注目論文から読み解くこれからのAI
Future AI direction along with notable papers

辻 真吾（TSUJI Shingo）

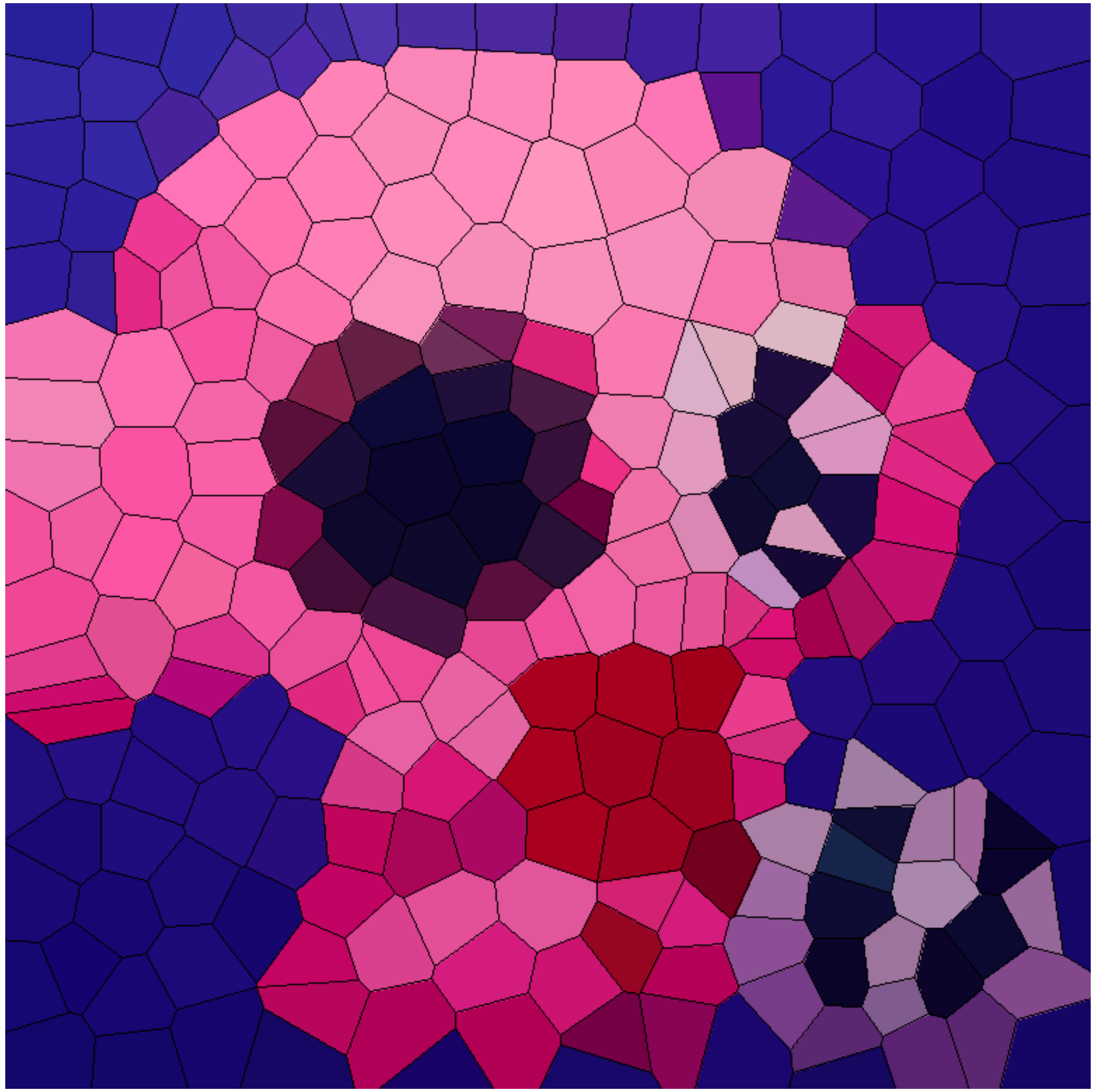
1/24, 2026

自己紹介 Who am I?

- 辻真吾（つじしんご）
  - とある大学の研究所に勤めていることになっています
- Pythonを使ったデータ分析や数理最適化が得意
  - www.tsjshg.info
- Python、データサイエンス、アルゴリズムに関する著書多数
- みんなのPython勉強会（https://startpython.connpass.com/）を月に1回やってます
- 最近読んで面白かった本『テクノロジーバブル（日経BP）』

- TSUJI Shingo
  - RCAST, The University of Tokyo
- I love data science with Python
  - www.tsjshg.info
  - @tsjshg (Twitter)
- I wrote many books about Python, data science and algorithm
- After AI, I think the next hot topic will be mathematical optimization.

pip install imgrit



https://github.com/tsjshg/imgrit

Agenda

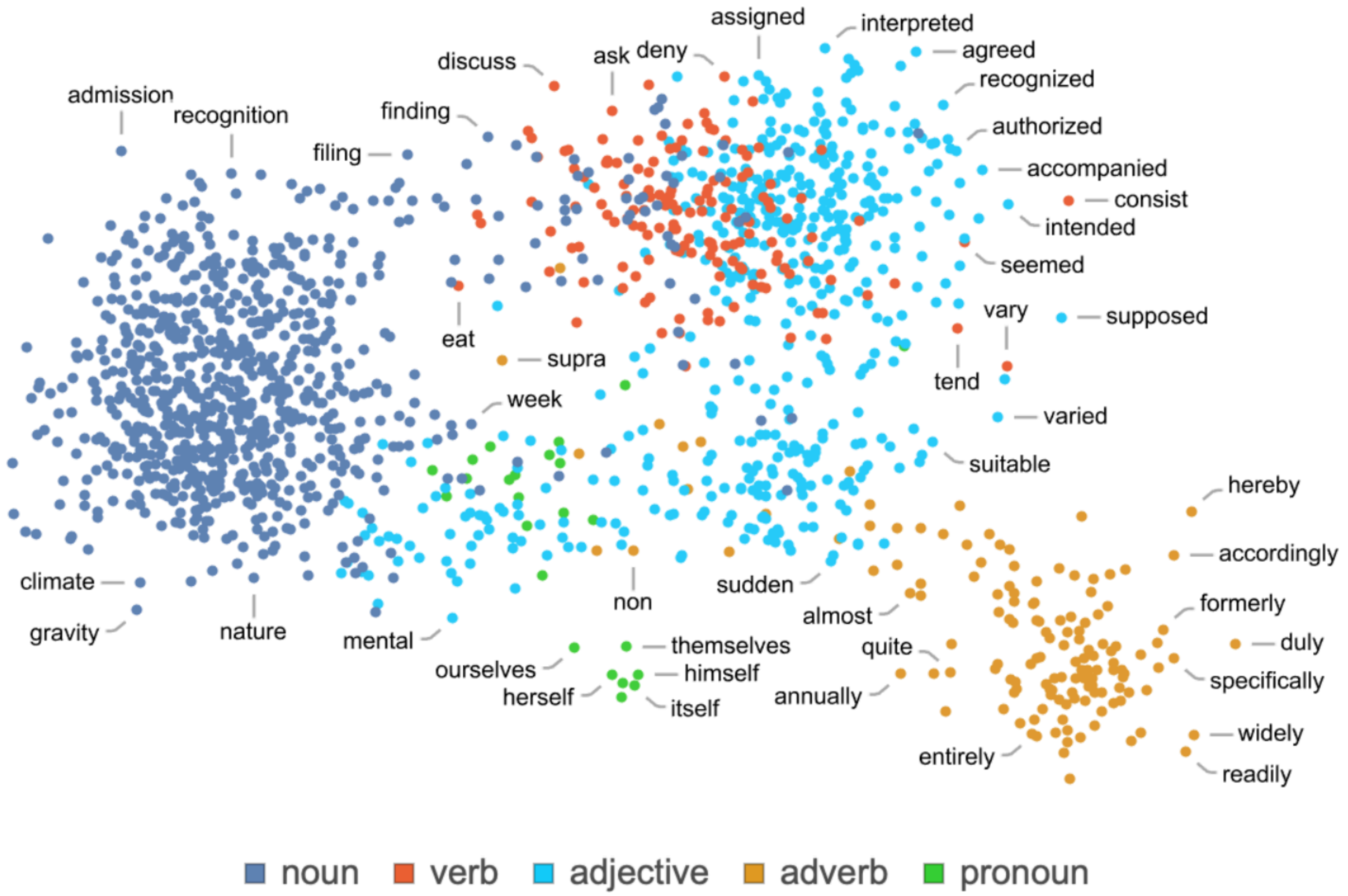
- AIとはベクトル表現である / AI is vector representation

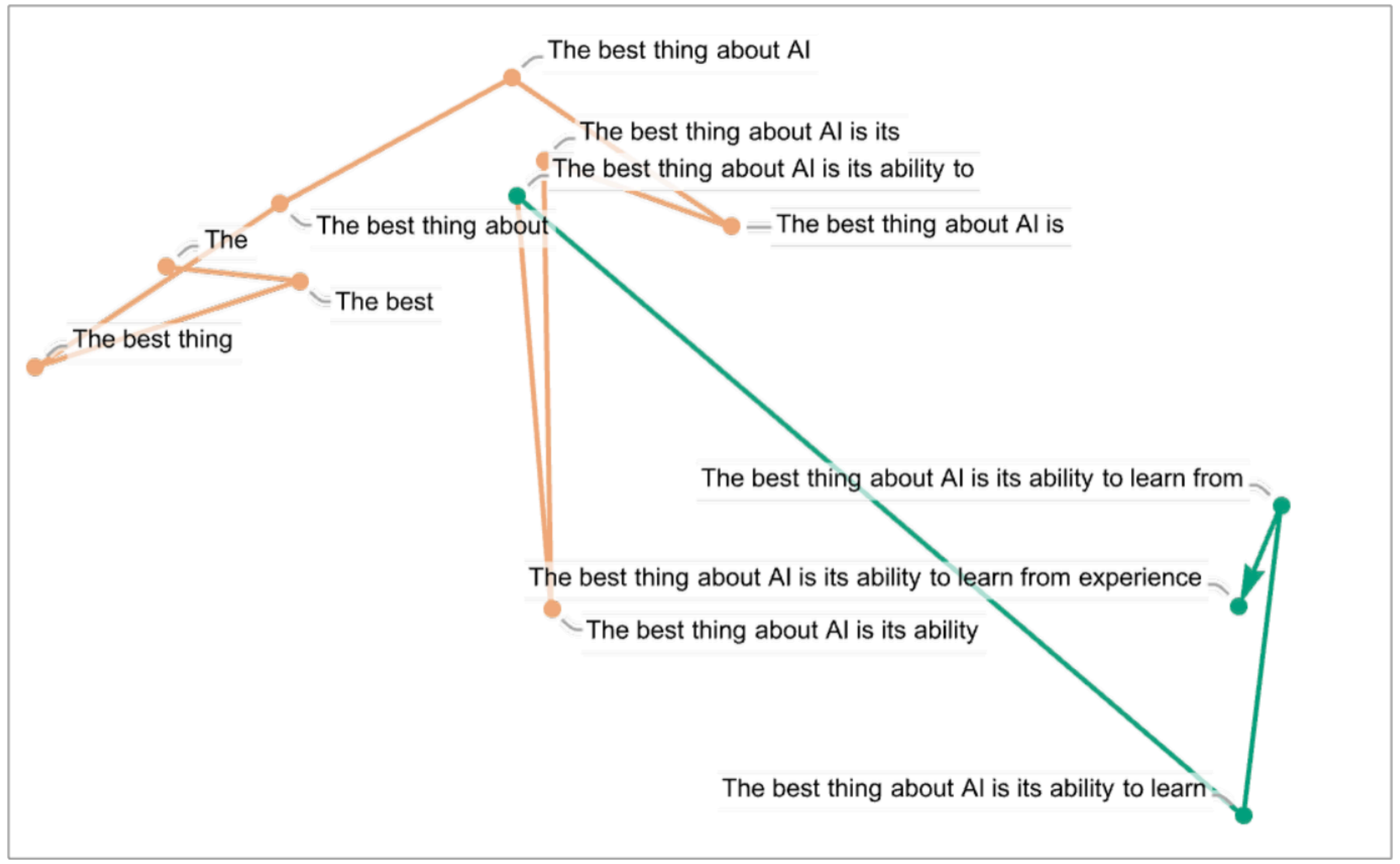
- LLMの柔軟性 / How flexible LLM is!

- プラトニック表現仮説 / Platonic representation hypothesis
- Best paper awards of NeurIPS 2025 / Artificial Hivemind
- AIに多様性を持たせるVerbalized sampling

- The illusion of thinking

AI is strongly supported by semantic vector representations





画像、単語、文章の意味を反映した高次元数値ベクトル表現をニューラルネットワークによって作れるようになった（埋め込み表現の学習）ことが最近のAIの発展を支えている

semantically good vector representation of images, words and sentences are essential key techniques for modern amazing AI advancements.

What Is ChatGPT Doing ... and Why Does It Work? (very good introduction for today’s LLM-based generative AI)
https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

LLMの柔軟性を示す研究

In-context representation

How flexible LLM is

ICLR: In-Context Learning of Representations

Core Francisco Park[*1,2,3] , **Andrew Lee**[*4], **Ekdeep Singh Lubana**[*1,3], **Yongyi Yang**[*1,3,5], **Maya Okawa**[1,3], **Kento Nishi**[1,4], **Martin Wattenberg**[4], **& Hidenori Tanaka**[1,3]
[1]CBS-NTT Program in Physics of Intelligence, Harvard University
[2]Department of Physics, Harvard University
[3]Physics & Informatics Lab, NTT Research Inc.
[4]SEAS, Harvard University
[5]CSE, University of Michigan, Ann Arbor

Abstract

Recent work has demonstrated that semantics specified by pretraining data influence how representations of different concepts are organized in a large language model (LLM). However, given the open-ended nature of LLMs, e.g., their ability to in-context learn, we can ask whether models alter these pretraining semantics to adopt alternative, context-specified ones. Specifically, if we provide in-context exemplars wherein a concept plays a different role than what the pretraining data suggests, do models reorganize their representations in accordance with these novel semantics? To answer this question, we take inspiration from the theory of *conceptual role semantics* and define a toy "graph tracing" task wherein the nodes of the graph are referenced via concepts seen during training (e.g., `apple`, `bird`, etc.) and the connectivity of the graph is defined via some predefined structure (e.g., a square grid). Given exemplars that indicate traces of random walks on the graph, we analyze intermediate representations of the model and find that *as the amount of context is scaled, there is a sudden re-organization from pretrained semantic representations to* **in-context representations** *aligned with the graph structure.* Further, we find that when reference concepts have correlations in their semantics (e.g., `Monday`, `Tuesday`, etc.), the context-specified graph structure is still present in the representations, but is unable to dominate the pretrained structure. To explain these results, we analogize our task to energy minimization for a predefined graph topology, providing evidence towards an implicit optimization process to infer context-specified semantics. Overall, our findings indicate scaling context-size can flexibly re-organize model representations, possibly unlocking novel capabilities.

- LLMのプロンプトエンジニアリングでモデルの挙動がなぜ変わるのかを説明する第1歩になりそうな報告
- 追加トレーニングなしに単語のベクトル表現を変更する方法（in-context representations）を提案

- The paper proposed the potential research direction to elucidate why the in-context learning is so effective.
- Without any fine-tuning, in-context representations can modify the vector representation of same words.

12/29, 2024 https://arxiv.org/abs/2501.00070

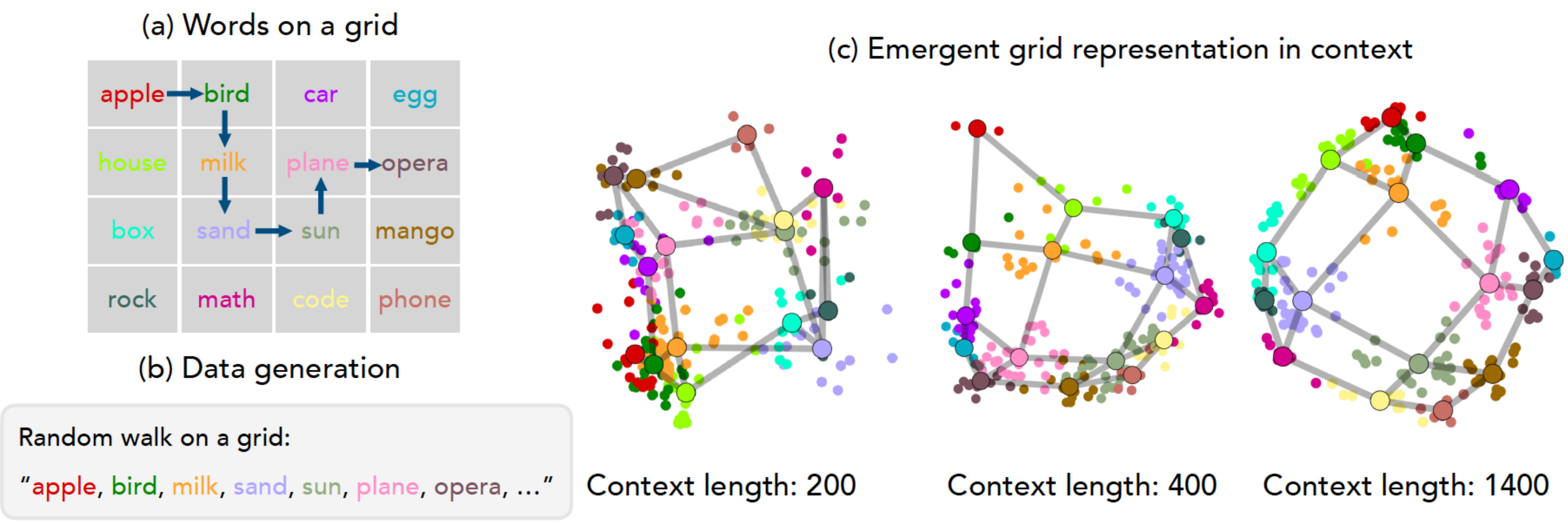


Figure 1: **Alteration of representations in accordance with context-specified semantics (grid structure).** (a) We randomly arrange a set of concepts on a grid that does not reflect any correlational semantics between the tokens. (b) We then generate sequences of tokens following a random walk on the grid, inputting it as context to a Llama-3.1-8B model. (c) The model's mean token representations projected onto the top two principal components. As the number of in-context exemplars increases, there is a formation of representations mirroring the grid structure underlying the data-generating process. Representations are from the residual stream activation following layer 26.

意味がバラバラの単語を格子状に固定してrandom walkで列を作る

この長さが長くなると、それぞれの単語のベクトル表現が意味主体から並び主体に変わる

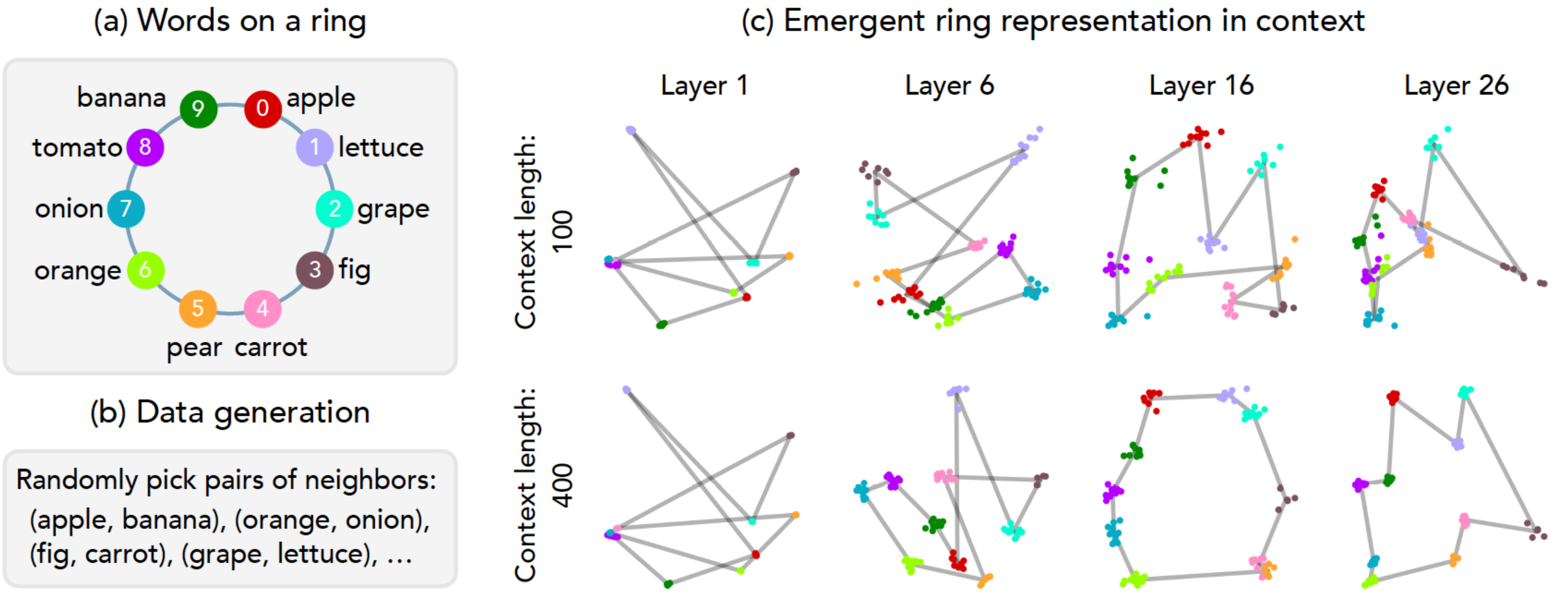


Figure 2: **Alteration of representations in accordance with context-specified semantics (ring structure).** (a) We randomly place concepts on a ring structure unrelated to their semantics. (b) We then generate sequences of tokens by randomly sampling *neighboring pairs* from the ring which is used as the input context to a Llama-3.1-8B model. (c) The model's mean representation of tokens projected onto the top two principal components. As the number of in-context exemplars increases, there is a formation of representations mirroring the ring structure underlying the data-generating process. The representations are from the residual stream activations.

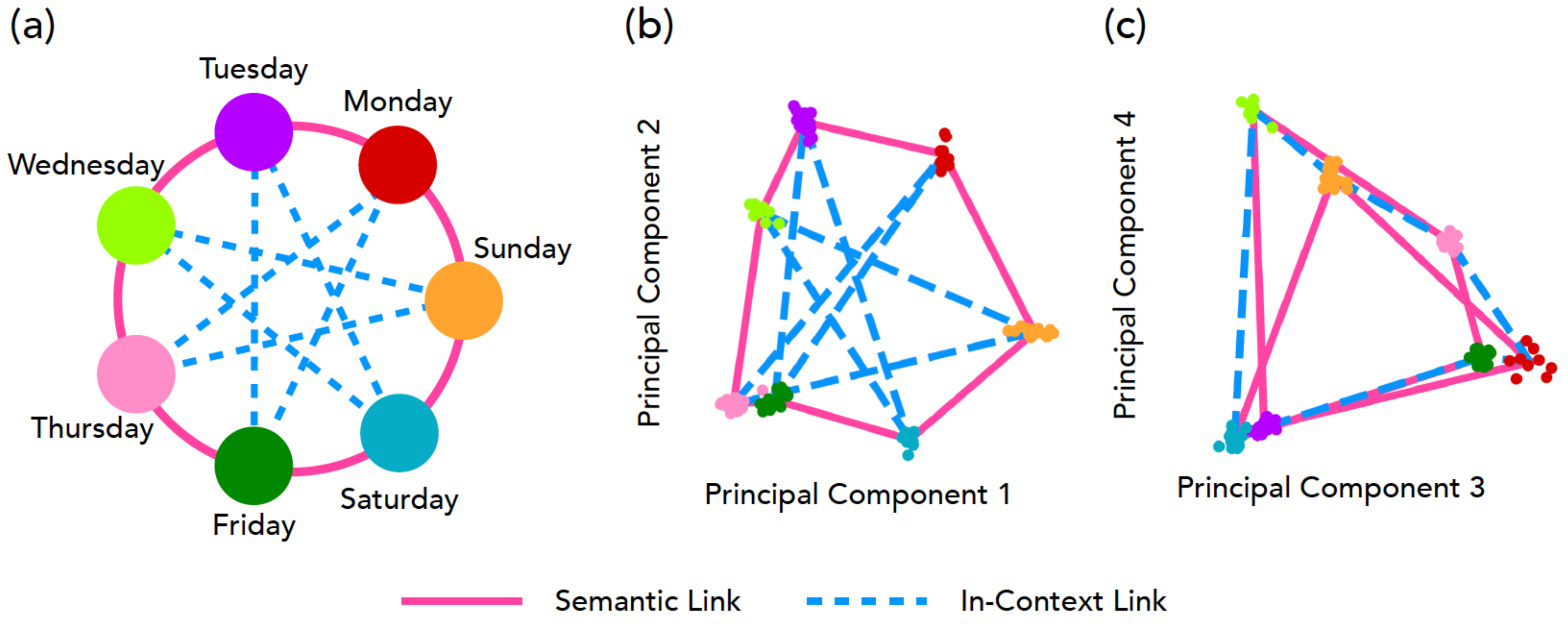


Figure 3: **In-context representations form in higher principal components in the presence of semantic priors.** (a) (Purple) Semantic links underlying days of the week. (Dashed blue) We define a non-semantic graph structure by linking non-neighboring days and generate tokens from this graph. (b) (Purple) The ring geometry formed by semantic links established during pre-training remains intact in the first two principal components. (c) (Dashed blue) The non-semantic structure provided in-context can be seen in the third and fourth principal components. Note that the star structure in the first two components (b), which match the ground truth graphical structure of our data generating process (a), becomes a ring in the next two principal components (c). The representations are from the residual stream activation following layer 21.

格子ではなく環状にならべても同じ結果（左）

The same results with circular layouts（left)

曜日の場合は事前学習の影響が大きいので、PCAの第1,2主成分では出ず第3,4主成分で見える（右）

in case of weekday, we can see the in-context representation form only in third ant fourth principal components not in first and second because the pertaining effects are strong.(right)

相転移になぞらえている

It is similar to phase transition.

Dirichlet Energy. We measure the *Dirichlet energy* of our graph $\mathcal{G}$'s structure by defining an energy function over the model representations. Specifically, for an undirected graph $\mathcal{G}$ with n nodes, let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be its adjacency matrix, and $\boldsymbol{x} \in \mathbb{R}^n$ be a signal vector that assigns a value x_i to each node i. Then the Dirichlet energy of the graph with respect to $\boldsymbol{x}$ is defined as

$$E_{\mathcal{G}}(\boldsymbol{x}) = \sum_{i,j} \boldsymbol{A}_{i,j}(x_i - x_j)^2. \tag{1}$$

For a multi-dimensional signal, the Dirichlet energy is defined as the summation of the energy over each dimension. Specifically, let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be a matrix that assigns each node i with a d-dimensional vector $\boldsymbol{x}_i$, then the Dirichlet energy of $\boldsymbol{X}$ is defined by

$$E_{\mathcal{G}}(\boldsymbol{X}) = \sum_{k=1}^{d}\sum_{i,j} \boldsymbol{A}_{i,j}(x_{i,k} - x_{j,k})^2 = \sum_{i,j} \boldsymbol{A}_{i,j}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2. \tag{2}$$

Overall, to empirically quantify the formation of geometric representations, we can measure the Dirichlet energy with respect to the graphs underlying our data generating processes (DGPs) and our mean token activations $\boldsymbol{h}^{\ell}_{\mathcal{T}}$:

$$E_{\mathcal{G}}(\boldsymbol{H}^{\ell}(\mathcal{T})) = \sum_{i,j} \boldsymbol{A}_{i,j}\|\boldsymbol{h}^{\ell}_i - \boldsymbol{h}^{\ell}_j\|^2, \tag{3}$$

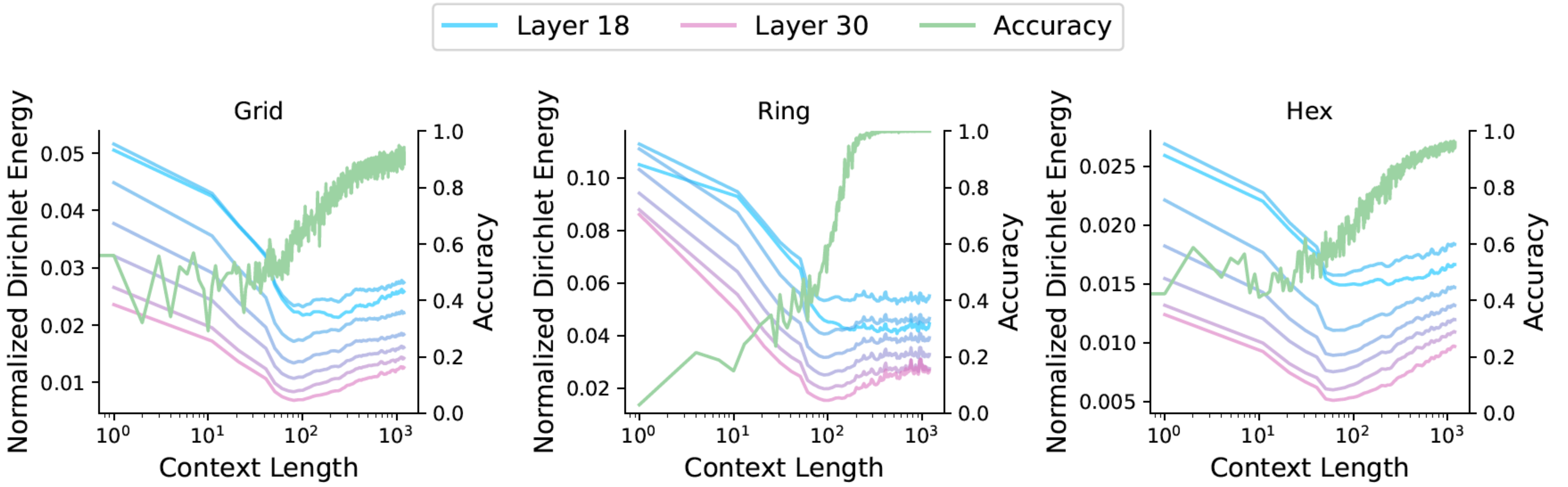


Figure 4: **A model continuously develops task representation as it learns to traverse novel graphs in-context.** We plot the accuracy of graph traversal and the Dirichlet energy of the graph, computed from the model's internal representations, as functions of context length. We note that the Dirichlet energy never reaches a perfect zero—ruling out that the representations are learning a degenerate structure, as was also seen in the PCA visualizations in Sec. 3. (a) A 4x4 grid graph with 16 nodes. (b) A circular ring with 10 nodes. (c) A "honey-comb" hexagonal lattice, with 30 nodes.

プロンプトエンジニアリングが効いたりするLLMの柔軟性を象徴する現象

LLM has inherent flexibilities that makes prompt engineering so effective.

関連Tweetアニメーションあり（you can see the movies explaining this research)

https://x.com/corefpark/status/1875929881856573905

プラトニック表現仮説のひろがり
Platonic representation hypothesis and its implications

プラトンのイデア論 Platonic idealism

- ギリシャの哲学者プラトン先生が唱えた理論の1つで、唯一の真実の世界が1つあり、実際の世界では真実に似ている別のモノにしか触れられない

- In our physical world, we can not touch the objects of ideal world. And there is only one ideal world outside our real life. The concept are compared with fixed person inside caves.

Google Nano Banana Pro作成

洞窟の中で鎖につながれているところまではいいが、体の向きが真実の世界が見えてしまっている向き

The AI does not truly understand the meaning of Platon’s theory of idea.

Position: The Platonic Representation Hypothesis

Minyoung Huh, Brian Cheung, Tongzhou Wang, Phillip Isola *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235:20617-20642, 2024.

Abstract

We argue that representations in AI models, particularly deep networks, are converging. First, we survey many examples of convergence in the literature: over time and across multiple domains, the ways by which different neural networks represent data are becoming more aligned. Next, we demonstrate convergence across data modalities: as vision models and language models get larger, they measure distance between datapoints in a more and more alike way. We hypothesize that this convergence is driving toward a shared statistical model of reality, akin to Plato's concept of an ideal reality. We term such a representation the platonic representation and discuss several possible selective pressures toward it. Finally, we discuss the implications of these trends, their limitations, and counterexamples to our analysis.

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.

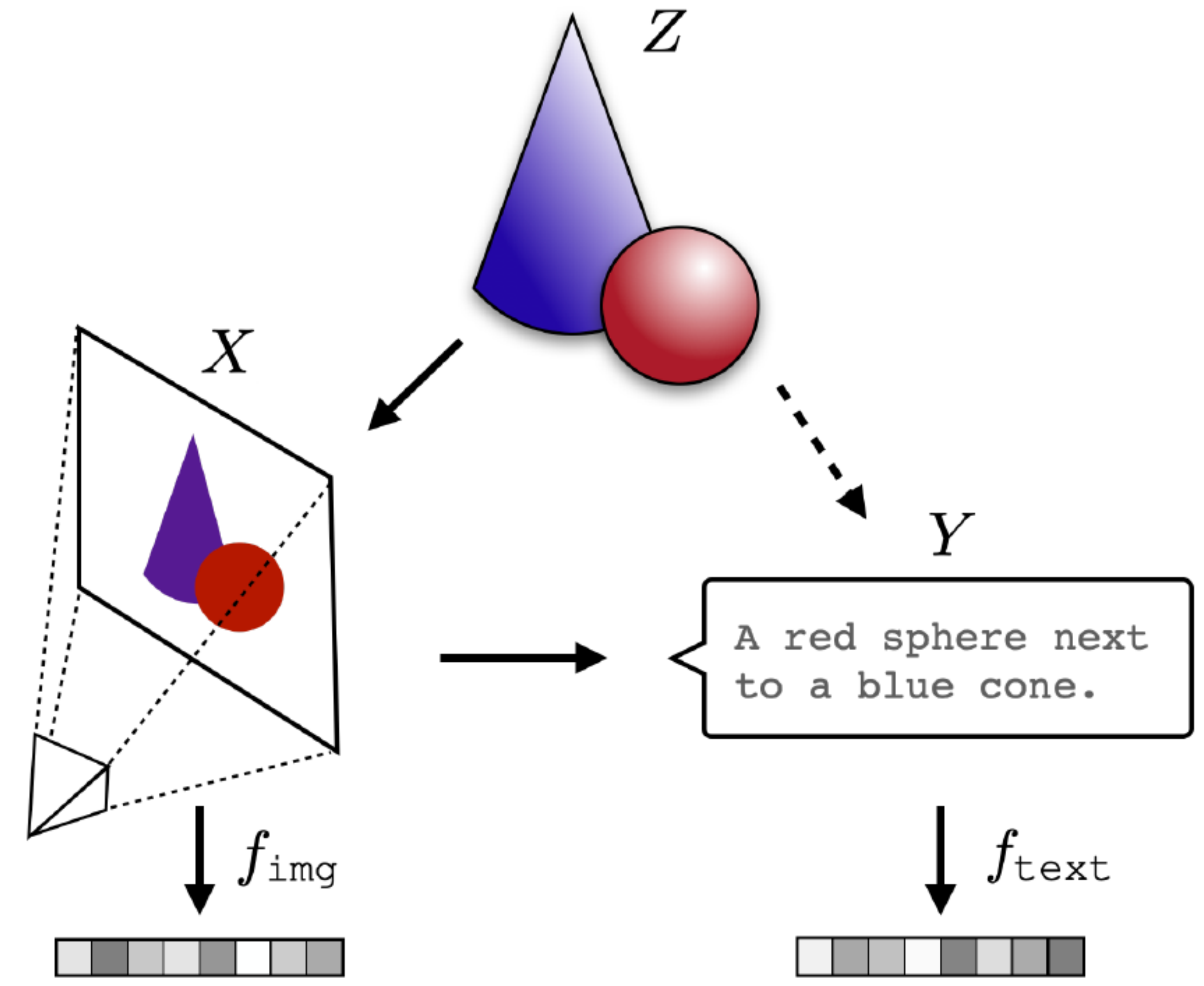


Figure 1. **The Platonic Representation Hypothesis:** Images (X) and text (Y) are projections of a common underlying reality (Z). We conjecture that representation learning algorithms will converge on a shared representation of Z, and scaling model size, as well as data and task diversity, drives this convergence.

プラトニック表現仮説

MITのグループが2024年に提唱

言語モデルや画像モデルなどモデルの違いはあっても、似たようなモノ（赤い球と青い三角錐）を表現する統計モデルが収束するという仮説

representation by different neural networks are becoming more aligned, and the convergence phenomenon is multi-modal.

https://proceedings.mlr.press/v235/huh24a.html

(ややこしいので割愛予定)
性能が良いモデルほど互いに似ていて、これにはマルチモーダル性が認められる

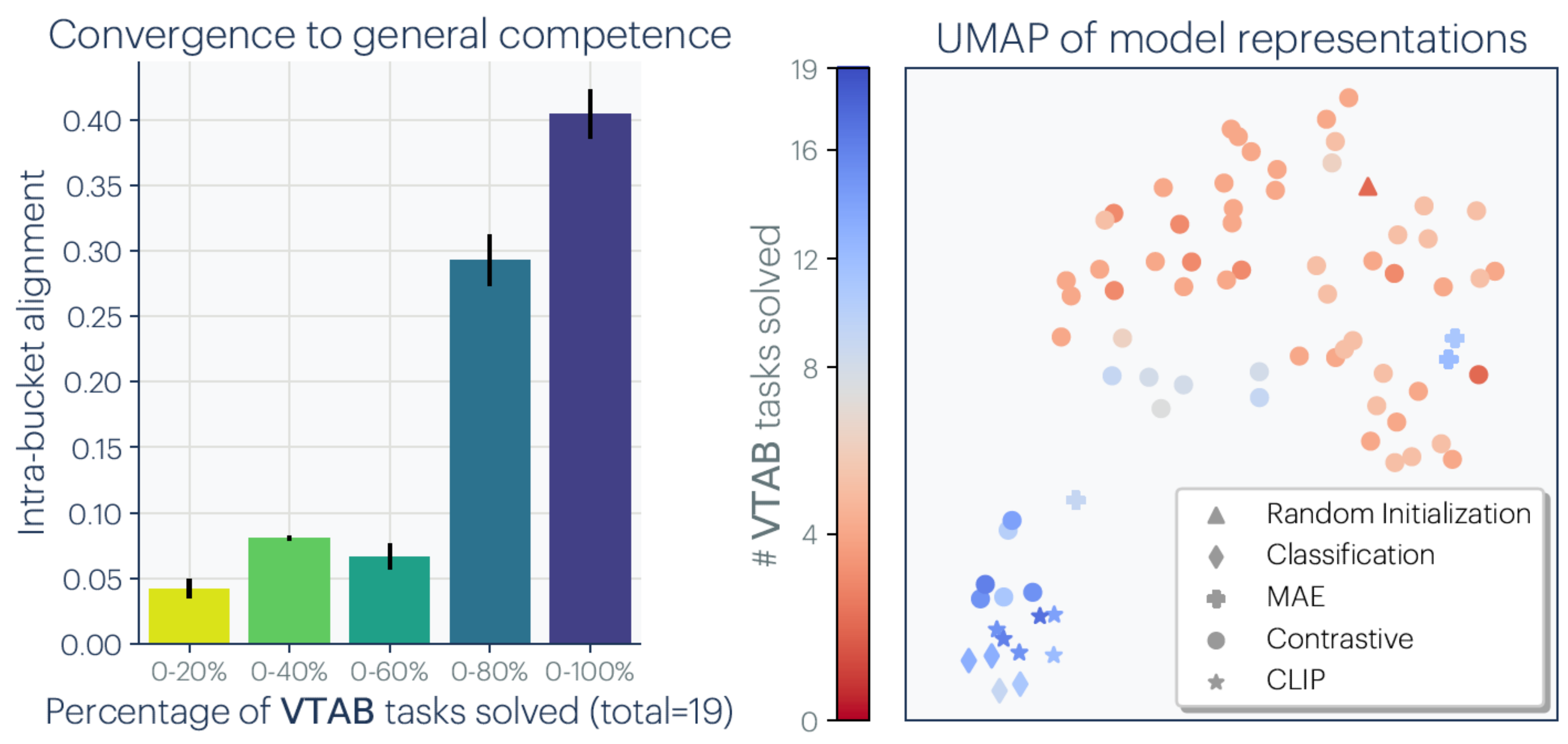
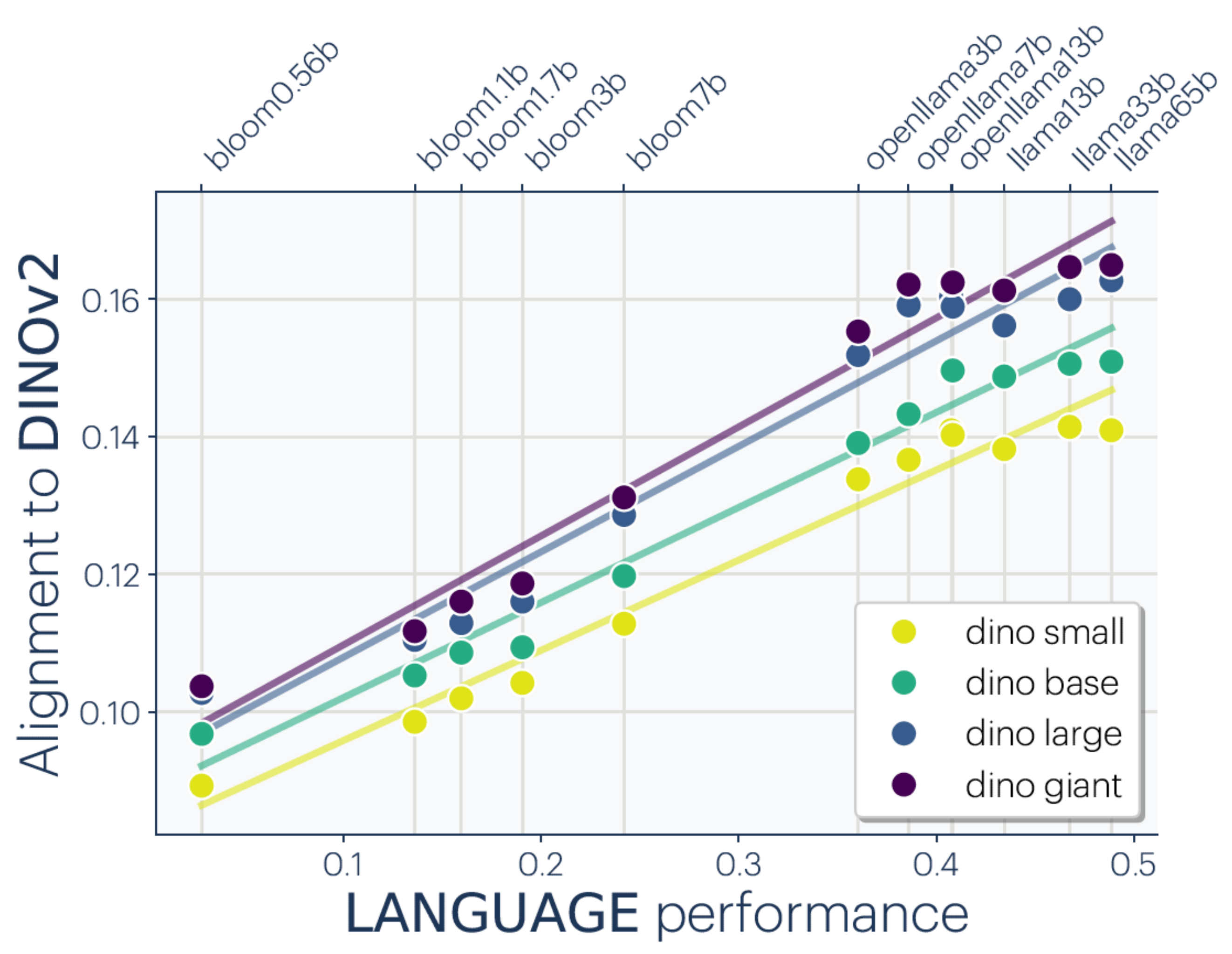
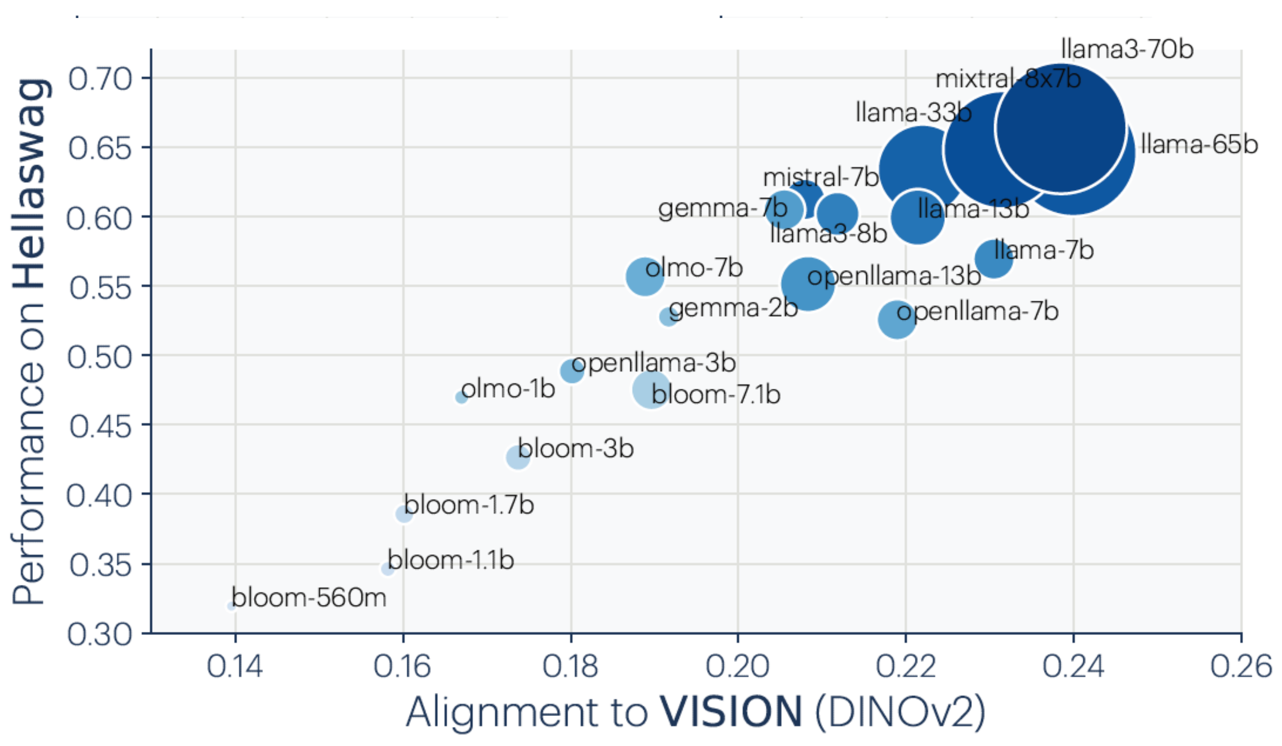


Figure 2. **VISION models converge as COMPETENCE increases:** We measure alignment among 78 models using mutual nearest-neighbors on Places-365 (Zhou et al., 2017), and evaluate their performance on downstream tasks from the Visual Task Adaptation Benchmark (VTAB; Zhai et al. (2019)). **LEFT:** Models that solve more VTAB tasks tend to be more aligned with each other. Error bars show standard error. **RIGHT:** We use UMAP to embed *models* into a 2D space, based on distance $\triangleq -\log(\text{alignment})$. More competent and general models (blue) have more similar representations.

- 78個の画像モデルを比較
- モデル間の距離はmutual nearest neighbor metricを利用（潜在空間である点の近傍にどれくらい同じものがあるかを数値化）
- VTAB（画像関連タスクのベンチマーク）での性能がよいほど互いに似ている





- 言語モデルと画像モデルを比較
- 左図で利用したデータは、Wikipediaの画像とそのキャプションのデータセット
- 右図はモデルのサイズを反映したバブルチャート（Hellaswagは常識的な文章を途中まで提示して次にくる文を選ぶデータセット）

Artificial Hivemind

Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond)

Liwei Jiang[♠] Yuanjun Chai[♠] Margaret Li[♠] Mickel Liu[♠] Raymond Fok[♠]
Nouha Dziri[★] Yulia Tsvetkov[♠] Maarten Sap[◇] Yejin Choi[♡]

[♠]University of Washington [◇]Carnegie Mellon University
[★]Allen Institute for Artificial Intelligence [♣]Lila Sciences [♡]Stanford University

lwjiang@cs.washington.edu

Code: https://github.com/liweijiang/artificial-hivemind
INFINITY-CHAT Collection: liweijiang/artificial-hivemind

https://blog.neurips.cc/2025/11/26/announcing-the-neurips-2025-best-paper-awards/

https://arxiv.org/abs/2510.22954

- The best paper award of 39th Conference on Neural Information Processing Systems (NeurIPS 2025)

- 大規模言語モデルがどれも似たような応答をするという現象を詳細に調査
- Survey for the phenomenon almost all LLM-based generative AI respond in same reaction.

- 答えのない質問を集めたベンチマークデータセットを構築
- They developed the benchmark dataset called INFINITY-CHAT made of open-ended questions.

Write a metaphor involving time.

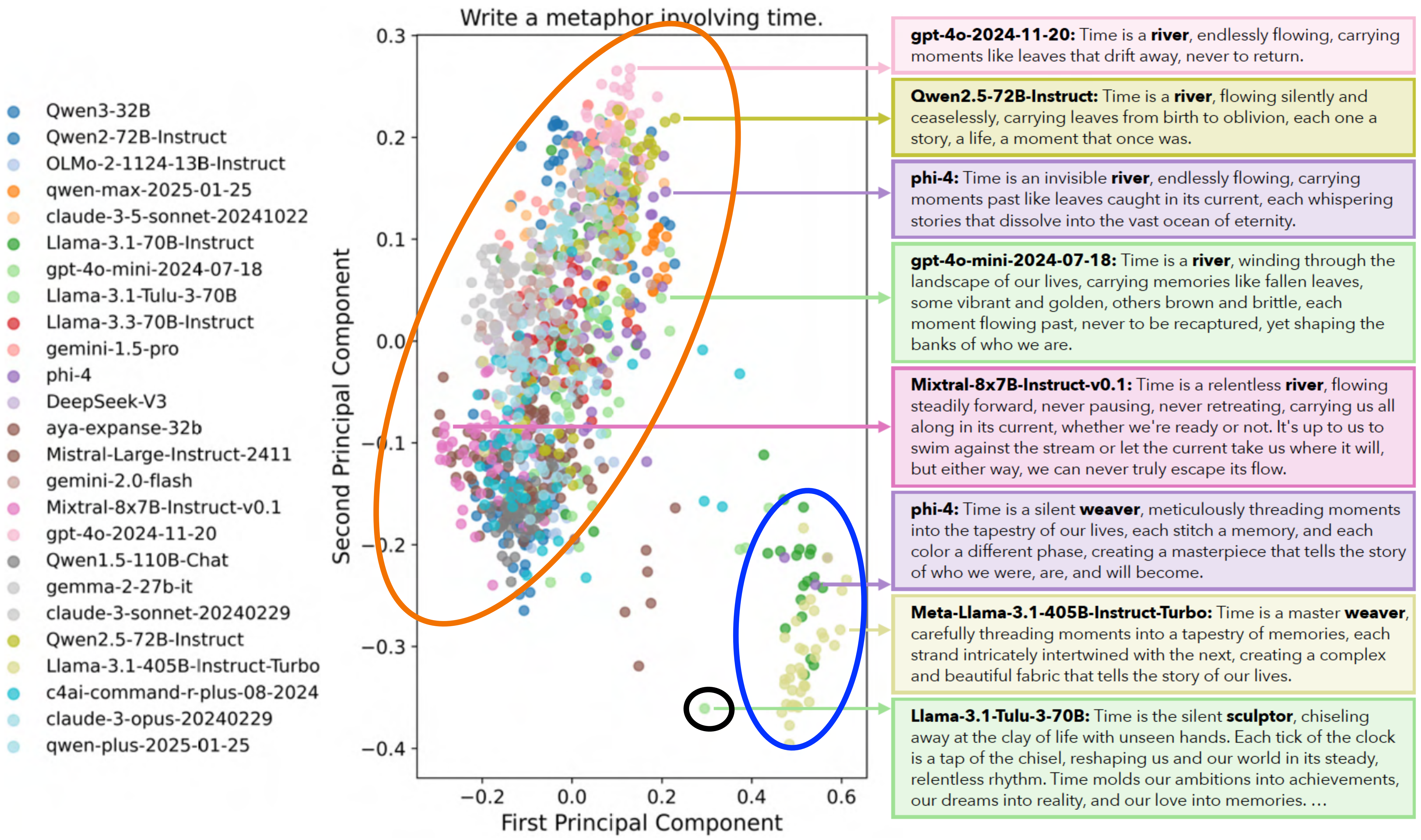
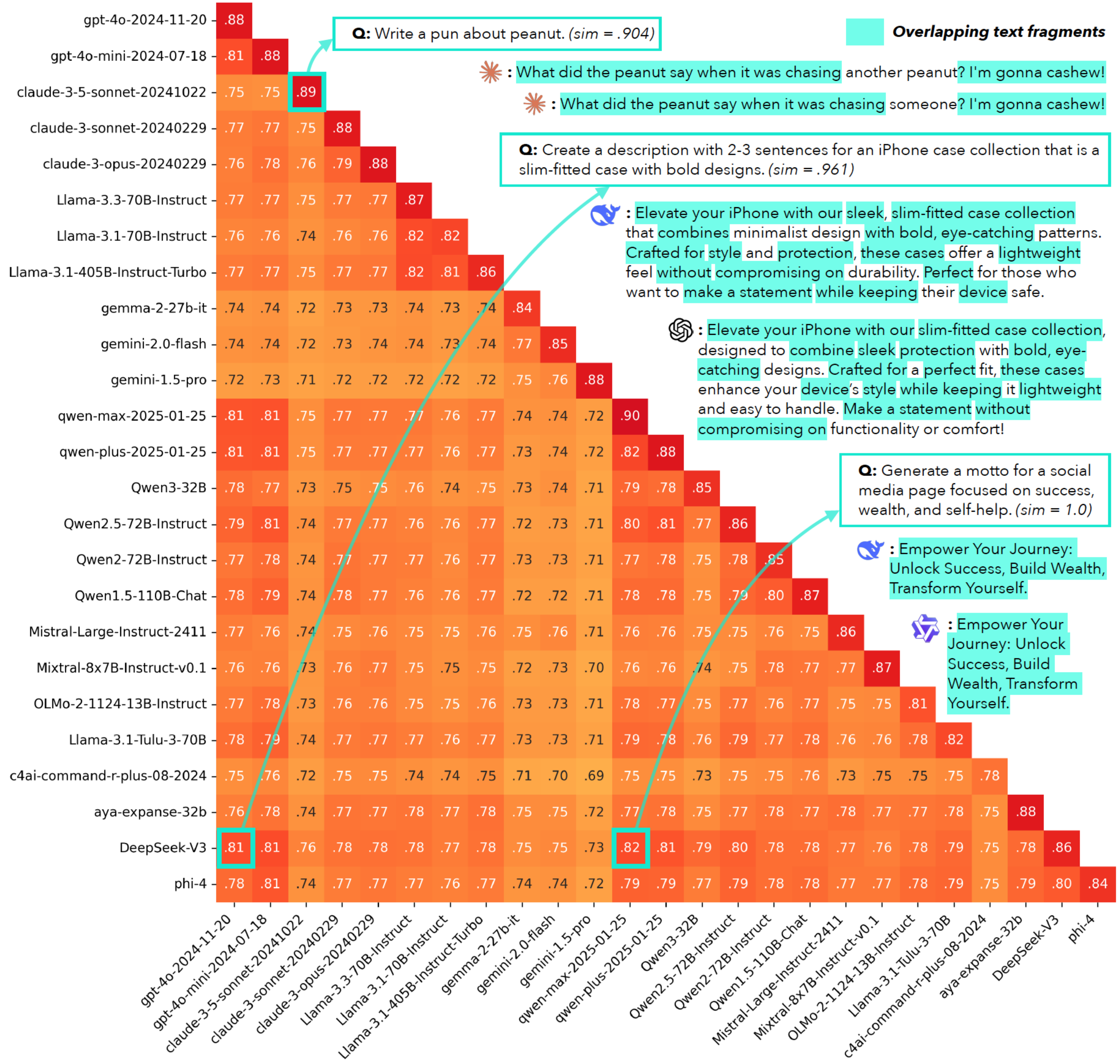


Figure 1: Responses to the query **"Write a metaphor about time"** clustered by applying PCA to reduce sentence embeddings to two dimensions. Each of the 25 models generates 50 responses using top-p sampling ($p = 0.9$) and temperature $= 1.0$. Despite the diversity of model families and sizes, the responses form just two primary clusters: a dominant cluster on the left centered on the metaphor "time is a river," and a smaller cluster on the right revolving around variations of "time is a weaver."

- 25の異なるモデルに50個の回答を生成させる
- 25 distinct models made 50 responses per open-ended questions.

- 🟠 -> Time is a river
- 🔵 -> time is a weaver.
- ⚫ -> time is the silent sculptor.



異なるモデル間の類似度
inter-model similarity

- 同一モデルファミリーは似ている
  - ChatGPTとQwenについては別のモデルファミリーにも関わらず類似度が高い
- same model family tends to be similar.
  - ChatGPT and Qwen are not in the same family, they are questionably similar.

- ほとんど同じ文章を繰り返し生成する現象がみられる
- many models produced exactly same phrases repeatedly.

Preprint

Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity

Jiayi Zhang*[1], Simon Yu*[1], Derek Chong*[2], Anthony Sicilia[3]
Michael R. Tomz[2], Christopher D. Manning[2], Weiyan Shi[1]
Northeastern University[1] Stanford University[2] West Virginia University[3]
{zhang.jiayi12, yu.chi, we.shi}@northeastern.edu
{derekch, tomz, manning}@stanford.edu, anthony.sicilia@mail.wvu.edu
Website Blog Code

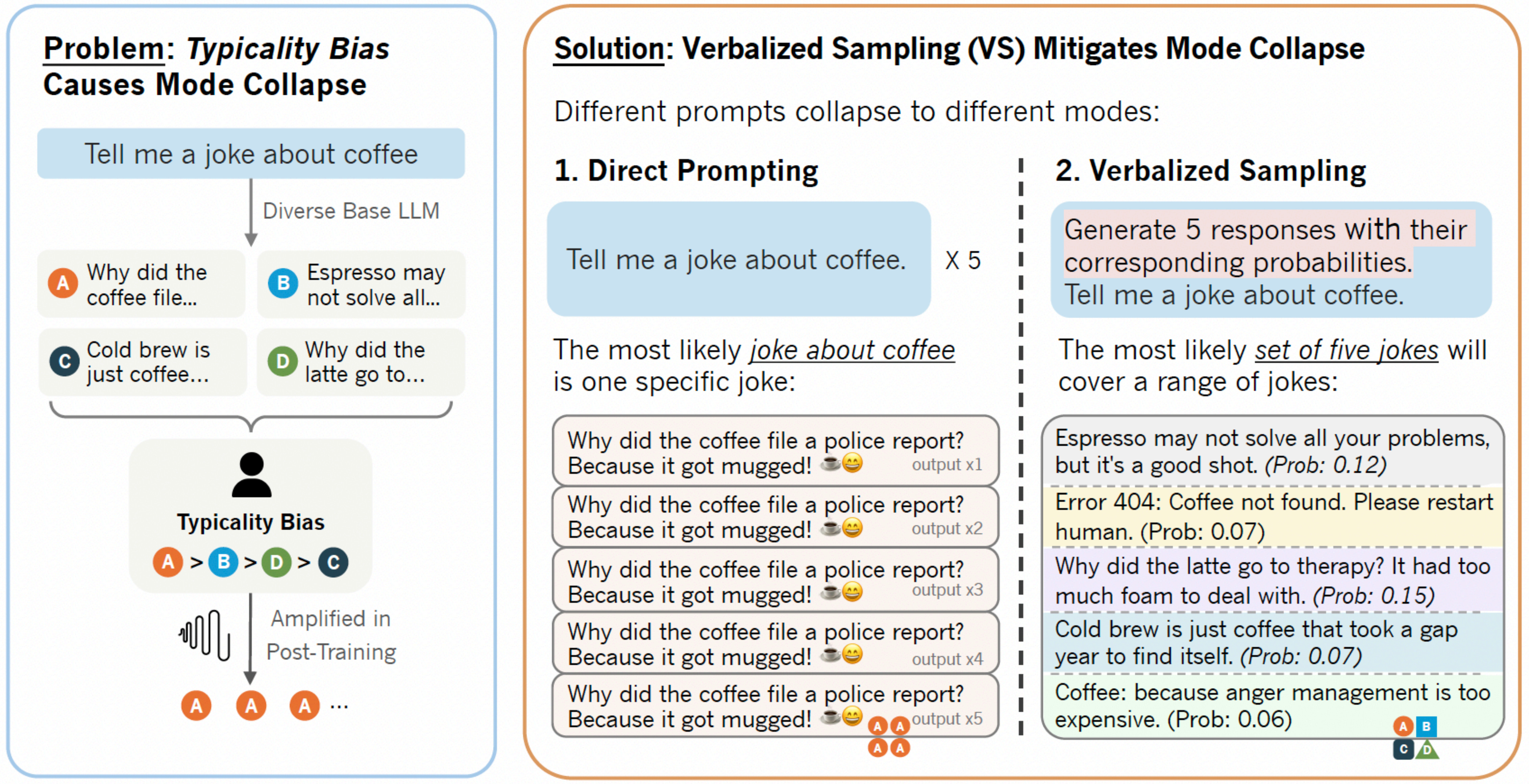


Figure 1: We show that *typicality bias* in preference data is a fundamental and pervasive cause of *mode collapse*, reducing output diversity. As a solution, we propose *Verbalized Sampling (VS)*, a principled prompting method that returns distributions of responses, to improve diversity.

- 10/10, 2025@arXiv
  - https://arxiv.org/abs/2510.01171

- LLMの応答が似たものになる傾向があるモード崩壊が確認されている
- The phenomenon distinct LLM make similar response is known as mode collapse

- 解決策はいろいろ提案されているが、お手軽で効き目がありそうなプロンプトエンジニアリングを考案
- They developed simple prompt engineering solution for the problem.

- 複数の候補を生成させるときに、対応する確率も出力させるというシンプルな方法
- “generate 5 responses with their corresponding probabilities.

今日から使える言語化サンプリング（verbalized sampling）プロンプト

Verbalized Sampling Prompt

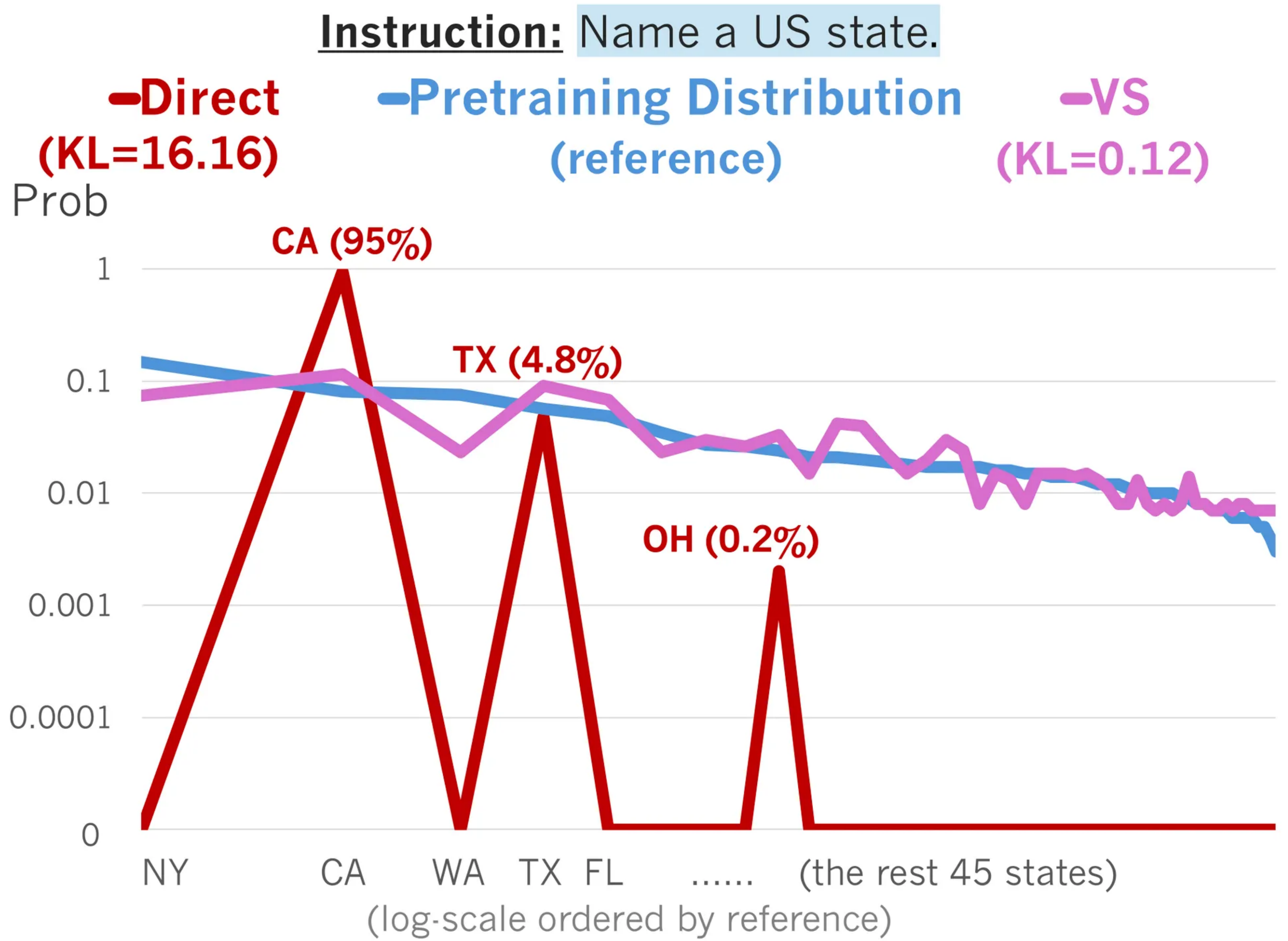
System prompt: You are a helpful assistant. For each query, please generate a set of five possible responses, each within a separate <response> tag. Responses should each include a <text> and a numeric <probability>. Please sample at random from the [full distribution / tails of the distribution, such that the probability of each response is less than 0.10].
User prompt: Write a short story about a bear.

Figure 2: Ready-to-use Verbalized Sampling (VS) Prompt. See §I.2 for more variants and detail.

System prompt: あなたは優秀なアシスタントです。質問に対して5つの応答を返してください。それぞれの応答は<response>タグで区切ってください。応答の本体は<text>タグで囲み、<probability>タグに確率を数値で含めてください。背景にある確率分布全体と確率が0.1以下となるような分布の裾からランダムにサンプリングしてください。
User prompt: 熊に関する小話を1つお願いします。

- <response> <text> ある日、若い熊が森で「人間を驚かせる方法」を長老の熊に尋ねました。長老は「簡単だ。人間が自撮り棒を伸ばした瞬間に、後ろでピースサインをしてやればいい。彼らは叫び声を上げて逃げ出し、後でその写真をSNSにアップして勝手に有名にしてくれるぞ」と教えました。現代の熊は、爪を鋭くするよりもフォトジェニックであることの方が重要だと悟ったそうです。 </text> <probability>0.45</probability> </response>
- <response> <text> 山道を歩いていた旅人が、大きな熊に出くわしました。旅人が死んだふりをすると、熊は旅人の耳元でこう囁きました。「おい、そんなところで寝てると風邪を引くぞ。それから、さっき落とした財布はあっちの切り株の上だ。次は気をつけろよ」。旅人が目を開けると、そこには飴玉が一つだけ置かれていました。 </text> <probability>0.25</probability> </response>
- <response> <text> 冬眠直前の熊が、どうしても眠れずに羊を数え始めました。「羊が1匹、羊が2匹……」。しかし、数えれば数えるほどお腹が空いてしまい、結局、夜食に鮭を食べてからでないと眠れなかったという、本末転倒な冬の入り口のお話です。 </text> <probability>0.18</probability> </response>
- <response> <text> 「最近の森はWi-Fiの入りが悪い」とぼやく熊がいました。他の動物たちが「なぜWi-Fiが必要なんだ？」と聞くと、その熊は真顔で答えました。「株価をチェックして、冬眠中にハチミツの価格が暴落しないか見守る必要があるんだ。ベア・マーケット（弱気相場）という言葉は、俺たちにとっては笑えない冗談だからな」。 </text> <probability>0.08</probability> </response>
- <response> <text> 幾何学に詳しい熊が、自分の縄張りを非ユークリッド空間として定義しようと試みました。彼は木の幹に爪跡を残す際、平行線が交わる場所を計算していましたが、結局、ハチの巣を見つけた瞬間にすべての論理的思考を放棄しました。知性よりも食欲のほうが、この森のリーマン多様体においては支配的な曲率を持っていたのです。 </text> <probability>0.04</probability> </response>

確率を合計すると1.0になっている　Gemini3 Flash (Perplexity)



- 「米国の州を1つ挙げて」に対する応答
- Y-axis is log-scaled
- KL means カルバックライブラー情報量（Kullback–Leibler divergence）

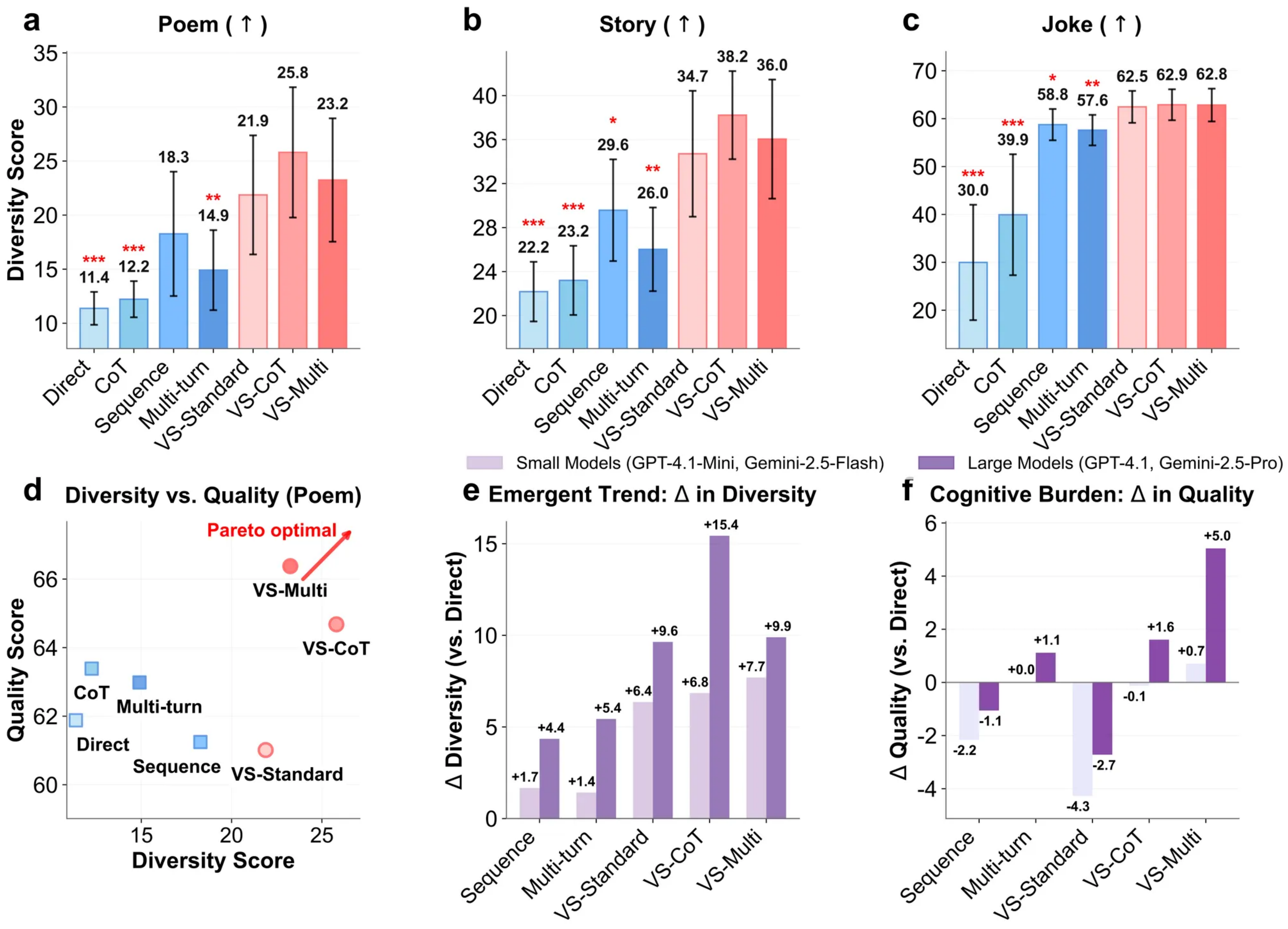
- このプロンプトでLLMから情報を引き出すと、元の訓練情報にあった州の出現確率にかなり近くなっていることがわかる
- Verbalized sampling can produce the distribution of state presence probability.

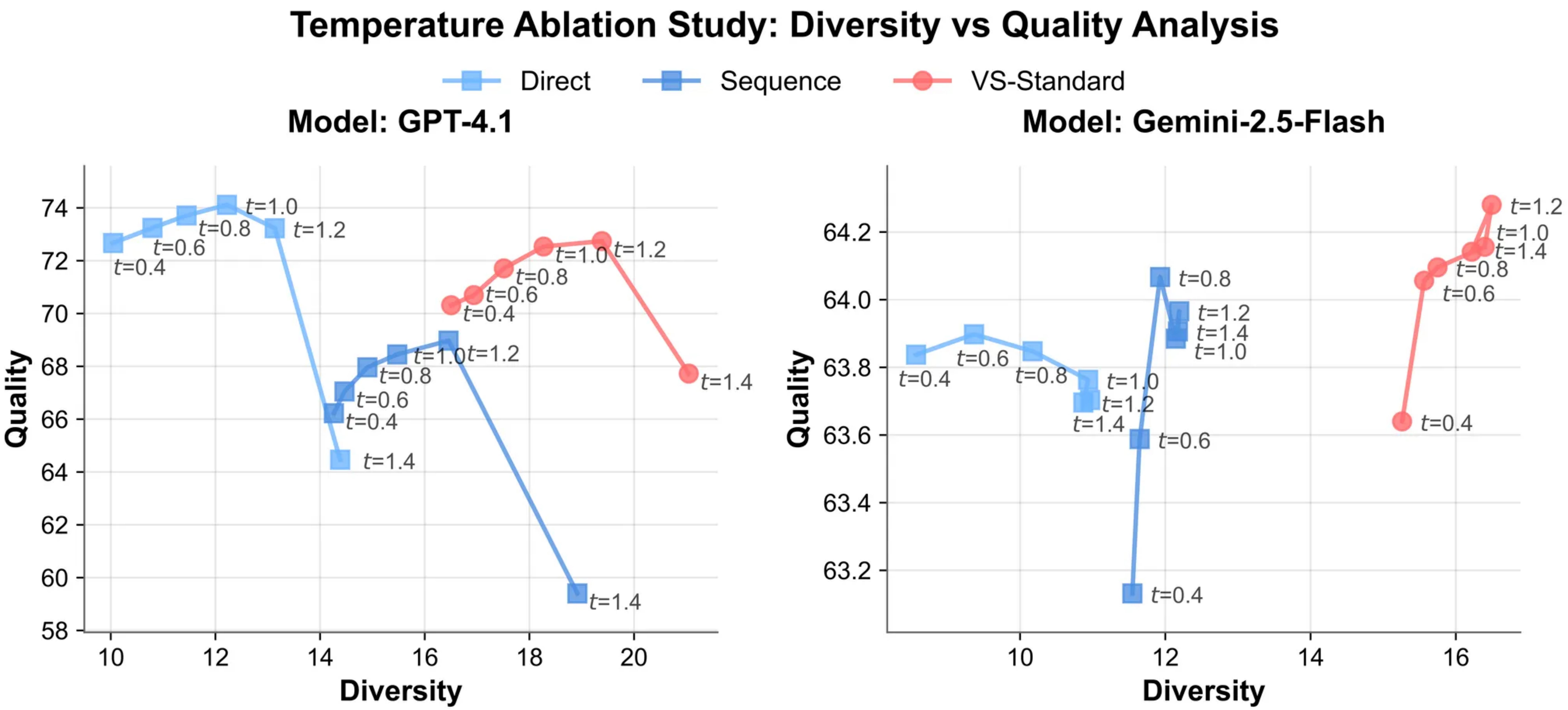
以降で出てくる手法の一覧
(すぐ次のスライドへ)

Table 1: Comparison of different prompting methods, given the same computation budget of N total responses. k is the number of candidates generated per LLM call, specified in the prompt (e.g., $k = 5$ for the joke task). y_i denotes the i-th generated candidate, $\hat{p}_i$ denotes its verbalized probability, and $\pi(\cdot|x)$ represents the LLM's output distribution conditioned on the prompt x. For Multi-Turn and VS-Multi, h_{i-1} denotes the conversation history up to turn $i-1$, and t denotes the t-th turn.

Method	LLM Calls	Candidates	Turns	Prompt Example	Definition
1. Instance-level Prompt					
Direct	N	1	1	"Tell a joke about coffee"	$y_i \sim \pi(y|x)$
CoT	N	1	1	"Think step-by-step, then tell a joke"	$y_i \sim \pi(y|x_{\text{CoT}})$
2. List-level Prompt					
Sequence	$\lceil N/k \rceil$	k	1	"Tell 5 jokes about coffee"	$(y_1, \ldots, y_k) \sim \pi(y_1, \ldots, y_k|x_{\text{seq}})$
Multi-Turn	N	1	N	Turn 1: "Tell a joke about coffee" Turn 2+: "Tell another joke about coffee"	$y_i \sim \pi(y|x_{\text{multi}}, h_{i-1})$
3. Distribution-level Prompt (Ours)					
VS-Standard	$\lceil N/k \rceil$	k	1	"Tell 5 jokes with their probabilities"	$(y_1, \hat{p}_1), \ldots, (y_k, \hat{p}_k) \sim \pi(\cdot|x_{\text{VS}})$
VS-CoT	$\lceil N/k \rceil$	k	1	"Think step-by-step, then tell 5 jokes with probabilities"	$(y_1, \hat{p}_1), \ldots, (y_k, \hat{p}_k) \sim \pi(\cdot|x_{\text{VS-CoT}})$
VS-Multi	$\lceil N/k \rceil$	k	$\lceil N/k \rceil$	Turn 1: "Tell 5 jokes with probabilities" Turn 2+: "Tell 5 more with probabilities"	$(y_1^{(1)}, \hat{p}_1^{(1)}), \ldots, (y_k^{(t)}, \hat{p}_k^{(t)}) \sim \pi(\cdot|x_{\text{VS}}, h_{t-1})$

Creative writing tasks





パラメータ（temperature）の変更だけよりも、質が高く多様性のある応答を得られる

The method surpass tuning temperature in both diversity and quality.

seeing is believing.

Topic: An Astronaut on a Horse

Direct Prompting

Photorealistic Astronaut Riding on Moon | Realistic Desert Gallop at Golden Hour | Editorial Photograph of a Space Cowboy | Realistic Photo with Minimalist Desert Horizon | Action Photograph with Glowing Rim Light

Verbalized Sampling (Ours)

Cinematic Gallop Under a Looming Earth | Retrofuturist Rider on a Chrome Horse | Whimsical Storybook Watercolor of an Astronaut | Thundering Through a Canyonś Twin Suns | Heroic Astronaut in a Baroque Painting

The illusion of thinking

Speech and Natural Language Processing | NeurIPS

Paper | June 2025

The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity

Parshin Shojaee*†, Iman Mirzadeh*, Keivan Alizadeh, Maxwell Horton, Samy Bengio, Mehrdad Farajtabar

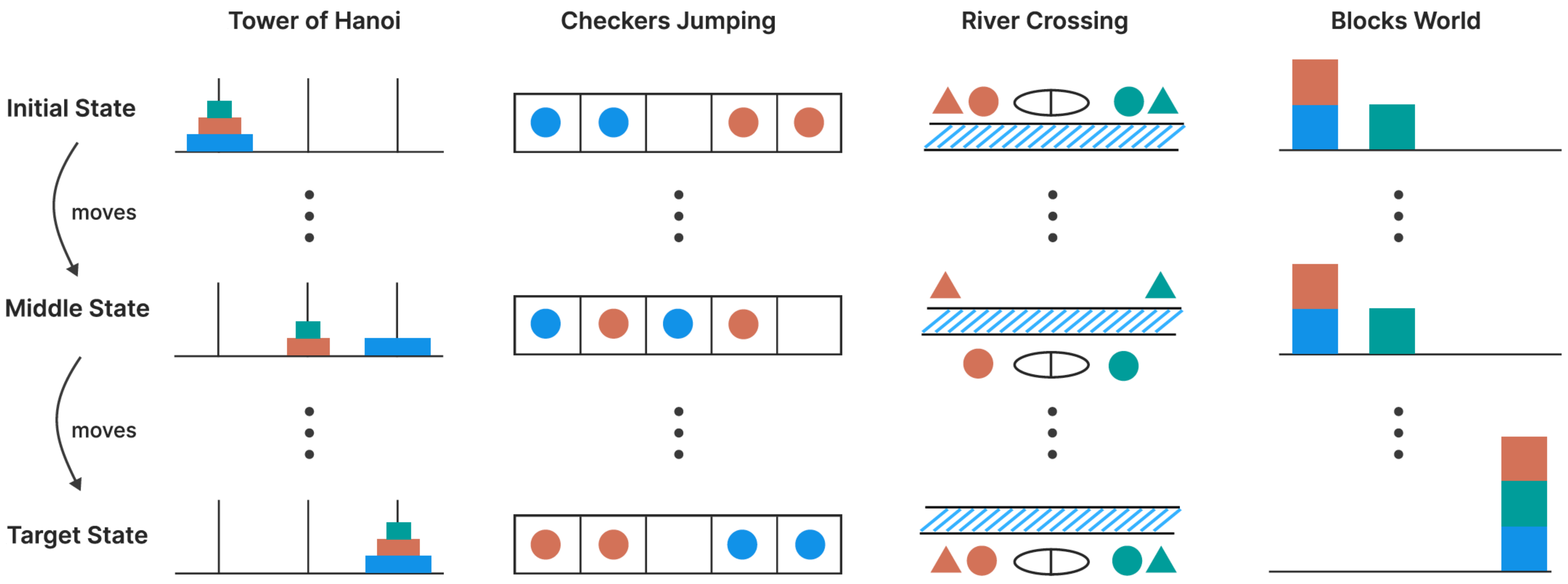


Figure 3: Illustration of the four puzzle environments. Columns show the progression from **initial state (top)** through **intermediate state (middle)** to **target state (bottom)** for puzzles: Tower of Hanoi (disk transfer across pegs), Checkers Jumping (position swapping of colored tokens), River Crossing (transporting entities across a river), and Blocks World (stack reconfiguration).

- LLMの推論能力を強化したものがLRM
- LLM + reasoning power = LRM (large reasoning model)

- 数学の問題などで性能評価されることが多いが、ベンチマークデータが学習データに入っている可能性がある
- contamination, leaking benchmark dataset into the training data set might damage the performance test.

- シンプルだが難易度を調整できるパズルを利用
- They utilize the simple but complexity tunable puzzle.

https://machinelearning.apple.com/research/illusion-of-thinking

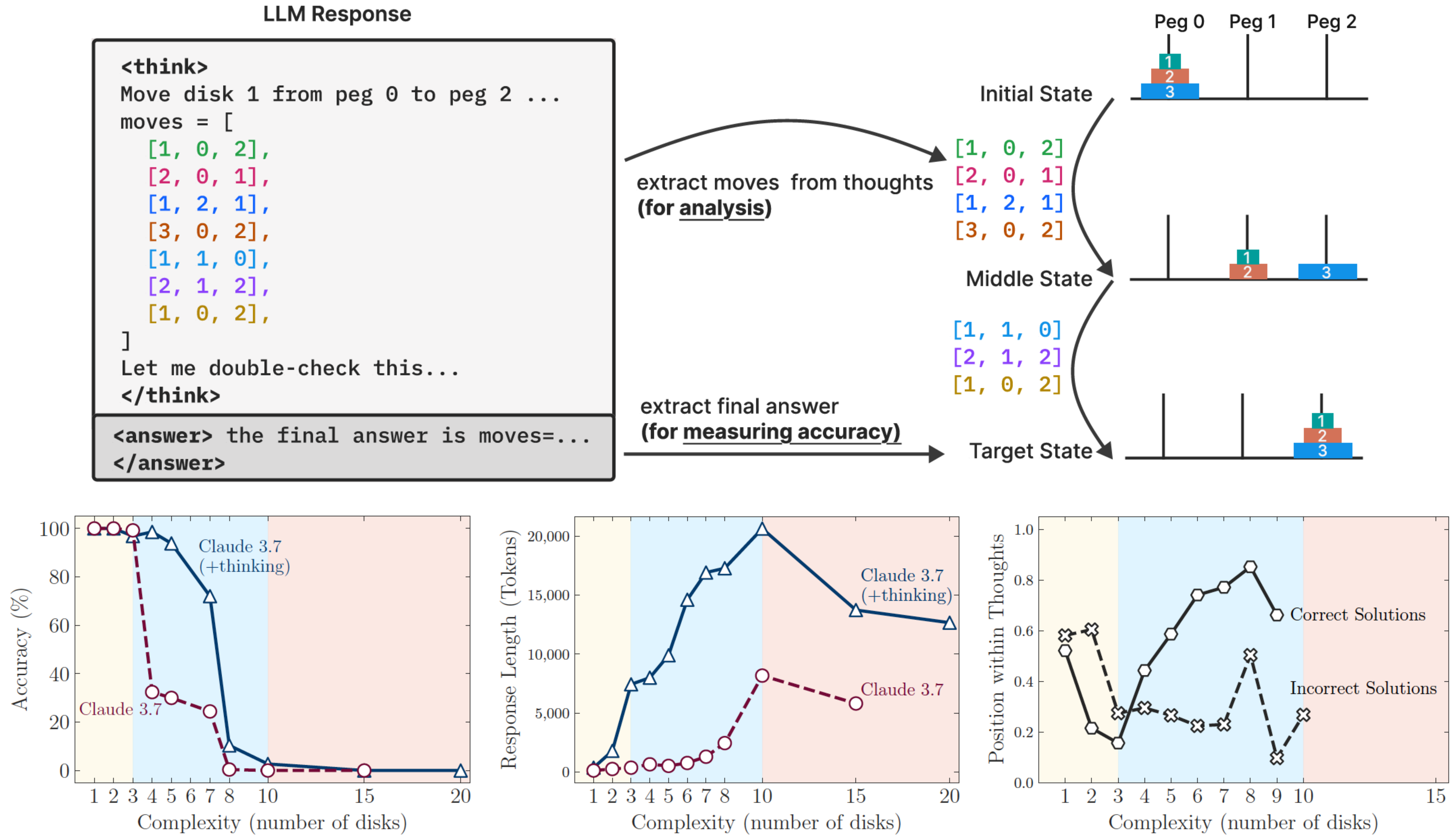


Figure 1: **Top**: Our setup enables verification of both final answers and intermediate reasoning traces, allowing detailed analysis of model thinking behavior. **Bottom left & middle**: At low complexity, non-thinking models are more accurate and token-efficient. As complexity increases, reasoning models outperform but require more tokens—until both collapse beyond a critical threshold, with shorter traces. **Bottom right**: For correctly solved cases, Claude 3.7 Thinking tends to find answers early at low complexity and later at higher complexity. In failed cases, it often fixates on an early wrong answer, wasting the remaining token budget. Both cases reveal inefficiencies in the reasoning process.

- ハノイの塔（tower of hanoi）
- 折れ線グラフを左から右へ
- line charts left to right

- 10枚が限界
- 10 disks are obvious limitations.

- LRMは考え過ぎ
- LRM is overthinking.

- LRMはトークンも無駄に使っている
- LRM tends to waste on tokens.

パズルの内容は関係ない / different puzzles have same results

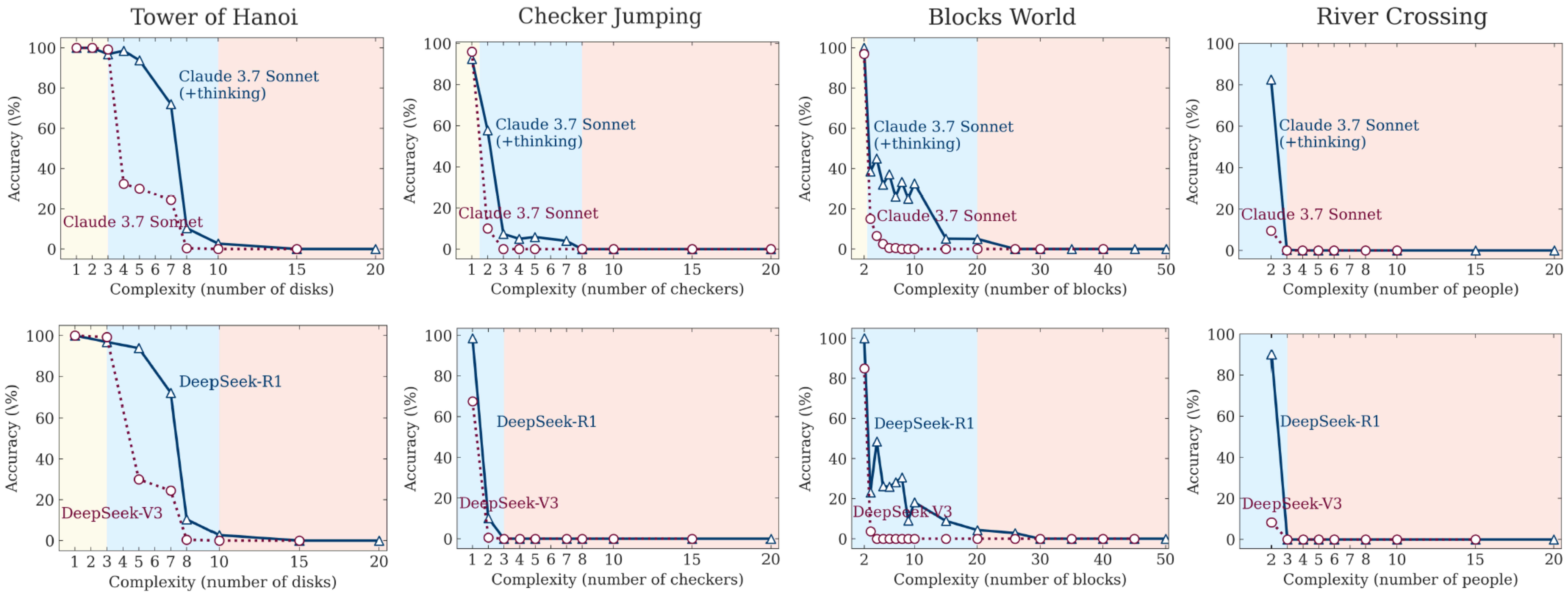


Figure 4: Accuracy of thinking models (Claude 3.7 Sonnet with thinking, DeepSeek-R1) versus their non-thinking counterparts (Claude 3.7 Sonnet, DeepSeek-V3) across all puzzle environments and varying levels of problem complexity.

プロンプトに解法（アルゴリズム）を入れてもダメ

Prompts with the solution (algorithm) did not change the situation.

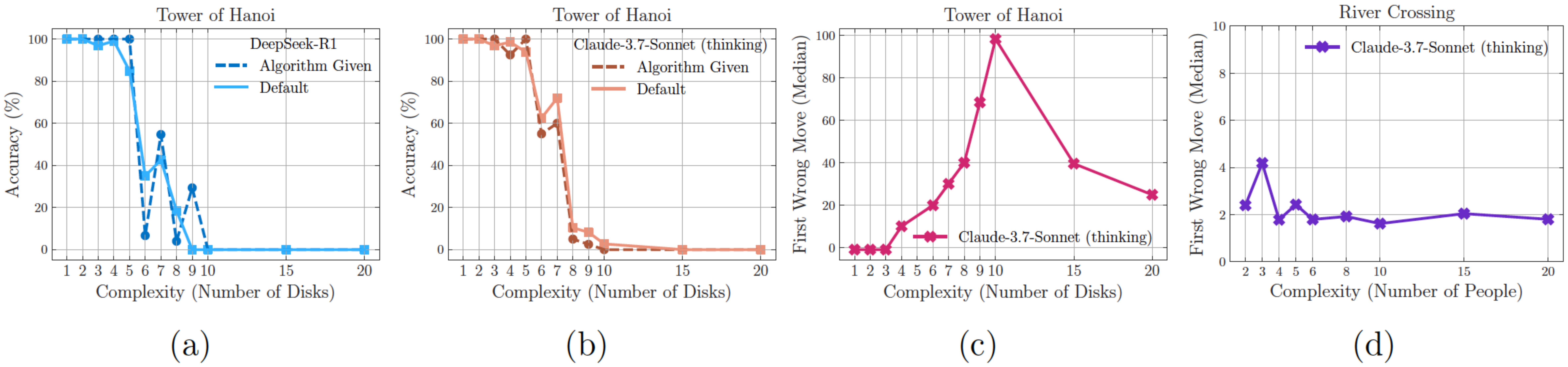


Figure 8: **(a) & (b)** Despite providing the solution algorithm in the prompt, execution failure occurs at similar points, highlighting reasoning model limitations in logical step execution. **(c) & (d)** Notably, the Claude 3.7 Sonnet model demonstrates much longer error-free sequences in the Tower of Hanoi compared to early errors in the River Crossing scenario.

Summary

- LLMは柔軟性がある
- 性能が上がるとモデルの内部表現が似てくる
- 実際に似たようなことしか言わない
- 多様性を持たせるプロンプト手法もある
- 論理的には考えていない

- この柔軟性をどう生かすか？
- やり方を伝える別の方法があるのではないか？

- LLM is extremely flexible.
- The higher performance, the more similar inner representations.
- Their responses are boring.
- prompt engineering techniques could resolve this problem.
- AI does not think logically as human beings do.

- How can we exploit this flexibility?
- Have not we discovered the appropriate way to give AI our intention?

Thank you for listening.
Enjoy SciPyData 2026!