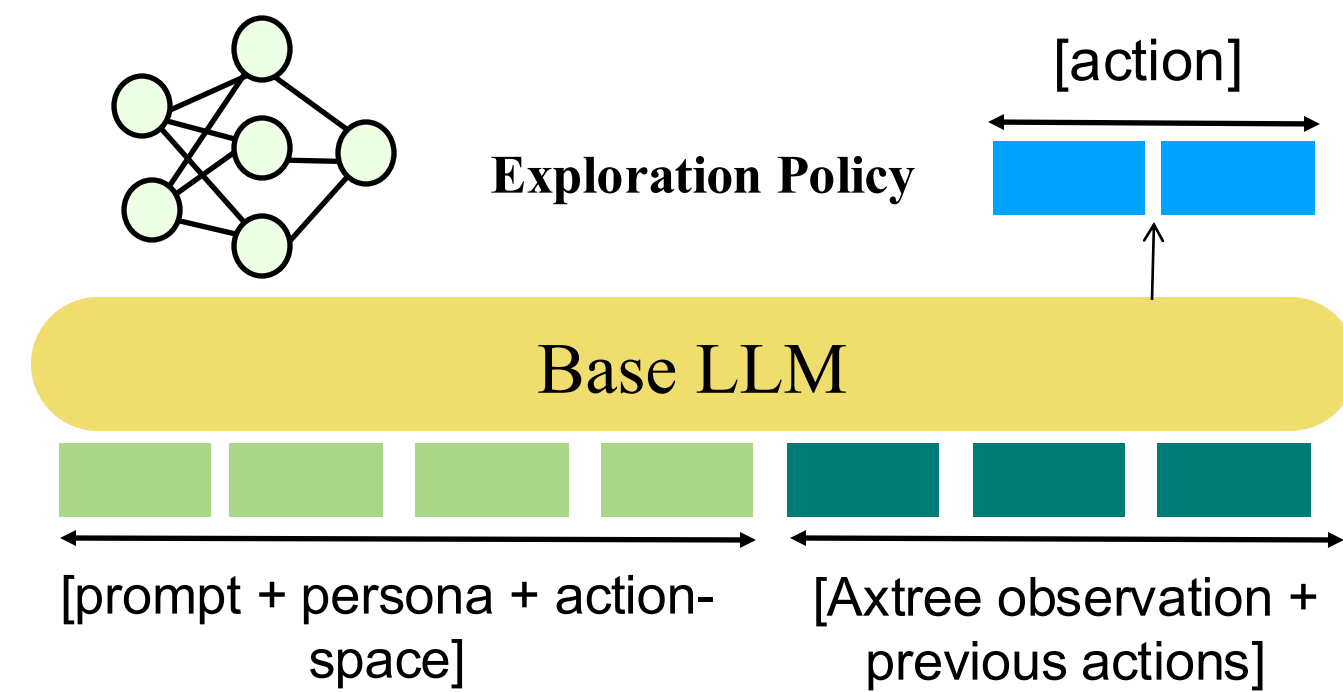
The Exploration Policy interacts the environment by generating actions from the action-space

NNetNav: Main Components



The Trajectory Labeler takes a summary of the trajectory-so-far and outputs a plausible goal for that trajectory.

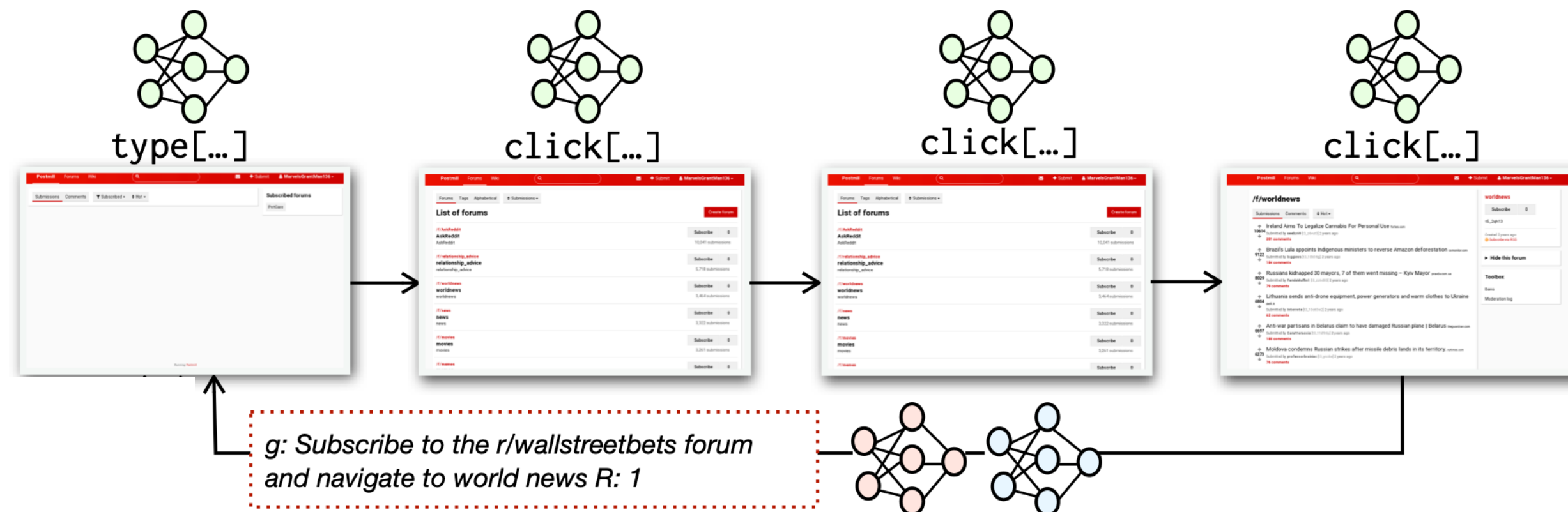
NNetNav: Main Components



The ORM scores the trajectory for the generated goal

NNetNav: Structuring Exploration for Goal Discovery

1. Interact with a website with a structured exploration policy + trajectory labeler



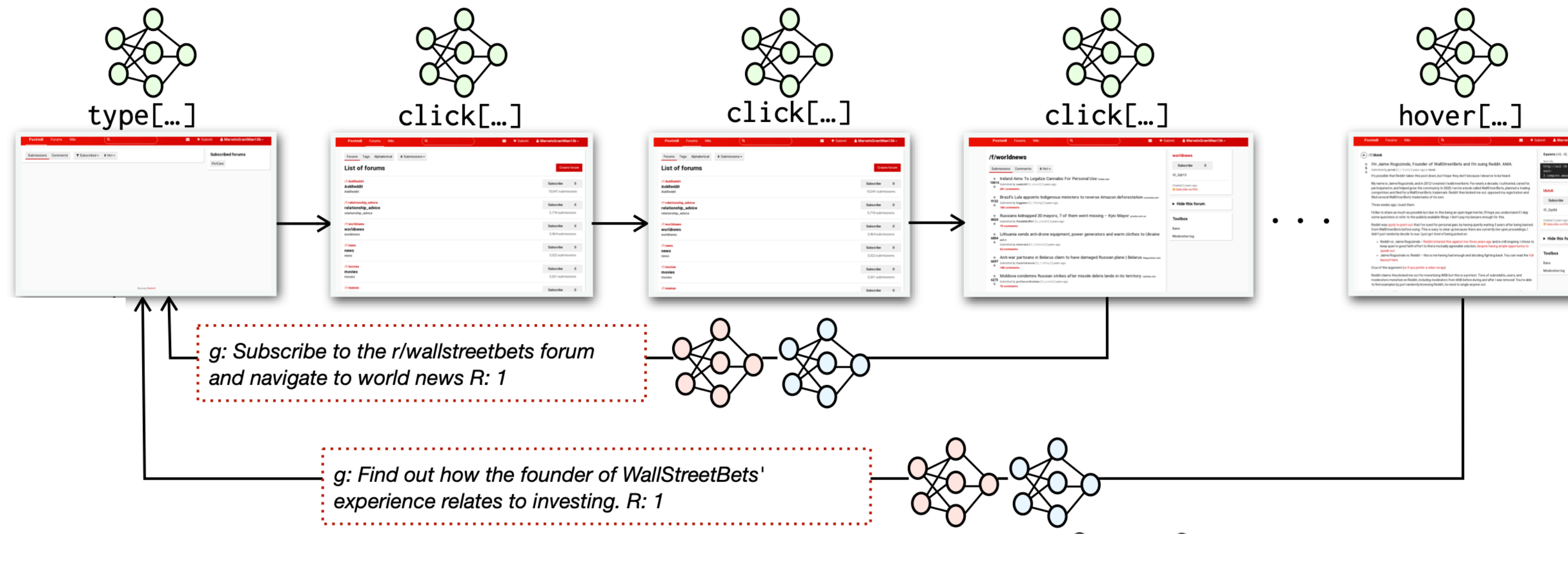
URL: <http://webarena-reddit.com>
Persona Type: Investor, Active on r/
investing and r/wallstreetbets

Pruning Function:

- ▶ Invoke Trajectory Labeler at fixed time-steps
- ▶ Assign reward to resulting (goal, trajectory) pair
- ▶ Continue if reward is 1
- ▶ Prune if reward is 0

NNetNav: Structuring Exploration for Goal Discovery

1. Interact with a website with a structured exploration policy + trajectory labeler

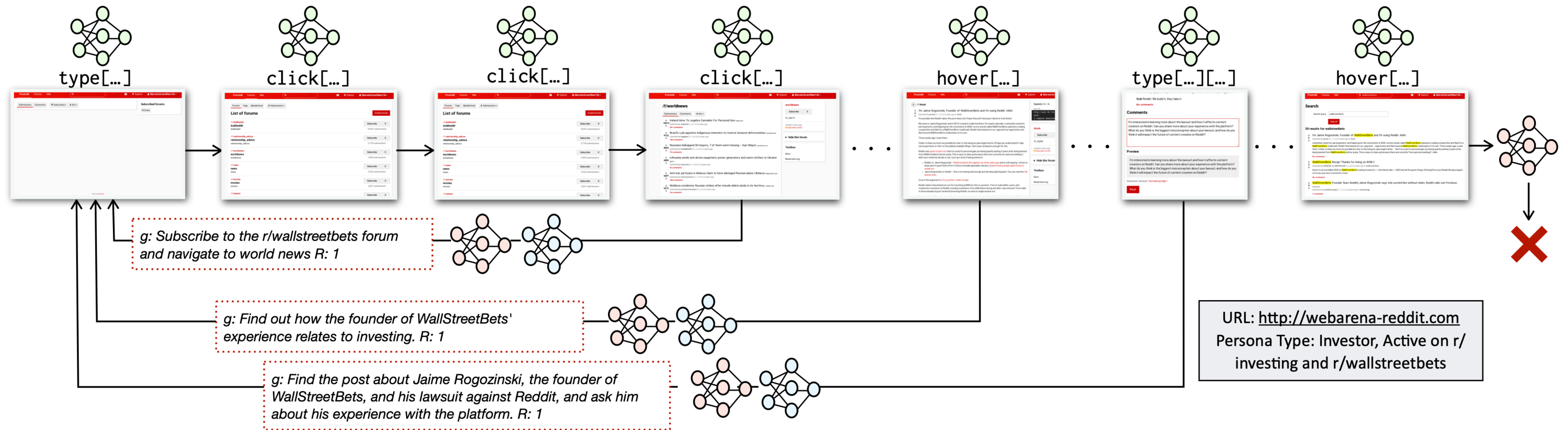


Pruning Function:

- ▶ Invoke Trajectory Labeler at fixed time-steps
- ▶ Assign reward to resulting (goal, trajectory) pair
- ▶ Continue if reward is 1
- ▶ Prune if reward is 0

NNetNav: Structuring Exploration for Goal Discovery

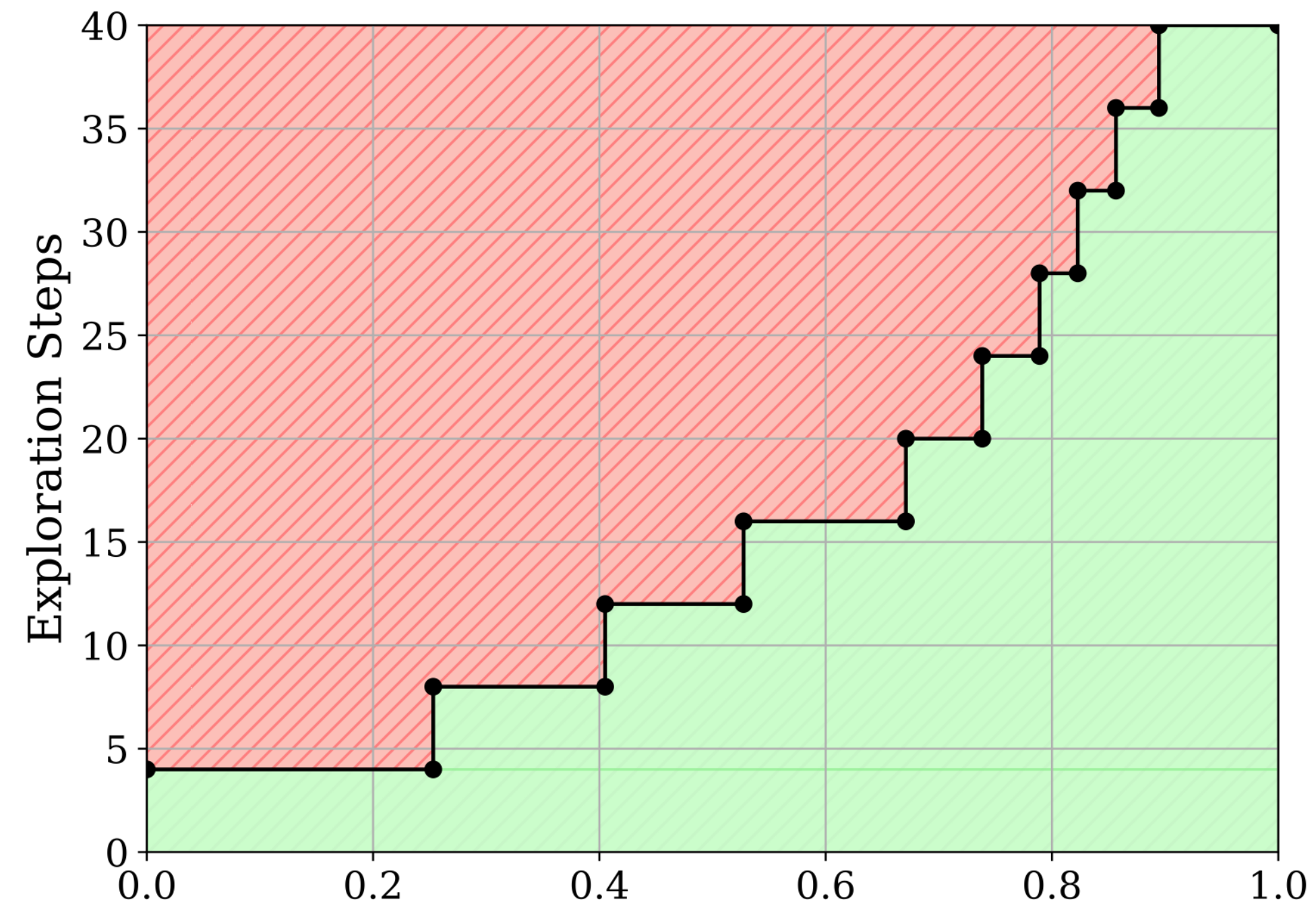
1. Interact with a website with a structured exploration policy + trajectory labeler



Pruning Function:

- ▶ Invoke Trajectory Labeler at fixed time-steps
- ▶ Assign reward to resulting (goal, trajectory) pair
- ▶ Continue if reward is 1
- ▶ Prune if reward is 0

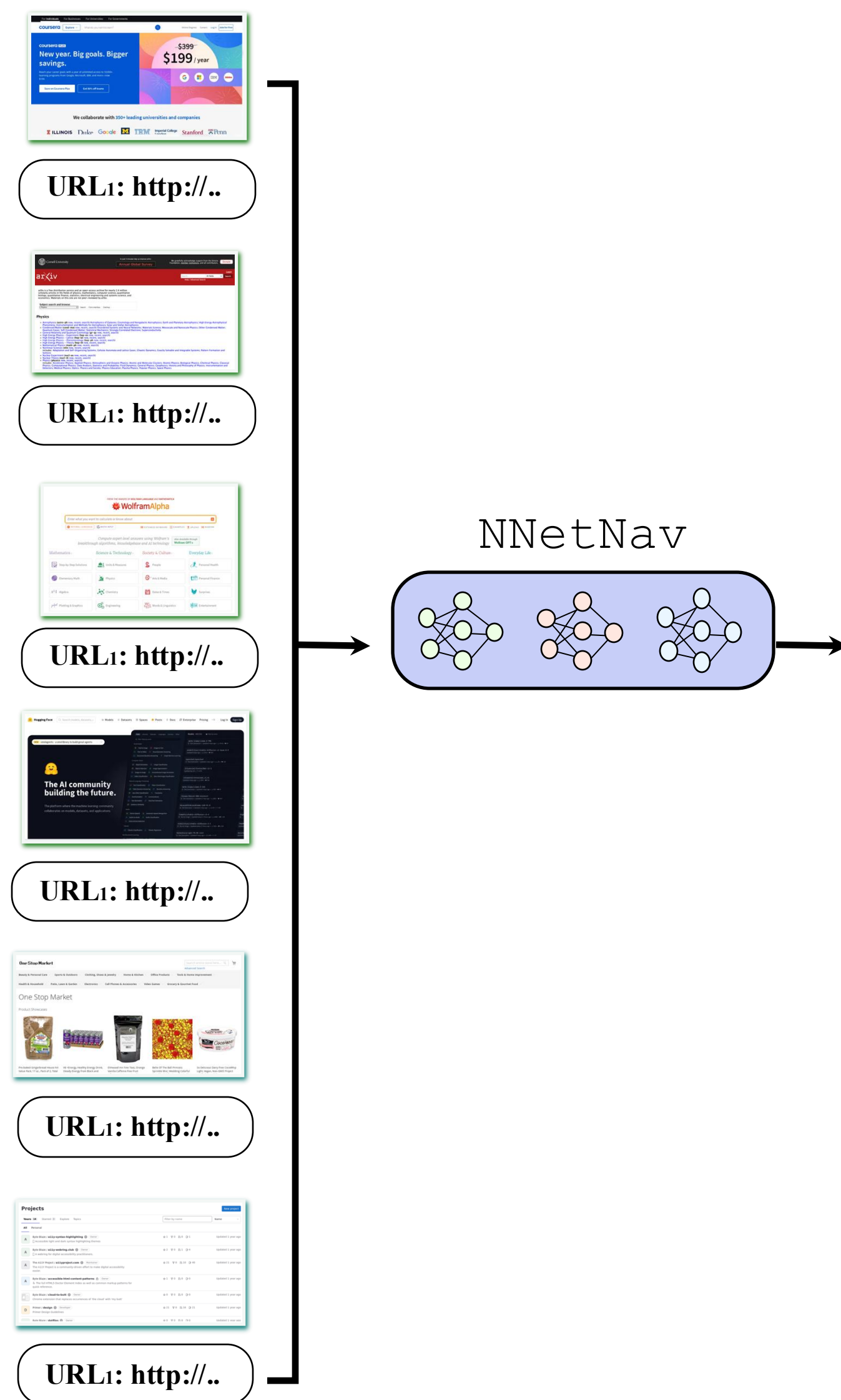
NNetNav: Structuring Exploration for Goal Discovery



NNetNav pruning leads to efficient exploration.

(area in red corresponds to actions that were prevented due to early pruning)

NNetNav so far



Understand the methodology presented in the research paper 'Spectroscopic and X-ray Modeling of the Strong Lensing Galaxy Cluster MACS J0138.0-2155'.

Find requirements and course details for the Master of Applied Data Science degree at the University of Michigan.

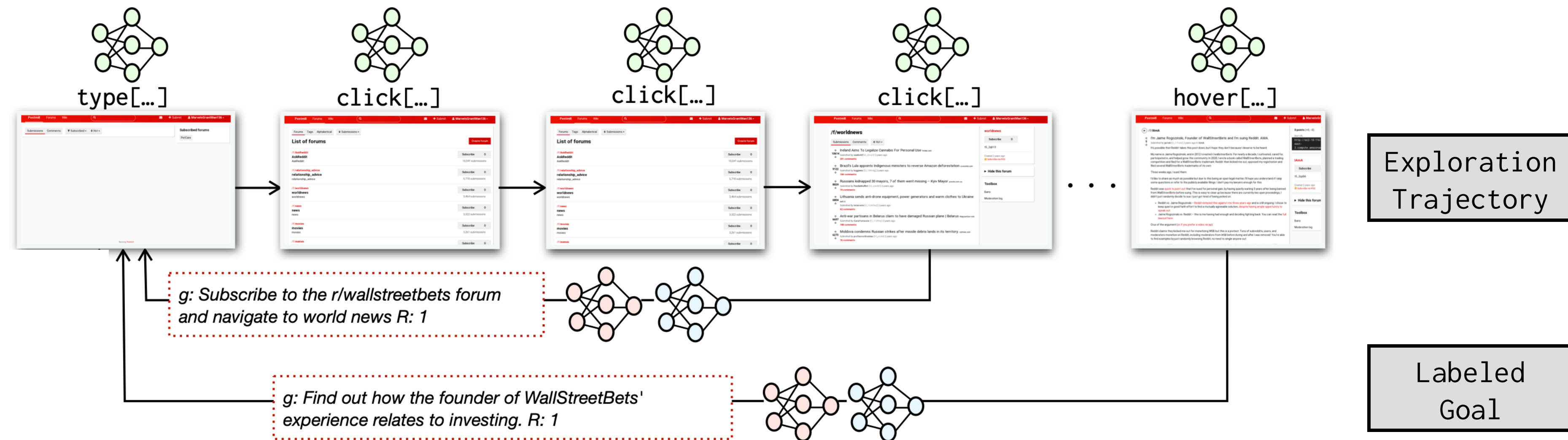
Find detailed information about the chemistry of water, including its structure and properties.

Find the status of your order #000000170 after comparing some gift ideas for skin care products on the One Stop Market website.

Create a new project using the NodeJS Express template and set the project name to "My Game Project", the project description to "A simple express.js project to demonstrate basic game development concepts"

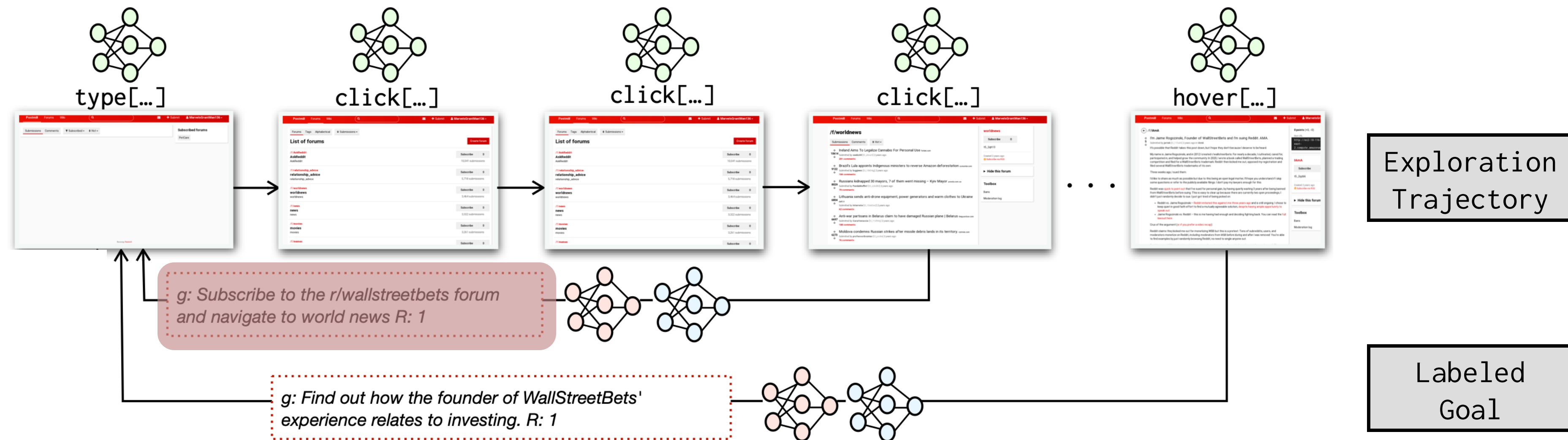
★ Fully unsupervised, open-ended goal discovery on arbitrary websites (using LLMs)!

NNetNav Final Goal: Learning a Mapping from Goals to Trajectories



Just use the exploration trace?

NNetNav Final Goal: Learning a Mapping from Goals to Trajectories



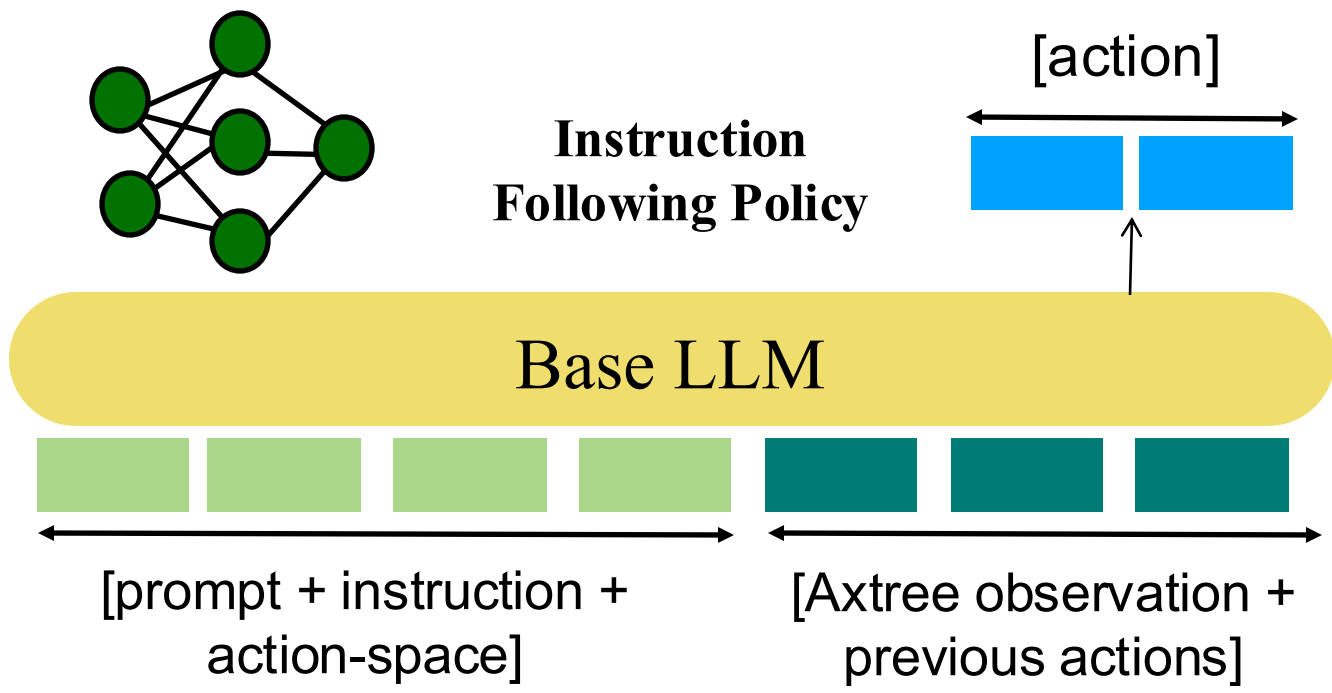
Just use the exploration trace?

✗ Exploration trajectory contains sub-optimal local decisions

NNetNav: Mapping Goals to Trajectories

2. Generate *goal-directed action* for each trajectory prefix, offline

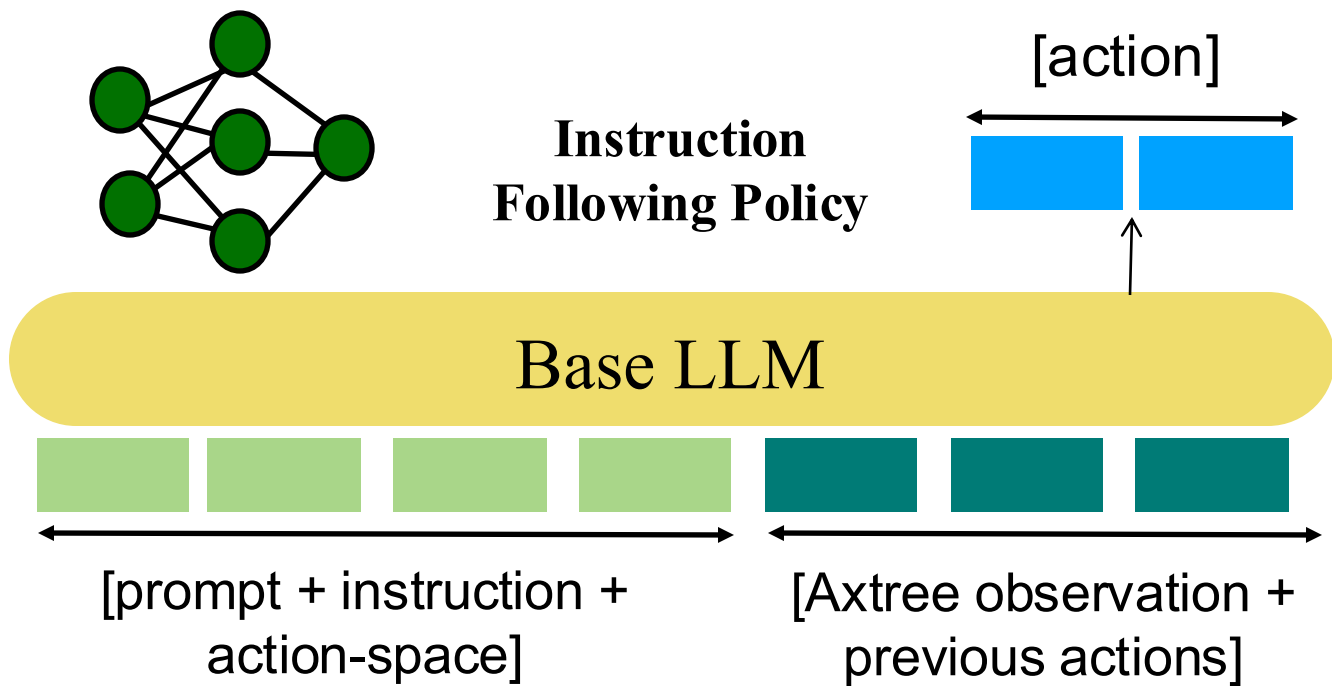
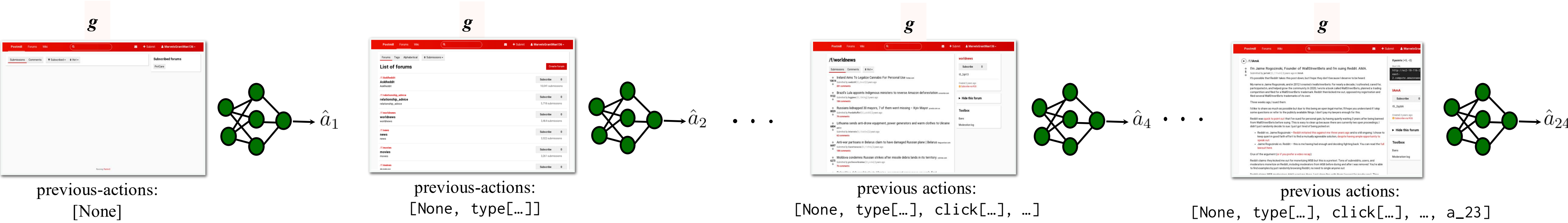
g: Find out how the founder of WallStreetBets' experience relates to investing.



NNetNav: Mapping Goals to Trajectories

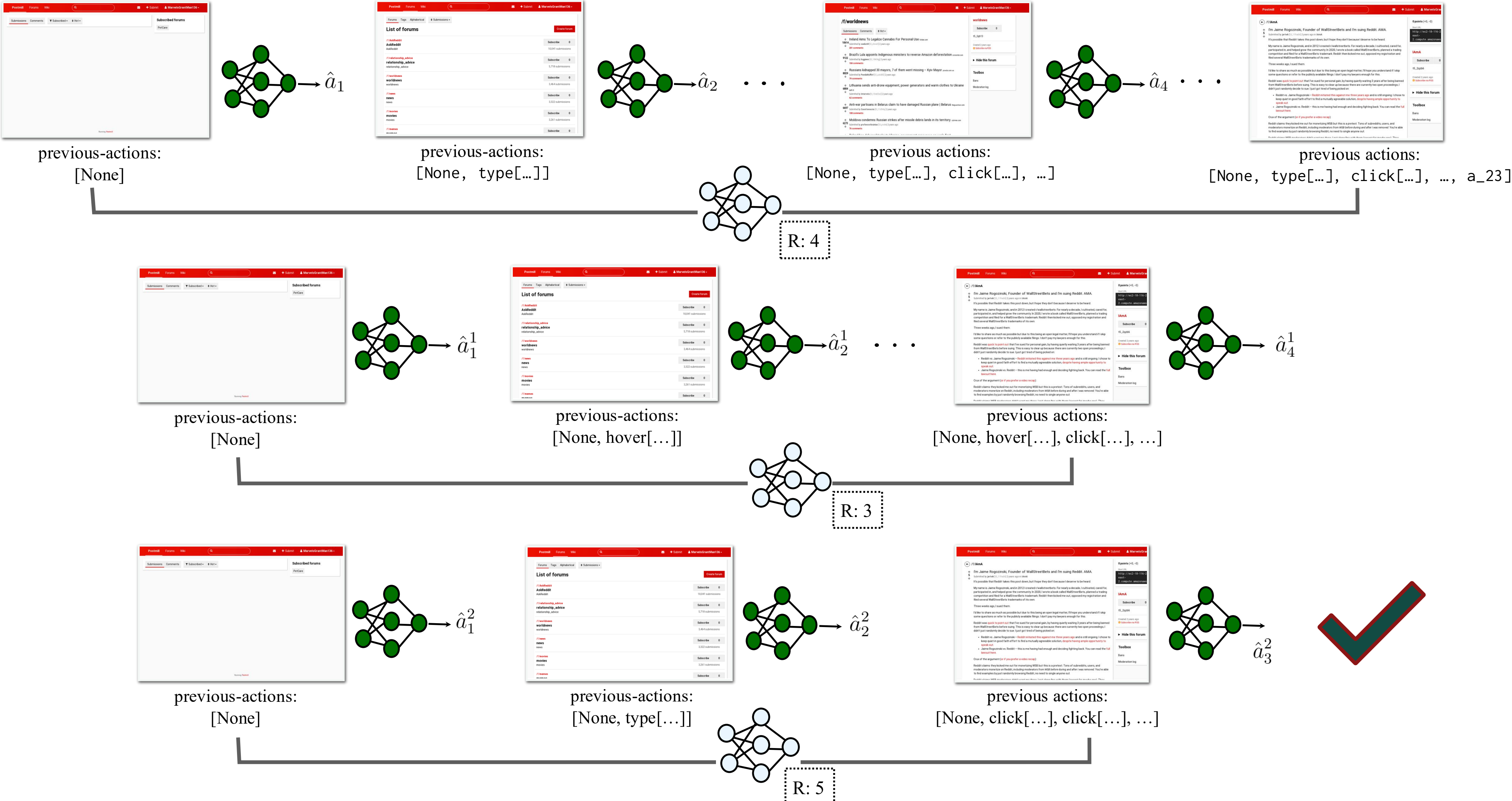
2. Generate goal-directed action for each trajectory prefix, offline

g: Find out how the founder of WallStreetBets' experience relates to investing.



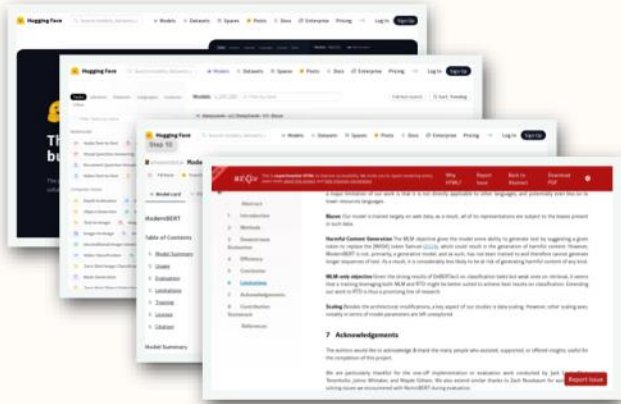
NNetNav: Mapping Goals to Trajectories

3. Best-of-k Rejection Sampling

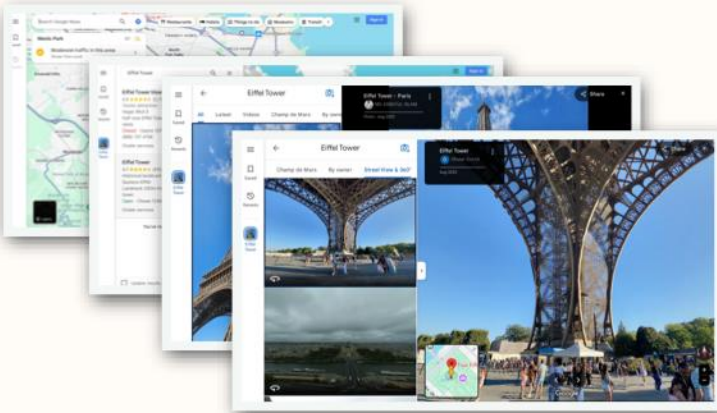


Scaling NNetNav to collect 100k training examples

We used NNetNav to collect ~10k demonstrations (around 100k examples) from 15 live websites and 5 self-hosted websites



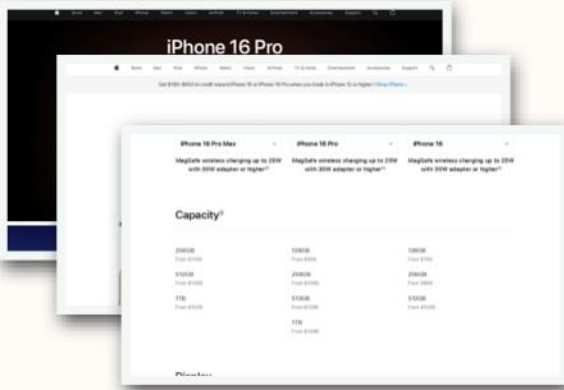
Read the research paper about the ModernBERT-base model from Hugging Face.



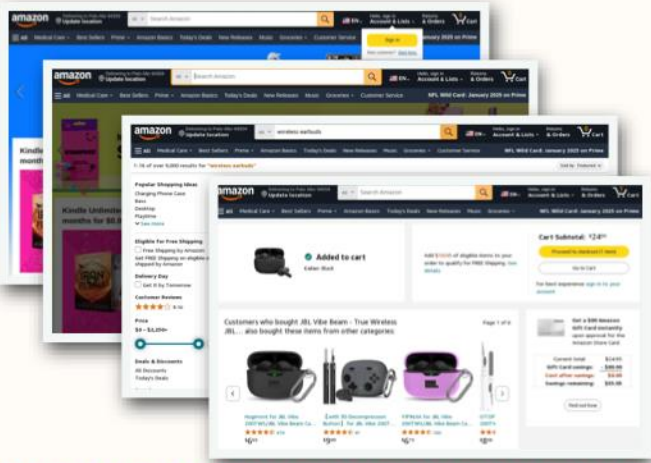
Explore the 360° and Street View images of the Eiffel Tower



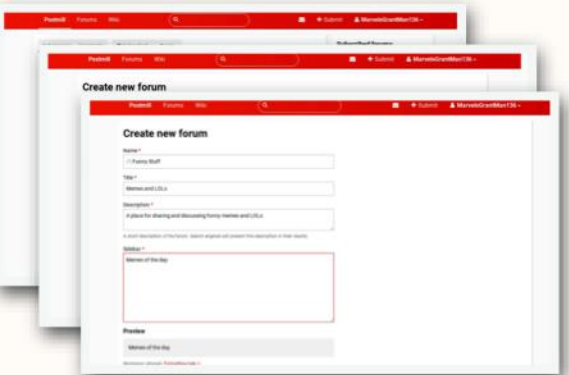
Read a review of Squid games season-2



Find the price of the iPhone 16 and compare it with the price of the iPhone 16 Pro.



Find a pair of wireless earbuds and add them to your cart on Amazon



Create a new forum called "Funny Stuff" with the title "Memes and LOLs", description "A place for sharing and discussing funny memes and LOLs", and sidebar "Memes of the day".

 **Datasets:**  stanfordnlp/nnetnav-wa

 like 0

 Following  Stanford NLP 144

Modalities:  Text

Formats:  json

Size: 10K - 100K

Libraries:  Datasets  pandas  Croissant + 1

License:  apache-2.0

 **Datasets:**  stanfordnlp/nnetnav-live

 like 0

 Following  Stanford NLP 144

Modalities:  Text

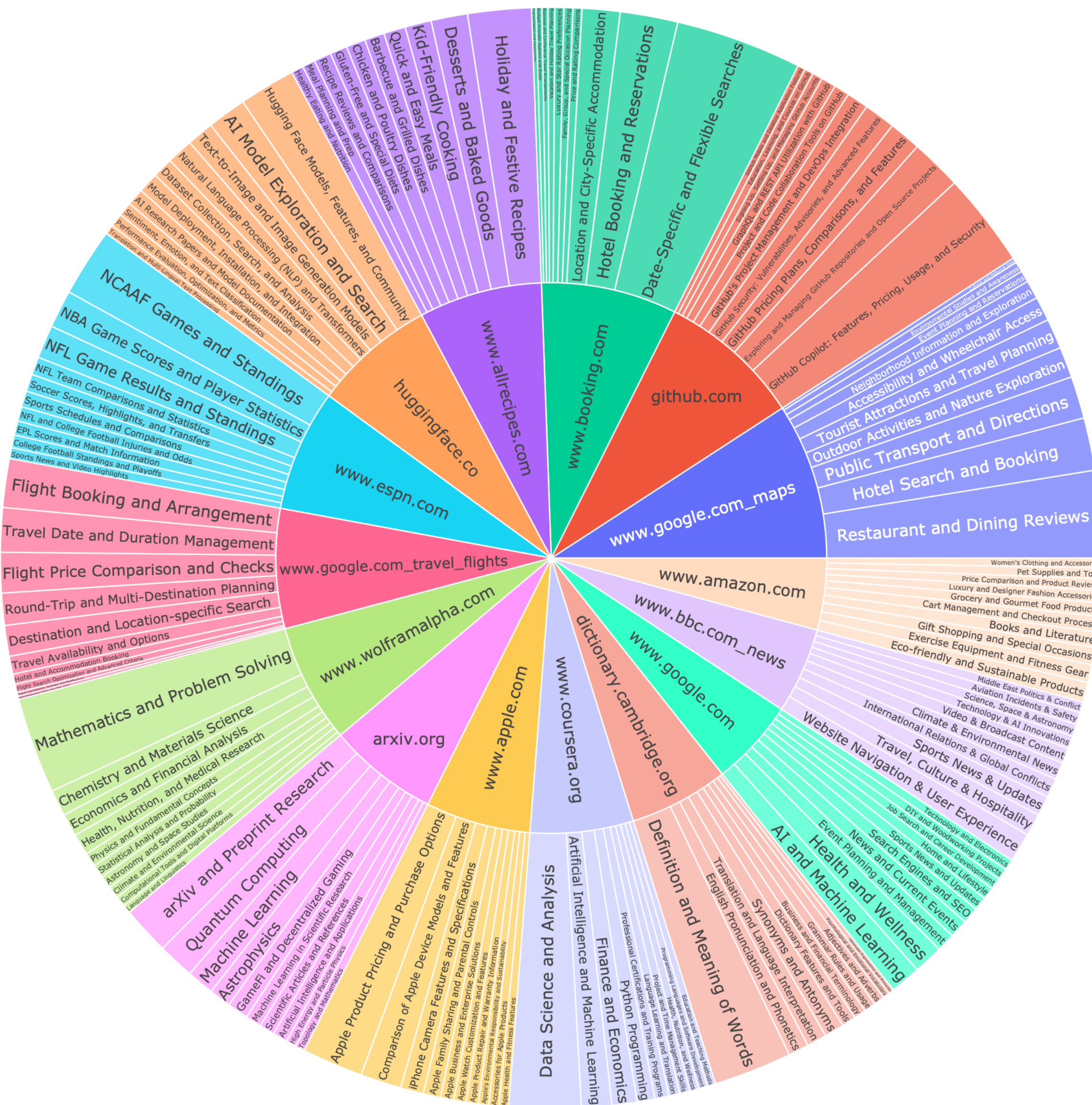
Formats:  json

Size: 10K - 100K

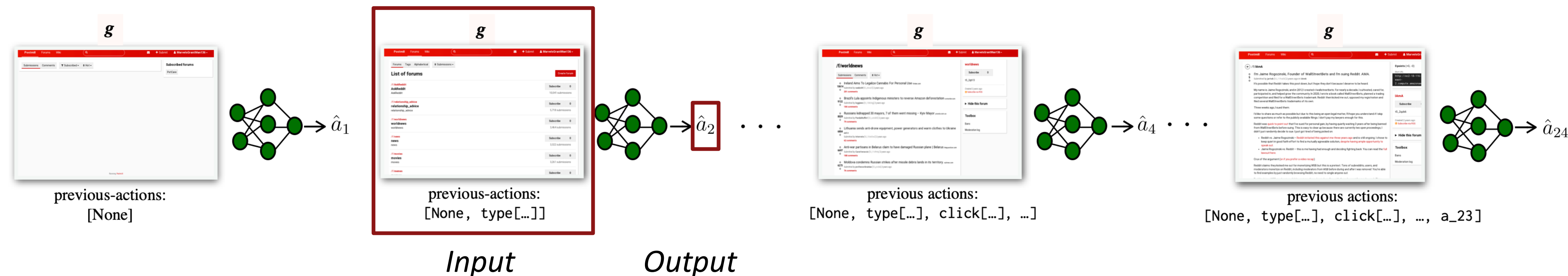
Libraries:  Datasets  pandas  Croissant + 1

License:  apache-2.0

These datasets contain a huge variety of intents covering 100s of use-cases



We used the NNetNav datasets for SFTing the Llama-3.1-8B-instruct model



stanfordnlp/llama8b-nnetnav-wa like 0 Follow Stanford NLP 144

PyTorch stanfordnlp/nnetnav-wa llama arxiv:2410.02907 License: apache-2.0

stanfordnlp/llama8b-nnetnav-live like 0 Follow Stanford NLP 144

PyTorch stanfordnlp/nnetnav-live llama arxiv:2410.02907 License: apache-2.0

Benchmarking Web Agents

We used the NNetNav datasets for SFTing the Llama-3.1-8B-instruct model, which is the *current SoTA* among open LLMs on WebVoyager and WebArena

WebVoyager: Information gathering tasks on live-websites, no logins allowed.

WebArena: Wide range of tasks, that involve changing website state, on self-hosted clones of real websites.

We used the NNetNav datasets for SFTing Llama-3.1-8B-instruct model, which is the *current SoTA* among open LLMs on WebVoyager and WebArena

Agent	#Params	WebArena SR	WebVoyager SR
GPT-4	Unk	14.1	33.5
Llama-3.1-8B	8B	1.0	4.4
Lai et al. 2024	7B	2.5	-
Ou et al. 2024	7B	6.3	-
Patel et al. 2024	72B	9.4	-
LLaVA-7b PAE + Claude (Zhou et al. 2024)	7B	-	22.3
LLaVa-34B-PAE + Claude (Zhou et al. 2024)	34B	-	33.0
Qwen2.5-7B-AgentTrek (Xu et al. 2024)	7B	10.5	-
Qwen2.5-32B-AgentTrek (Xu et al. 2024)	32B	16.3	-
LLama70b	70B	7.1	31.3
LLama8b-NNetNav-WA	8B	16.3	28.1
LLama8B-NNetNav-Live	8B	9.5	35.2
LLama8B-NNetNav-All	8B	14.9	34.1

Ours

We used the NNetNav datasets for SFTing Llama-3.1-8B-instruct model, which is the *current SoTA* among open LLMs on WebVoyager and WebArena

Agent	#Params	WebArena SR	WebVoyager SR
GPT-4	Unk	14.1	33.5
Llama-3.1-8B	8B	1.0	4.4
Lai et al. 2024	7B	2.5	-
Ou et al. 2024	7B	6.3	-
Patel et al. 2024	72B	9.4	-
LLaVA-7b PAE + Claude (Zhou et al. 2024)	7B	-	22.3
LLaVa-34B-PAE + Claude (Zhou et al. 2024)	34B	-	33.0
Qwen2.5-7B-AgentTrek (Xu et al. 2024)	7B	10.5	-
Qwen2.5-32B-AgentTrek (Xu et al. 2024)	32B	16.3	-
LLama70b	70B	7.1	31.3
LLama8b-NNetNav-WA	8B	16.3	28.1
LLama8B-NNetNav-Live	8B	9.5	35.2
LLama8B-NNetNav-All	8B	14.9	34.1

*outperforms zero-shot
GPT-4*

Ours

We used the NNetNav datasets for SFTing Llama-3.1-8B-instruct model, which is the *current SoTA* among open LLMs on WebVoyager and WebArena

Agent	#Params	WebArena SR	WebVoyager SR
GPT-4	Unk	14.1	33.5
Llama-3.1-8B	8B	1.0	4.4
Lai et al. 2024	7B	2.5	-
Ou et al. 2024	7B	6.3	-
Patel et al. 2024	72B	9.4	-
LLaVA-7b PAE + Claude (Zhou et al. 2024)	7B	-	22.3
LLaVa-34B-PAE + Claude (Zhou et al. 2024)	34B	-	33.0
Qwen2.5-7B-AgentTrek (Xu et al. 2024)	7B	10.5	-
Qwen2.5-32B-AgentTrek (Xu et al. 2024)	32B	16.3	-
LLama70b	70B	7.1	31.3
LLama8b-NNetNav-WA	8B	16.3	28.1
LLama8B-NNetNav-Live	8B	9.5	35.2
LLama8B-NNetNav-All	8B	14.9	34.1

outperforms zero-shot GPT-4

OSS VLM models not as good?

Ours

We used the NNetNav datasets for SFTing Llama-3.1-8B-instruct model, which is the *current SoTA* among open LLMs on WebVoyager and WebArena

Agent	#Params	WebArena SR	WebVoyager SR
GPT-4	Unk	14.1	33.5
Llama-3.1-8B	8B	1.0	4.4
Lai et al. 2024	7B	2.5	-
Ou et al. 2024	7B	6.3	-
Patel et al. 2024	72B	9.4	-
LLaVA-7b PAE + Claude (Zhou et al. 2024)	7B	-	22.3
LLaVa-34B-PAE + Claude (Zhou et al. 2024)	34B	-	33.0
Qwen2.5-7B-AgentTrek (Xu et al. 2024)	7B	10.5	-
Qwen2.5-32B-AgentTrek (Xu et al. 2024)	32B	16.3	-
LLama70b	70B	7.1	31.3
LLama8b-NNetNav-WA	8B	16.3	28.1
LLama8B-NNetNav-Live	8B	9.5	35.2
LLama8B-NNetNav-All	8B	14.9	34.1

outperforms zero-shot GPT-4

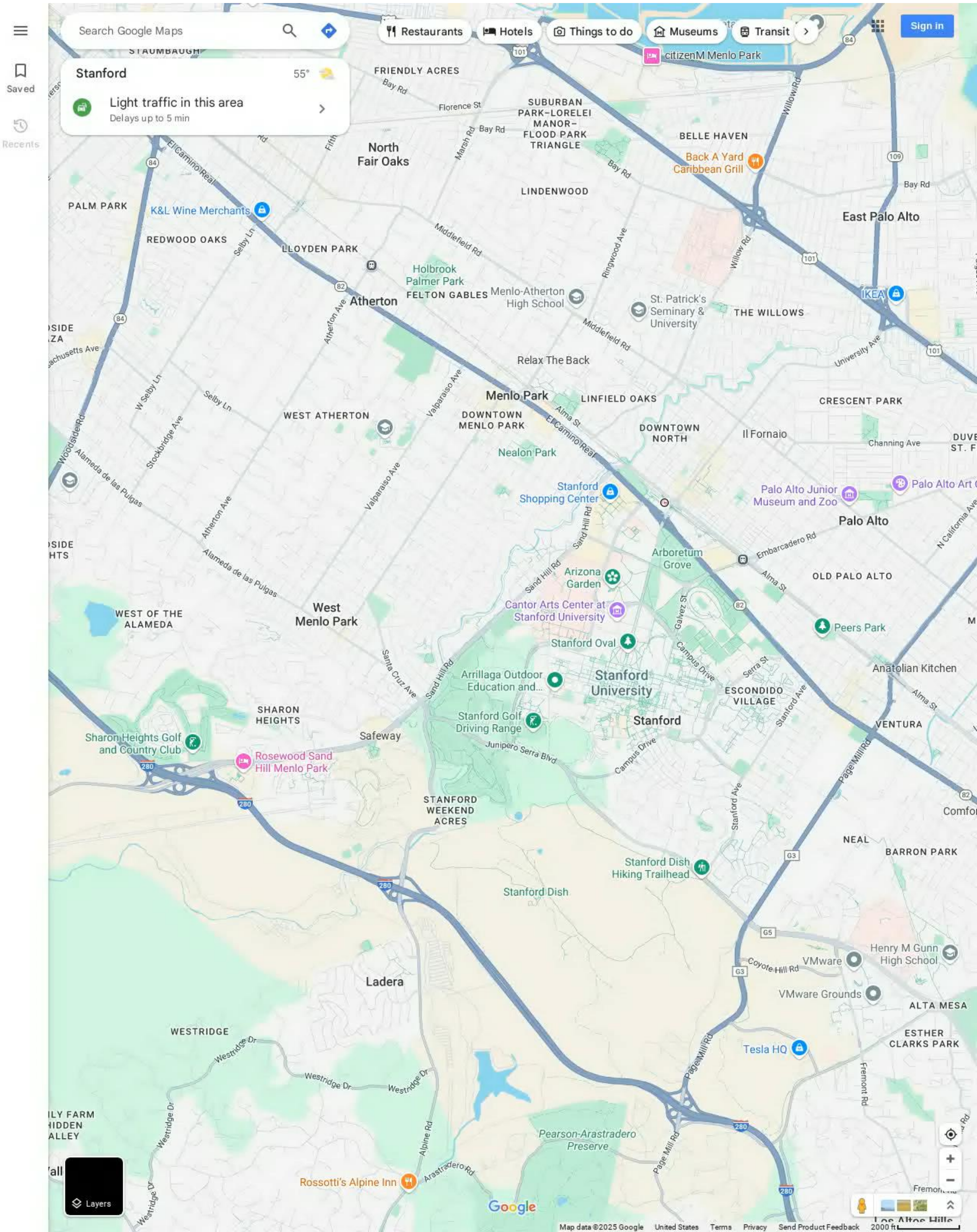
OSS VLM models not as good?

NNetNav is not just distillation

Ours

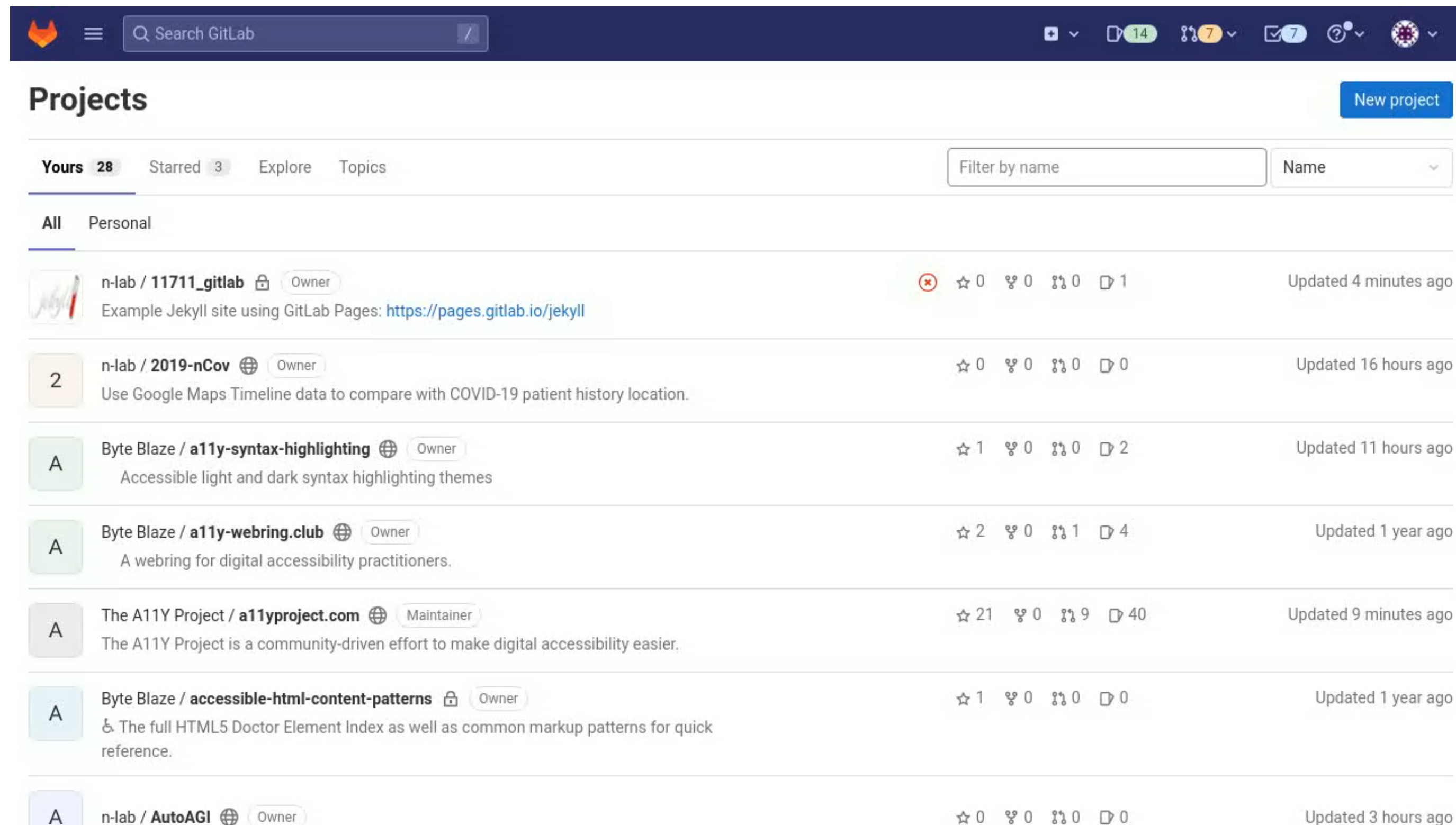
Some rollouts from LLama8b-NNetNav-Live

Find a place to climb within 2 miles of zip code 90028.



Some rollouts from LLama8b-NNetNav-WA

For safety reasons, we do not let the model login to live websites, limiting the kinds of tasks it can do. To test for tasks that require login, we evaluate on WebArena.



Assigning issues on GitHub

Some rollouts from LLama8b-NNetNav-WA

For safety reasons, we do not let the model login to live websites, limiting the kinds of tasks it can do. To test for tasks that require login, we evaluate on WebArena.

DASHBOARD

SALES

CATALOG

CUSTOMERS

MARKETING

CONTENT

REPORTS

STORES

SYSTEM

FIND PARTNERS & EXTENSIONS

Failed to synchronize data to the Magento Business Intelligence service. [Retry Synchronization](#)

System Messages: 1

Dashboard

Search, Notifications, User: admin

Scope: All Store Views

Reload Data

Advanced Reporting

Gain new insights and take command of your business' performance, using our dynamic product, order, and customer reports tailored to your customer data.

Go to Advanced Reporting

Lifetime Sales

\$0.00

Average Order

\$0.00

Chart is disabled. To enable the chart, click [here](#).

Revenue

\$0.00

Tax

\$0.00

Shipping

\$0.00

Quantity

0

Last Orders

Customer	Items	Total
Sarah Miller	5	\$0.00
Grace Nguyen	4	\$190.00

Bestsellers

Most Viewed Products

New Customers

Customers

Product	Price	Quantity
Sprite Stasis Ball 65 cm	\$27.00	6

Creating a shipping report
★: self-correction!

NNetNav Summary

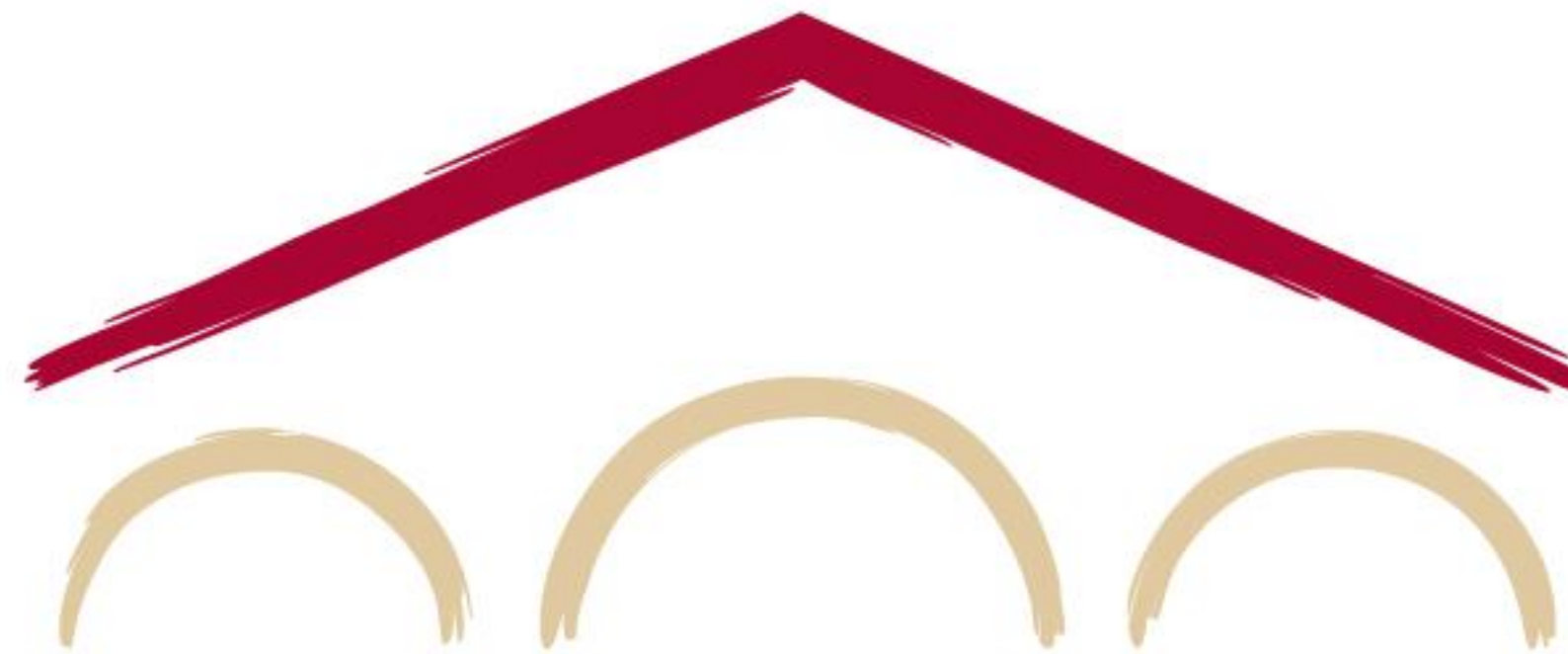
NNetNav is a method of learning by unsupervised interaction. It builds demonstrations by interacting with live websites by:

1. Open-ended goal discovery with a hierarchically biased exploration policy
2. Mapping goals to trajectories with best-of-k rejection sampling

NNetNav-trained Llama-8b became the SoTA among unsupervised methods using OSS models, outperforming zero-shot GPT-4



The Surprising Victory of NLP: From Philosophy to Agentic Language Models



Christopher Manning

@chrmanning ✿ manning@stanford.edu

Founder, Stanford NLP Group ✿ Senior Fellow, HAI

KDD 2025 ✿ Toronto

