ICLR: In-Context Learning of Representations

2025.1.17 DMQA Open Seminar

장건희

Introduction

장건희
Geonhui Jang

학력

고려대학교 공과대학 산업경영공학부 (2018.03-2023.02)

고려대학교 공과대학 산업경영공학과 (2023.03-2025.02)

- 석사과정
- 데이터마이닝 및 품질애널리틱스 (DMQA) 연구실
- 지도 교수: 김성범 교수님

관심 분야 Computer Vision, Diffusion Models, LLMs

In-context Learning of LLMs

❖ **In-context Learning (ICL)**

- Pretrained LLM이 **추가적인 파라미터 업데이트 없이 주어진 맥락(context)을 기반으로 task를 수행**하는 능력
- 사전에 학습한 prior를 바탕으로 주어진 context의 패턴을 파악하며 응답을 생성하는 과정
- Context가 자연어로 구성되어 있어 LLM에게 지식을 전달하기 용이 & 인간이 유추를 통해 의사결정하는 것과 유사

입력 (w/o context)

서울의 관광지를 추천해줘.

응답

서울에는 명소가 많습니다.
경복궁을 방문해보세요.

입력 (w/ context)

다음 예시를 참고해서 서울의 관광지를 추천해줘.

파리: 에펠탑 - 파리를 대표하는 랜드마크로, 아름다운 야경과 함께 즐길 수 있습니다.
루브르 박물관 - 세계적인 예술 작품들을 감상할 수 있는 곳입니다.

서울:

응답

남산서울타워 - 서울의 멋진 전망을 감상할 수 있는 랜드마크입니다.
경복궁 - 서울의 전통적인 궁궐로, 한국의 역사와 문화를 체험할 수 있습니다.

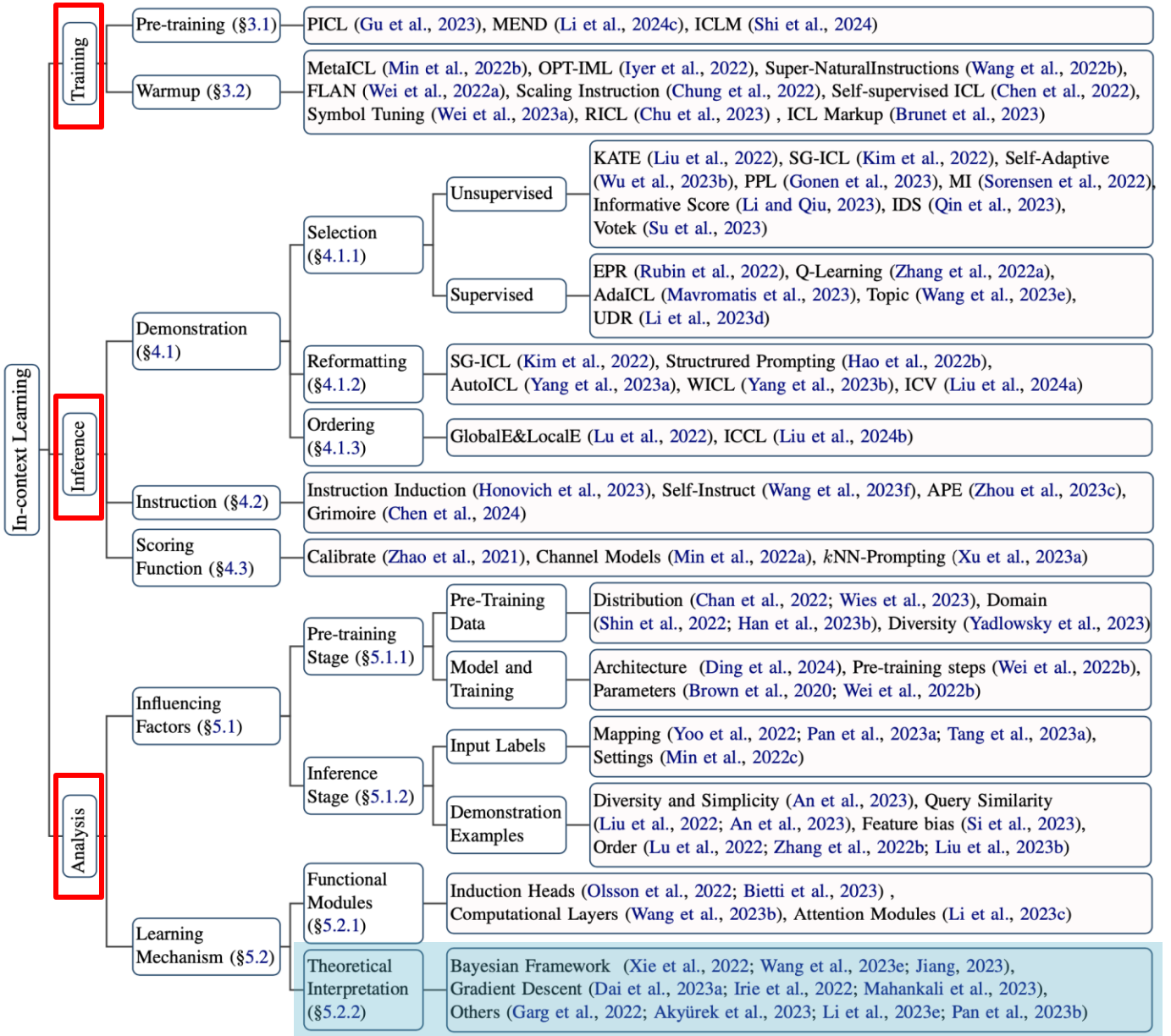
<In-context Learning 예시>

Dong, Qingxiu, et al. "A survey on in-context learning." *arXiv preprint arXiv:2301.00234* (2022).

In-context Learning of LLMs

❖ **In-context Learning (ICL)**

- 다양한 관점에서 동작 원리 분석
  - ✓ **Training**: 어떻게 학습해야 ICL을 더 잘 할까?
  - ✓ **Inference**: 추론 시 어떻게 prompt를 구성해야 ICL을 더 잘 할까?
  - ✓ **Analysis**: 무엇이 ICL을 동작하게 할까?



Dong, Qingxiu, et al. "A survey on in-context learning." *arXiv preprint arXiv:2301.00234* (2022).

In-context Learning of LLMs

❖ In-context Learning (ICL)

- 다양한 관점에서 동작 원리 분석
  - ✓ **Training**: 어떻게 학습해야 ICL을 더 잘 할까?
  - ✓ **Inference**: 추론 시 어떻게 prompt를 구성해야 ICL을 더 잘 할까?
  - ✓ **Analysis**: 무엇이 ICL을 동작하게 할까?

Theoretical Interpretation (§5.2.2)	Bayesian Framework (Xie et al., 2022; Wang et al., 2023e; Jiang, 2023), Gradient Descent (Dai et al., 2023a; Irie et al., 2022; Mahankali et al., 2023), Others (Garg et al., 2022; Akyürek et al., 2023; Li et al., 2023e; Pan et al., 2023b)

- Bayesian Framework: ICL을 Bayesian inference로 해석
- Gradient Descent: ICL이 gradient descent로서 동작함을 보임
- Others: 다양한 task에서 ICL이 어떻게 성능을 향상시키는지 분석

사전학습에서 봐왔던 개념이 아닌 새로 정의된 개념을 context로 입력하는 경우 모델은 이를 어떻게 해석할까?

Dong, Qingxiu, et al. "A survey on in-context learning." *arXiv preprint arXiv:2301.00234* (2022).

ICLR: In-context Learning of Representations

❖ 문제 정의

- 예를 들어 새로 출시된 제품 이름이 "strawberry"라고 LLM에게 설명하는 경우 LLM은 이 context를 기반으로 "strawberry"의 **representations**을 기존의 <과일>이 아닌 <제품>으로 적절히 바꿔야 한다. LLM은 이러한 능력을 갖추고 있는가?
- 즉, 사전학습에서 봐왔던 의미가 아닌 새로운 의미의 context가 주어지면 모델은 이들의 representations을 context에 따라 적절히 재구성할 수 있는가?

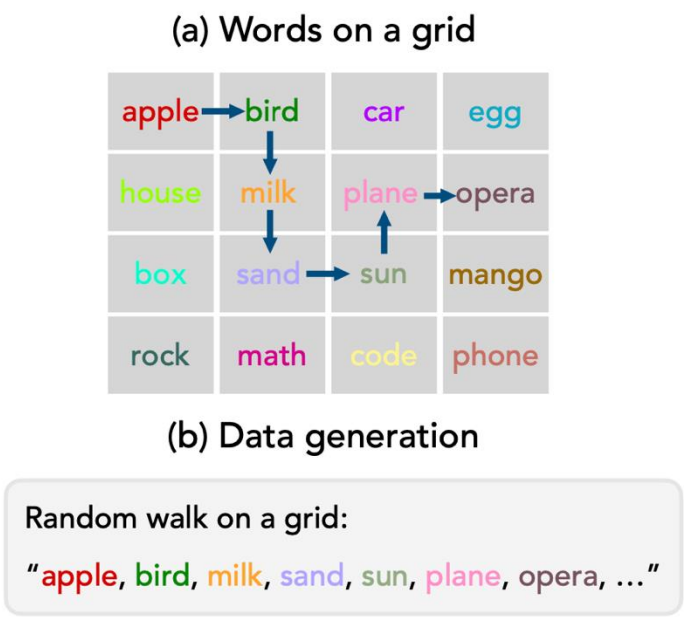
※ **Representations** = 모델 내부 token features = 모델이 단어를 어떻게 인식하는지 나타냄

Strawberry

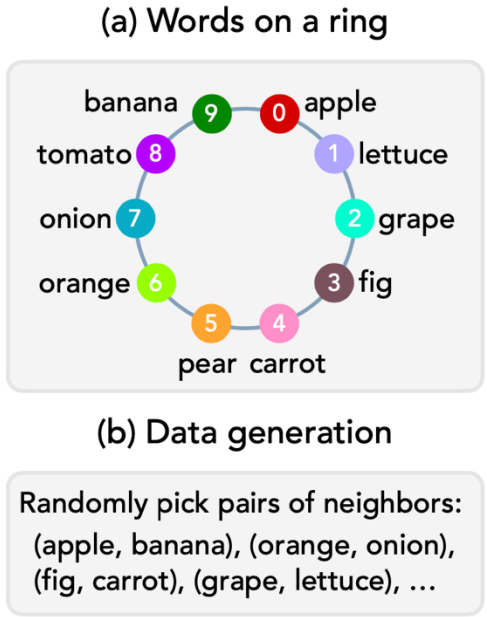
ICLR: In-context Learning of Representations

❖ 분석 방법

- **Graph tracing** 작업을 정의 (square grid, ring, hexagonal grid)
  1. 일반적인 다양한 단어들 선별 후 그래프에 무작위 배치
  2. 모델에 입력할 무작위 단어 시퀀스 context 제작
  3. 이 context가 주어졌을 때 **모델 representations이 해당 그래프 구조를 반영**하는지 분석

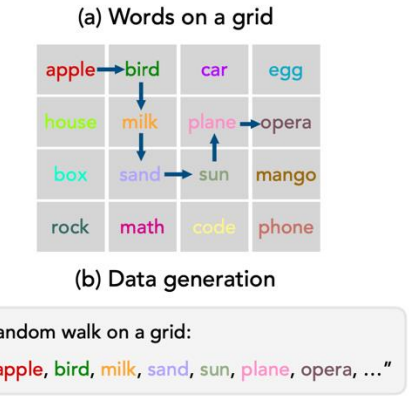


<Square Grid Graph 시퀀스 제작 예시>



<Ring Graph 시퀀스 제작 예시>

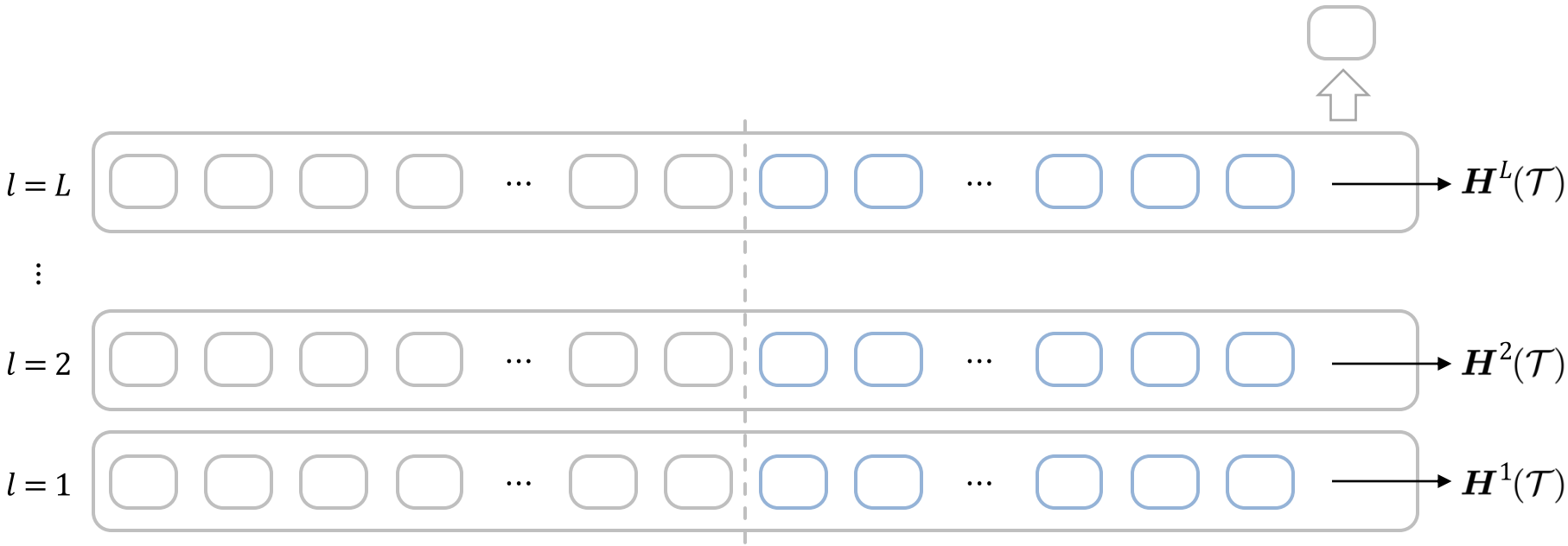
ICLR: In-context Learning of Representations



❖ **Representations 계산**

- Tokens (nodes) $\mathcal{T} = \{\overset{1}{\text{apple}}, \overset{2}{\text{bird}}, \overset{3}{\text{car}}, \cdots, \overset{n-1}{\text{code}}, \overset{n}{\text{phone}}\}$

(Next token prediction)



Context: apple → bird → milk → sand → ⋯ → math → rock → box → sand → ⋯ → plane → car → egg

이전 N_w(=50)개 length의 토큰별 평균 activation 계산

Stack of mean token representations:

$$\boldsymbol{H}^{\ell}(\mathcal{T}) \in \mathbb{R}^{n \times d}$$

$\boldsymbol{H}^{\ell}(\mathcal{T})$ 에 PCA를 적용해

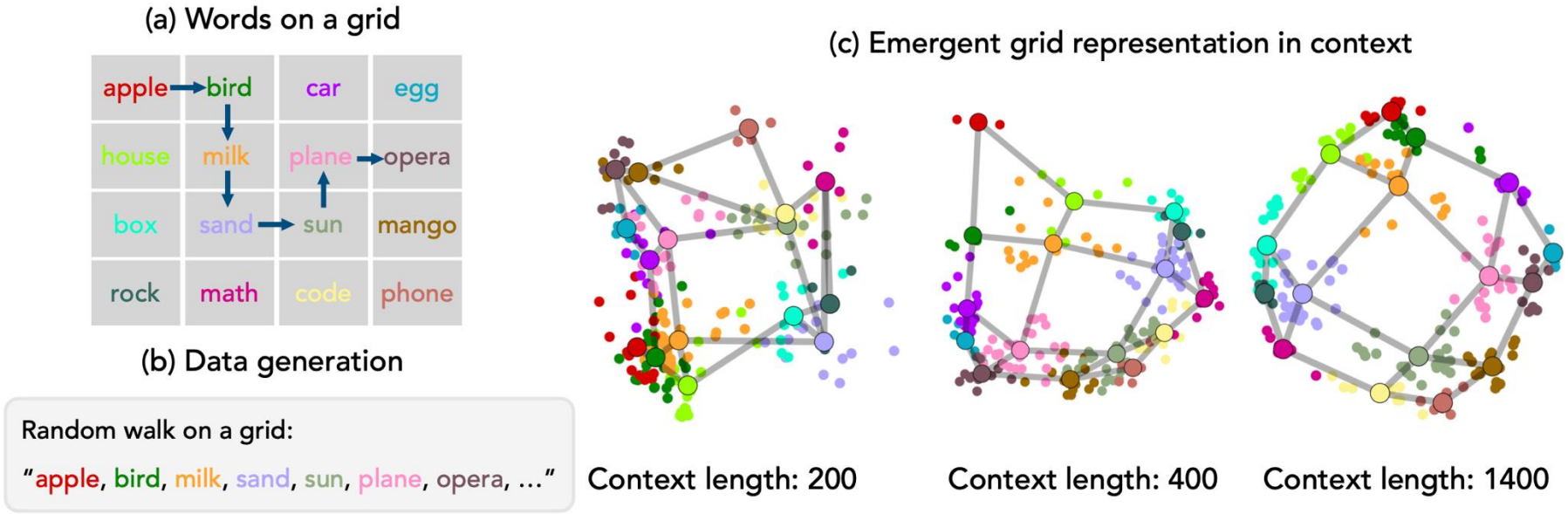
첫 두 개의 주성분으로 representations 시각

Results

❖ Square Grid Graph

- PCA 2차원 시각화 결과, context에 담긴 그래프 구조에 따라 representations이 형성됨
- **Context length가 길어질수록 representations의 형태가 격자 구조를 더 잘 반영**
- 중심이 팽창되어 보이는 것은 중앙 2×2 노드들이 random walk에 의해 더 많이 방문되기 때문

※ $\boldsymbol{H}^{26}(\mathcal{T})$ (26번째 레이어) 시각화

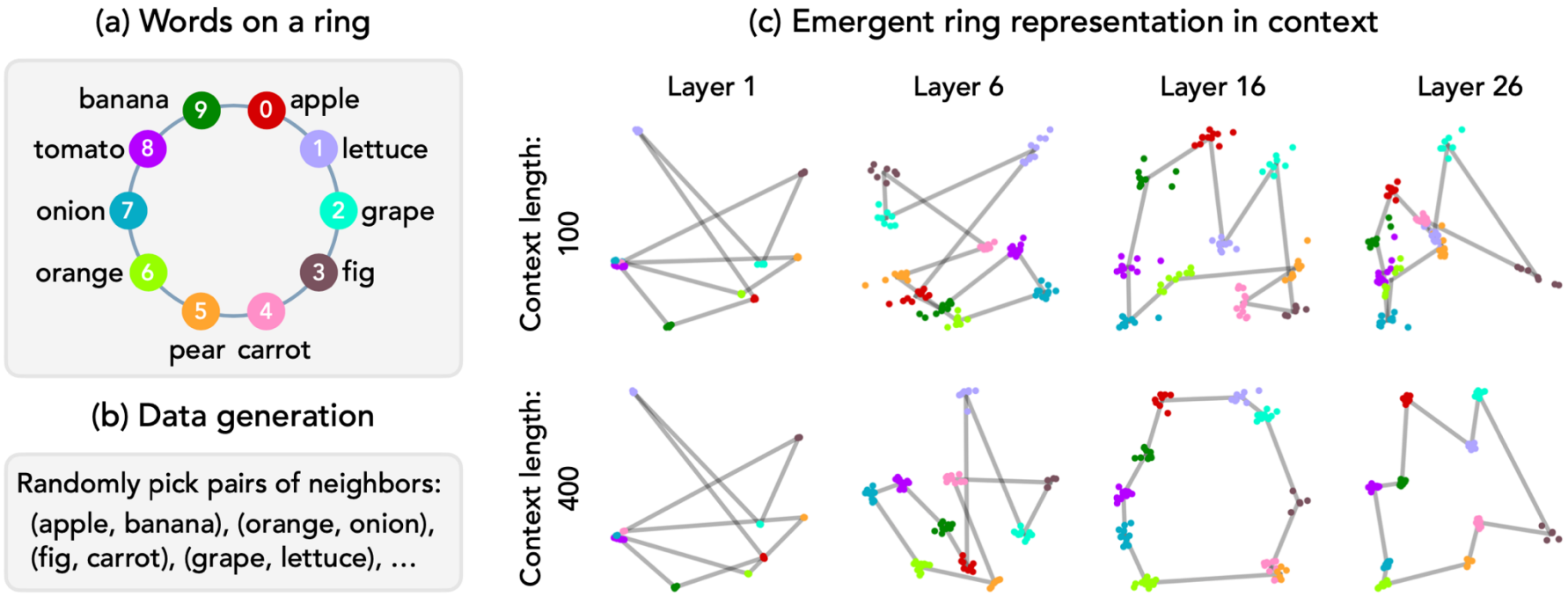


<Square Grid Graph Visualization>

Results

❖ Ring Graph

- PCA 2차원 시각화 결과, context에 담긴 그래프 구조에 따라 representations이 형성됨
- **Context length가 길어질수록 representations의 형태가 격자 구조를 더 잘 반영**
- 초기 레이어의 representations은 prior의 영향을 받음 & 후반 레이어의 representations은 context에 의해 조정

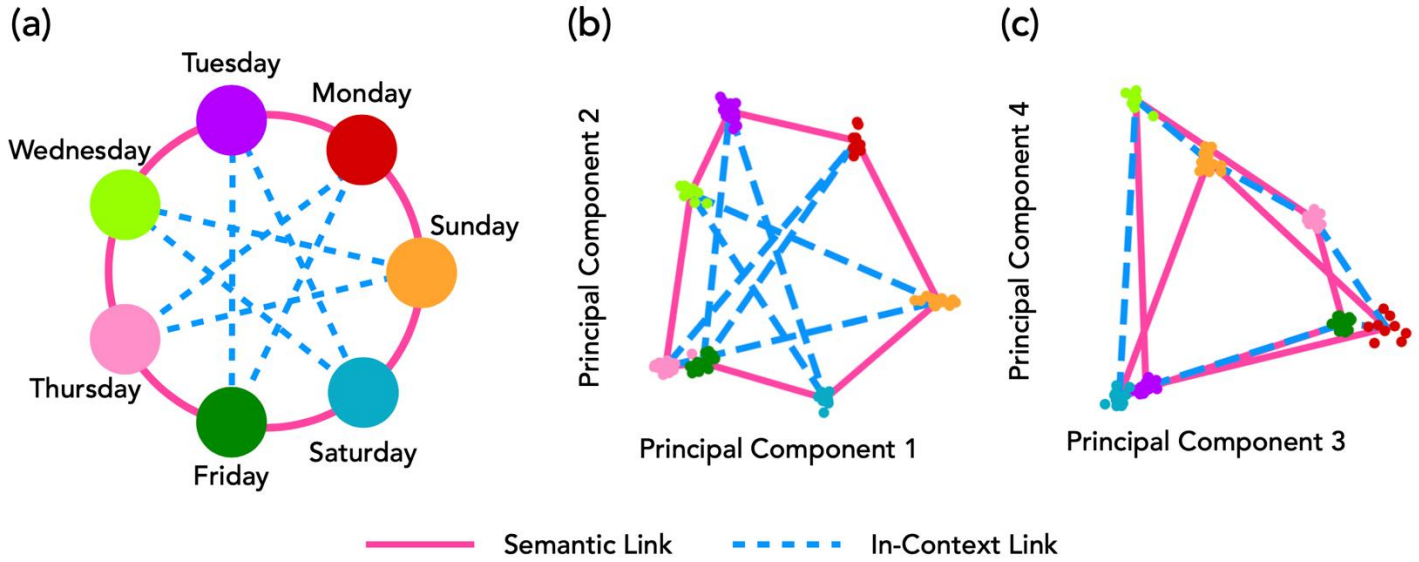


<Ring Graph Visualization>

Results

❖ Semantic Prior vs. In-context Task

- 의미적으로 상관관계가 존재하는 개념에 대해 실험함으로써, **in-context representations이 strong pretrained prior를 지배할 수 있는지** 분석
- 7개의 요일 단어를 랜덤하게 섞고 7-node ring graph 구성
  - ✓ 첫 두 주성분에서는 기존 semantic ring이 여전히 나타남
  - ✓ **3, 4번째 주성분에서 context 구조가 드러남**
- → Context 구조가 representations에 드러나긴 하지만, prior를 완전히 지배하지는 못함

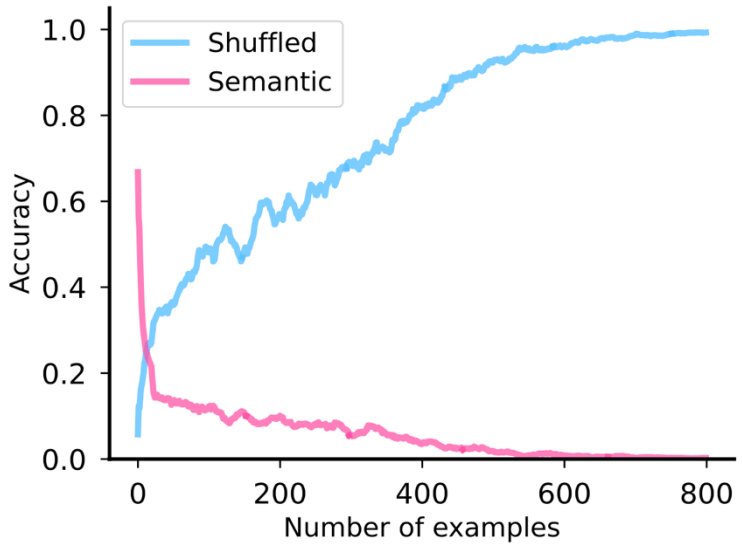


<Days of the Week Ring Graph Visualization>

Results

❖ Semantic Prior vs. In-context Task

- 의미적으로 상관관계가 존재하는 개념에 대해 실험함으로써, **in-context representations이 strong pretrained prior를 지배할 수 있는지** 분석
- 7개의 요일 단어를 랜덤하게 섞고 7-node ring graph 구성
  - ✓ 첫 두 주성분에서는 기존 semantic ring이 여전히 나타남
  - ✓ **3, 4번째 주성분에서 context 구조가 드러남**
- → Context 구조가 representations에 드러나긴 하지만, prior를 완전히 지배하지는 못함
- Context length가 길어질수록 인접 노드 예측 정확도가 in-context (Shuffled)는 증가, prior (Semantic)는 감소



<Context Length에 따른 인접 노드 예측 정확도>

Emergent Re-organization of Representations by Context Length

❖ **Context Length에 따른 in-context Task 성능 향상을 정량적으로 분석**

1) **Dirichlet Energy**: 그래프의 smoothness 측정

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

- $\boldsymbol{h}_i^{\ell}$: $\boldsymbol{H}^{\ell}(\mathcal{T})$ 의 i번째 행 (=노드 i의 mean token activations d차원 벡터)
- $\boldsymbol{A} \in \mathbb{R}^{n \times n}$: adjacency matrix
- 즉, 여기서 Dirichlet energy는 그래프 상 인접한 노드 사이 representations의 차이 측정
- **모델이 토큰의 그래프 구조를 올바르게 포착하면 Dirichlet energy는 감소**
- 모든 노드가 동일한 값을 갖는 trivial solutions은 실험에서 등장하지 않음

$$\mathcal{T} = \{\overset{1}{\text{apple}}, \overset{2}{\text{bird}}, \overset{3}{\text{car}}, \cdots, \overset{n-1}{\text{code}}, \overset{n(=16)}{\text{phone}}\}$$

(a) Words on a grid

Stack of mean token representations:

$$\boldsymbol{H}^{\ell}(\mathcal{T}) \in \mathbb{R}^{n \times d}$$

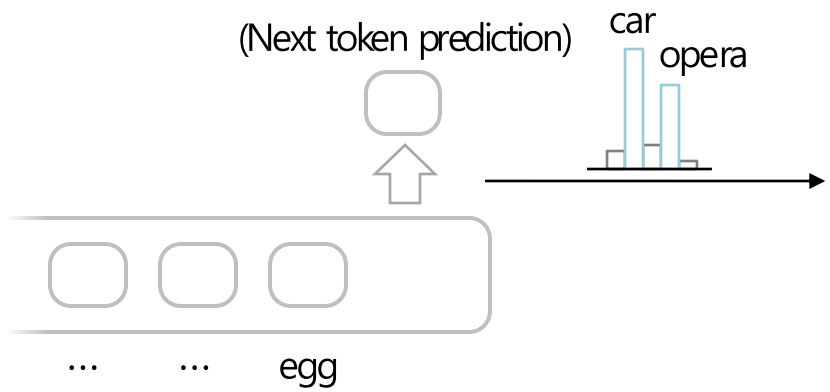
	apple	bird	car	⋯	code	phone
apple	0	1	0		0	0
bird	1	0	1		0	0
car	0	1	0		0	0
⋮				⋱		
code	0	0	0		0	1
phone	0	0	0		1	0

$\boldsymbol{A} \in \mathbb{R}^{n \times n}$: adjacency matrix

Emergent Re-organization of Representations by Context Length

❖ **Context Length에 따른 in-context Task 성능 향상을 정량적으로 분석**

2) Rule-following Accuracy: 인접 노드 예측 정확도



Rule-following Accuracy

$= P(\text{car}|c_{0:N-1}) + P(\text{opera}|c_{0:N-1})$

$\mathcal{T} = \{\overset{1}{\text{apple}}, \overset{2}{\text{bird}}, \overset{3}{\text{car}}, \cdots, \overset{n-1}{\text{code}}, \overset{n(=16)}{\text{phone}}\}$

(a) Words on a grid

Stack of mean token representations:

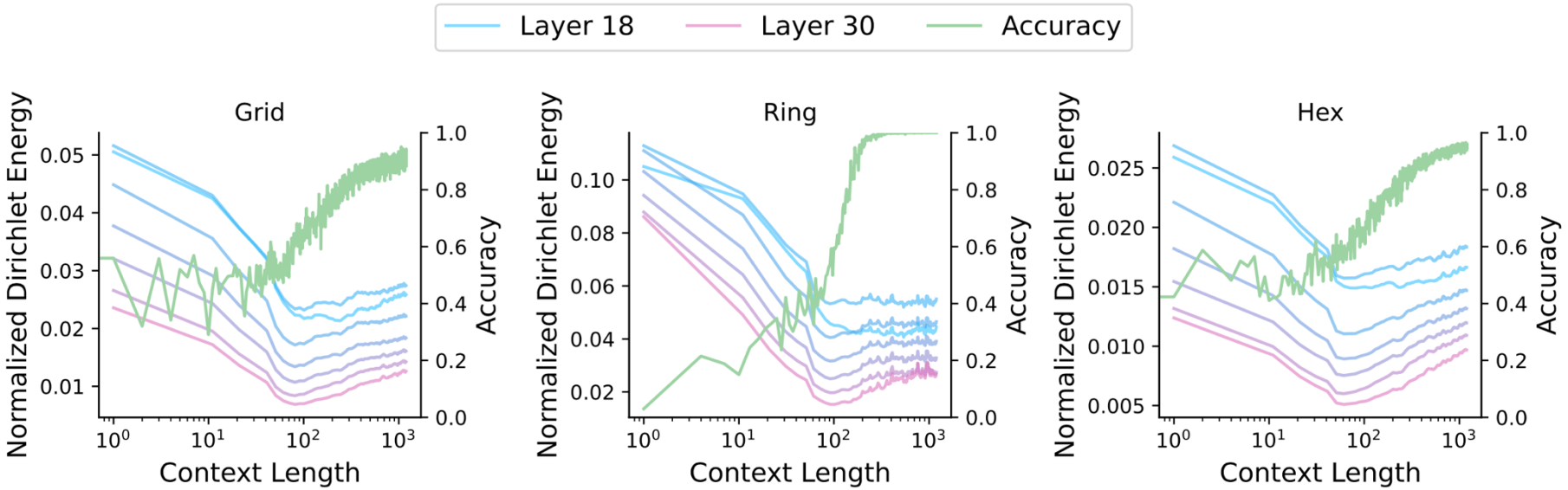
$\boldsymbol{H}^{\ell}(\mathcal{T}) \in \mathbb{R}^{n \times d}$

- 그래프 구조가 car-egg-opera이고 현재 state가 egg일 때, 다음 노드를 car 또는 opera라고 예측할 확률의 합
- **모델이 얼마나 그래프 구조를 잘 포착하고 있는지 측정**

Emergent Re-organization of Representations by Context Length

❖ **Context Length에 따른 in-context Task 성능 향상을 정량적으로 분석**

- Dirichlet energy가 최솟값을 갖는 지점 직후 accuracy가 크게 증가
- 즉, 모델이 데이터 구조를 올바르게 포착하면 정확한 예측 가능
- **Context 양이 많아지면 representations의 재구성이 발생**하고(emergent), 이는 모델이 **in-context graph tracing task를 잘 수행하도록** 만듦



<Context Length에 따른 Dirichlet Energy 및 Rule-following Accuracy>

Emergent Re-organization of Representations by Context Length

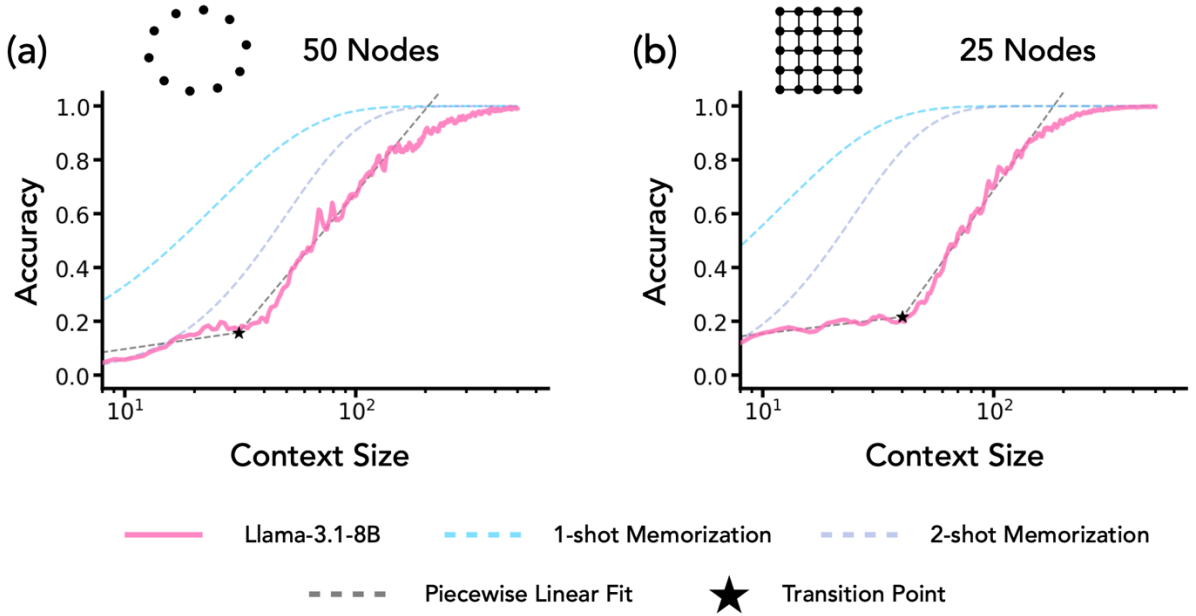
❖ **Is there a Trivial Solution at play?**

- 모델이 context 시퀀스를 단순히 암기하는 것인지 파악
  - ✓ 1-shot memorization: context에서 token을 적어도 한 번 보면 100%의 accuracy로 next token을 예측하는 가상 모델
  - ✓ 2-shot memorization: context에서 token을 적어도 두 번 보면 100%의 accuracy로 next token을 예측하는 가상 모델
- 1-shot 또는 2-shot memorization curve는 LLM의 행동을 설명하지 못함 → **"모델의 accuracy는 단순한 암기로 설명될 수 없다"**
- 또한 accuracy가 천천히 증가하다가 중간부터 급증하는 two phase 양상

$$p_{\text{seen1}}(\mathbf{x}) = 1 - \left(\frac{n-1}{n}\right)^{l}$$

$$p_{\text{seen2}}(\mathbf{x}) = p_{\text{seen1}}(\mathbf{x}) - l\left(\frac{1}{n}\right)^{1}\left(\frac{n-1}{n}\right)^{(l-1)}$$

<Memorization 이론적 모델링>



<Context Length에 따른 Rule-following Accuracy>

Re-organizing Representations through Energy Minimization

❖ **Energy Minimization Hypothesis**

- LLM이 데이터의 올바른 representations 구조를 찾기 위해 내부적으로 에너지 최소화 과정을 수행한다는 가설
- ICL이 gradient descent로서 implicit optimization을 수행한다는 이전 연구[1]와 관련

$$E_{\mathcal{G}}(\boldsymbol{H}^{\ell}(\mathcal{T})) = \sum_{i,j} \boldsymbol{A}_{i,j} \|\boldsymbol{h}_i^{\ell} - \boldsymbol{h}_j^{\ell}\|^2 = \left(\boldsymbol{h}^{\ell}\right)^T \boldsymbol{L}\boldsymbol{h}^{\ell} \qquad \boldsymbol{L}\text{: Laplacian matrix } (= \boldsymbol{D} - \boldsymbol{A})$$

- Laplacian matrix $\boldsymbol{L}$의 eigenvector들이 Dirichlet energy를 최소화한다는 것이 알려져 있음
  - ✓ $\boldsymbol{L}$의 eigenvalue: $0 = \lambda_1, \lambda_2, \cdots, \lambda_{n-1}, \lambda_n$
  - ✓ $\boldsymbol{L}$의 eigenvector: $\mathbf{1} = z_1, z_2, \cdots, z_{n-1}, z_n$
- (2-D) Spectral embeddings: 각 노드 i에 대해 $(z_{2,i}, z_{3,i})$는 그래프 구조를 잘 반영할 수 있는 좌표로서 사용

$\mathcal{T}$ = {apple, bird, car, ⋯, code, phone}
(1, 2, 3, $n-1$, $n(=16)$)

(a) Words on a grid

apple → bird		car	egg
house	milk	plane → opera	
box	sand → sun		mango
rock	math	code	phone

Stack of mean token representations:
$\boldsymbol{H}^{\ell}(\mathcal{T}) \in \mathbb{R}^{n \times d}$

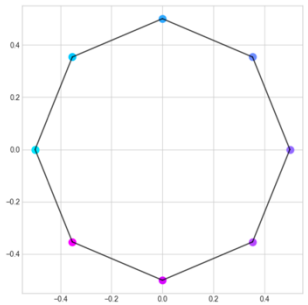


Figure 6: **Spectral embedding of a ring graph.**

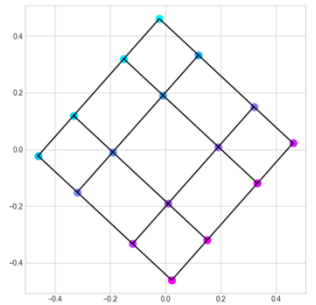


Figure 7: **Spectral embedding of a grid graph.**

<Spectral Embeddings $(z_{2,i}, z_{3,i})$ 시각화>

[1] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joa˜o Sacramento, Alexander Mordv- intsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In International Conference on Machine Learning, pp. 35151–35174. PMLR, 2023a.

Re-organizing Representations through Energy Minimization

❖ Energy Minimization Hypothesis

- LLM이 데이터의 올바른 representations 구조를 찾기 위해 내부적으로 에너지 최소화 과정을 수행한다는 가설
- ICL이 gradient descent로서 implicit optimization을 수행한다는 이전 연구[1]와 관련

$$E_{\mathcal{G}}(\boldsymbol{H}^{\ell}(\mathcal{T})) = \sum_{i,j} \boldsymbol{A}_{i,j} \|\boldsymbol{h}_i^{\ell} - \boldsymbol{h}_j^{\ell}\|^2 = \left(\boldsymbol{h}^{\ell}\right)^T \boldsymbol{L}\boldsymbol{h}^{\ell} \qquad \boldsymbol{L}\text{: Laplacian matrix } (= \boldsymbol{D} - \boldsymbol{A})$$

- Laplacian matrix $\boldsymbol{L}$의 eigenvector들이 Dirichlet energy를 최소화한다는 것이 알려져 있음
  - ✓ $\boldsymbol{L}$의 eigenvalue: $0 = \lambda_1, \lambda_2, \cdots, \lambda_{n-1}, \lambda_n$
  - ✓ $\boldsymbol{L}$의 eigenvector: $\mathbf{1} = z_1, z_2, \cdots, z_{n-1}, z_n$
- (2-D) Spectral embeddings: 각 노드 i에 대해 $(z_{2,i}, z_{3,i})$는 그래프 구조를 잘 반영할 수 있는 좌표로서 사용

$$\mathcal{T} = \{\underset{1}{\text{apple}}, \underset{2}{\text{bird}}, \underset{3}{\text{car}}, \cdots, \underset{n-1}{\text{code}}, \underset{n(=16)}{\text{phone}}\}$$

(a) Words on a grid

apple → bird		car	egg
house	milk	plane → opera	
box	sand → sun		mango
rock	math	code	phone

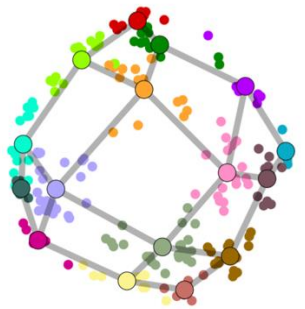
Stack of mean token representations:

$$\boldsymbol{H}^{\ell}(\mathcal{T}) \in \mathbb{R}^{n \times d}$$

❖ Theorem

- "모델 representations $\boldsymbol{H}$가 Dirichlet energy를 최소화하고 비자명한 해일 경우, $\boldsymbol{H}$의 첫 두 개 주성분은 spectral embeddings z_2와 z_3이다"
- 따라서, $\boldsymbol{H}$에 PCA를 적용해 시각화한 결과는 모델이 Dirichlet energy를 얼마나 줄이는지 보여줌
- Context에 따른 모델의 유추는 implicit energy minimization으로 해석 가능

<Spectral Embeddings $(z_{2,i}, z_{3,i})$>



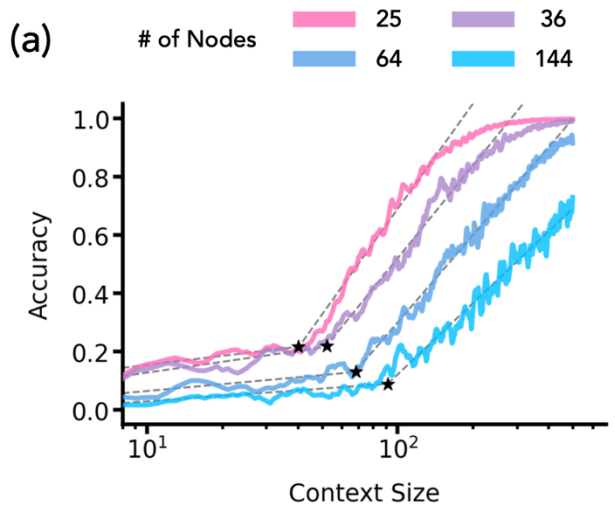
<모델 representations $\boldsymbol{H}$>

[1] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joa˜o Sacramento, Alexander Mordv- intsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In International Conference on Machine Learning, pp. 35151–35174. PMLR, 2023a.

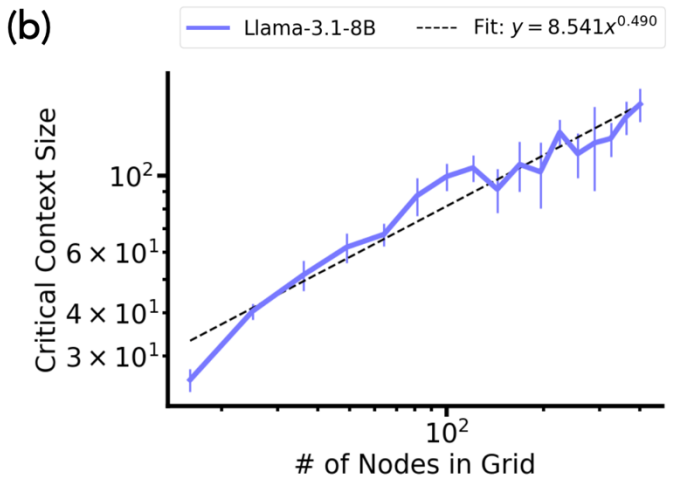
Re-organizing Representations through Energy Minimization

❖ **In-context Emergence**

- Context 크기에 따른 accuracy 측정 시 공통적으로 중간에 양상이 변하는 **second-order phase transition** 존재
- Transition point는 노드 개수에 대해 power-law scaling trand를 따름
- 이 two-phase 양상은 **bond-percolation on a graph** [1, 2] 와 연관지을 수 있음
  - ✓ bond-percolation: 그래프 노드 간 엣지를 일정 확률로 연결하거나 제거하면서 전체 그래프 연결성을 관찰



<Node 개수별 context Length에 따른 Rule-following Accuracy>



<Node 개수에 따른 transition point>

[1] M. E. J. Newman. The Structure and Function of Complex Networks. SIAM Review, 45(2):167–256, January 2003. ISSN 0036-1445, 1095-7200.
[2] H. Hooyberghs, B. Van Schaeybroeck, and J. O. Indekeu. Percolation on bipartite scale-free net- works. Physica A: Statistical Mechanics and its Applications, 389(15):2920–2929, August 2010. ISSN 0378-4371.

Conclusion

❖ Contributions

- **Graph tracing task**: LLM이 그래프 구조로 이루어진 노드들의 in-context 관계를 포착하고 예측할 수 있는지 분석
- **Context 크기 확장에 따른 representations의 재구성**: context 크기가 확장되면서 representations이 그래프 구조를 반영하도록 재구성되는 현상을 관찰
- **에너지 최소화 기반 의미 추론**: Dirichlet energy 감소를 통해 LLM이 context에 따라 representations을 재구성하는 메커니즘을 정량적으로 설명

- 입력 context를 그래프 구조로 정의 + 모델의 representations이 그래프 모양으로 시각화되는 것을 보임 + 그래프 이론 쪽에서 잘 정의된 개념들을 가지

  → "**모델이 context를 이해, 포착하는 방식**을 그래프 이론에서 **에너지를 최소화(smoothness를 최대화)하는 방식**으로 해석할 수 있다"

Thank you