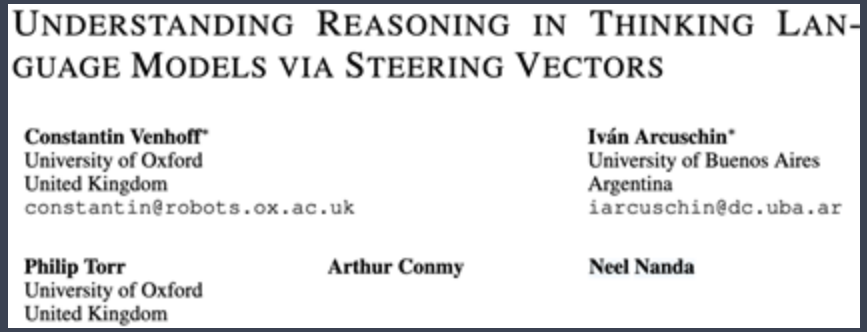
# Reasoning Interpretability

2025 하계 세미나

김동준



UNDERSTANDING REASONING IN THINKING LANGUAGE MODELS VIA STEERING VECTORS

**Constantin Venhoff***
University of Oxford
United Kingdom
constantin@robots.ox.ac.uk

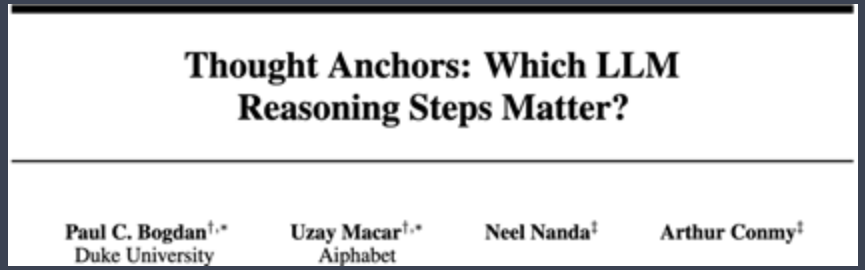
**Iván Arcuschin***
University of Buenos Aires
Argentina
iarcuschin@dc.uba.ar

**Philip Torr**
University of Oxford
United Kingdom

**Arthur Conmy**

**Neel Nanda**

- ICLR 2025 Workshop on Reasoning and Planning for LLMs
- Focus: Steering Vector로 언어 모델의 internal activations를 직접 조작하여, 모델의 추론 행동을 바꾸고 제어할 수 있는가?
- Key insight: 모델의 reasoning 과정은 더 작은 개별 행동들에 해당하는 특정 내부 활성화 패턴을 직접 조작함으로써 제어할 수 있음



Thought Anchors: Which LLM Reasoning Steps Matter?

**Paul C. Bogdan**[†,*]
Duke University

**Uzay Macar**[†,*]
Aiphabet

**Neel Nanda**[‡]

**Arthur Conmy**[‡]

- arXiv 2025
- Focus: 모델의 긴 CoT에서, 최종 답변을 결정하는 데 불균형적으로 큰 영향을 미치는 가장 중요한 단계들을 어떻게 식별할 수 있는가?
- Key insight: 모델의 최종 답변은 모든 reasoning 단계가 동등하게 기여하는 것이 아니라, 소수의 핵심적 단계에 의해 불균형적으로 결정됨

# Understanding Reasoning in Thinking Language Models via Steering Vectors

# Motivation & Contribution

**RQ:** RLM들의 Reasoning Behavior를 Steering Vector를 통해 바꿀 수 있을까

1. Reasoning Step에 중요한 부분들을 담당하는 SV들을 뽑고 Control 하는 방법 제시
2. DeepSeek R1 Distill 모델들을 500개 태스크에 대해 Steering한 결과

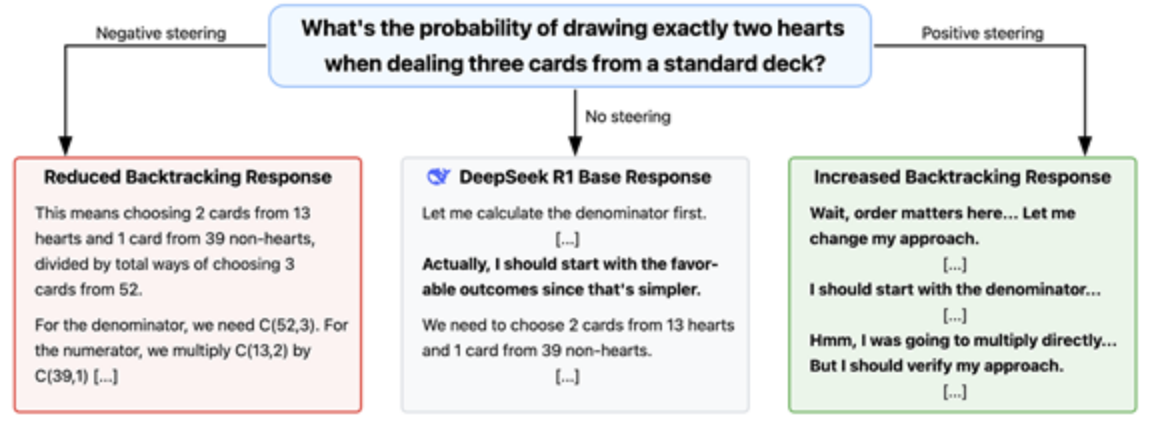
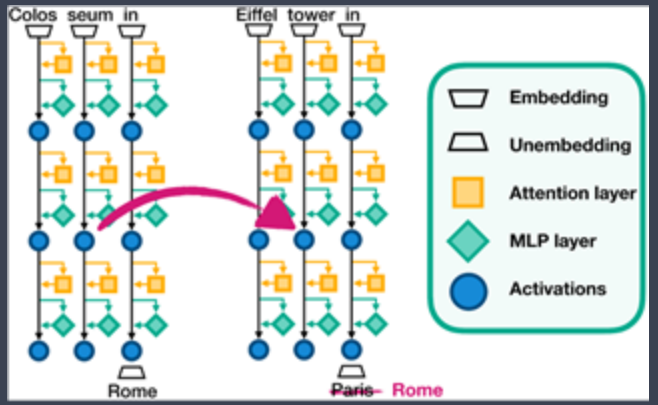



Figure 1: *Steering* on DeepSeek-R1's *backtracking* feature vector changes the model's behavior. Depending on whether we add or subtract this vector to the activations at inference time, the model increases or decreases its tendency to abandon its current approach and explore alternative strategies for the task at hand. Highlighted sections indicate instances of this behavior.

# Methodology

- Step 1: 모델의 어떤 부분들이 이러한 Behavior를 담당하는지 알아내야 함
  - Relevant Layers
    - 일반적으로 사용되는 Activation Patching 대신 더 효율적인 Attribution Patching 사용함



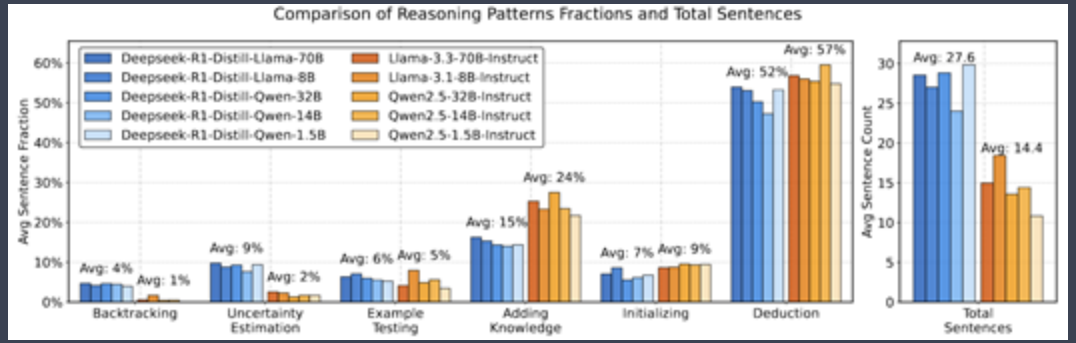

- Step 2: 찾은 layer를 통해 Steering Vector를 만들어야 함
  - Difference of Means method

$$\mathbf{u} = \frac{1}{|D_+|} \sum_{p_i \in D_+} \mathbf{a}(p_i) - \frac{1}{|D_-|} \sum_{p_j \in D_-} \mathbf{a}(p_j)$$

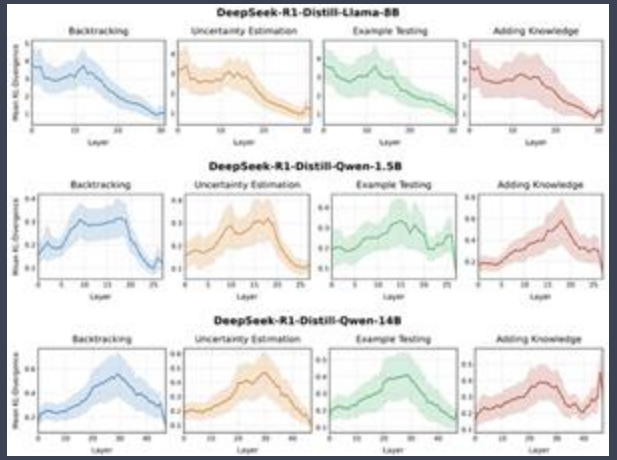
- Step 3: Inference Time Steering

# Methodology

- Hypothesis: Reasoning Behavior들은 여러 간단한 behavior들로 나뉘어진다
    - **Initialization:** 모델이 query를 반복하고 초기 생각을 명시하며, 일반적으로 reasoning chain의 시작 부분에서 발생함
    - **Deduction:** 모델이 현재의 접근 방식과 가정을 기반으로 결론을 도출함
    - **Knowledge Augmentation:** 모델이 자신의 추론을 정교화하기 위해 외부 지식을 통합함
    - **Example Testing:** 모델이 자신의 작동 가설을 검증하기 위해 예시나 시나리오를 생성함
    - **Uncertainty Estimation:** 모델이 자신의 추론에 관한 확신이나 불확실성을 명시적으로 진술함
    - **Backtracking:** 모델이 현재의 접근 방식을 포기하고 대안 전략을 탐색함

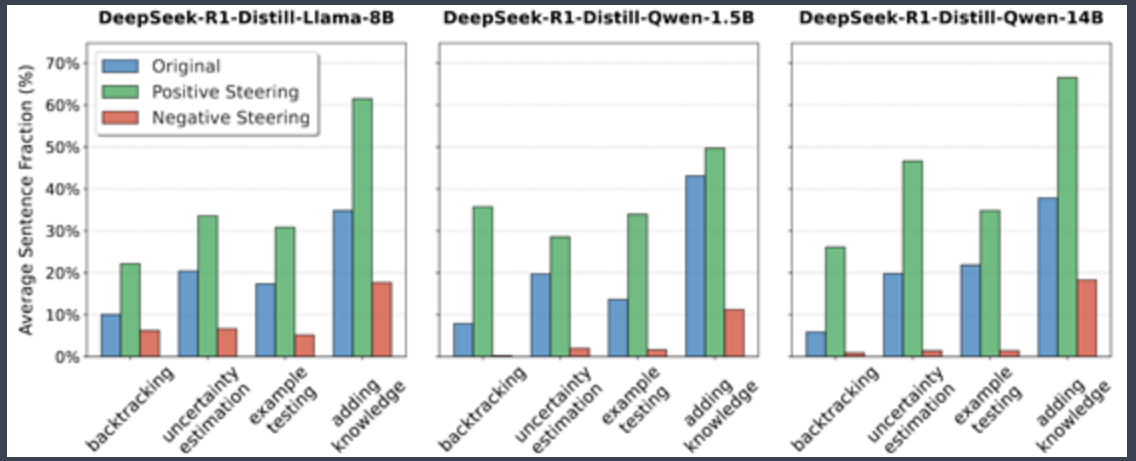


# Results




| Behavioral Category | DeepSeek-R1-Distill Llama-8B | DeepSeek-R1-Distill Qwen-1.5B | DeepSeek-R1-Distill Qwen-14B |
|---|---|---|---|
| Uncertainty Estimation | 12 | 18 | 29 |
| Example Testing | 12 | 15 | 29 |
| Backtracking | 12 | 17 | 29 |
| Adding Knowledge | 12 | 18 | 24 |

Table 2: Selected layers for each behavioral category based on attribution patching results. For each model, we select the layer with the maximum score from the attribution patching experiments, ignoring early layers that are highly correlated with embedding tokens.



# Thought Anchors: Which LLM Reasoning Steps Matter?

# Motivation & Contribution

**RQ:** RLM들의 긴 Reasoning Trace에서, 최종 답변을 결정하는 데 가장 중요한 단계를 어떻게 찾아낼 수 있을까?

- Thought Anchors: Reasoning 방향과 최종 답변에 결정적인 영향을 미치는 가장 중요하고 핵심적인 문장
  - 이 "Anchor"들은 최종 결과에 불균형적으로 큰 영향을 미침
  - 식별함은 CoT의 효율성과 reliability를 향상시킬 수 있음
- 기존의 MI 방법들은 긴 response를 다룰 수 없다는 한계가 있음 => Sentence-Level Analysis

1. Black-Box Counterfactual Resampling: 특정 문장을 대체하고, 그로 인해 모델의 최종 답변 분포에 나타나는 변화를 관찰하여 문장의 중요도를 측정하는 방법
2. White-Box Attention Aggregation: 모델의 내부 어텐션 패턴을 분석하여, 중요한 broadcasting 문장에 지속적으로 집중하는 특화된 receiver head를 찾아내는 상관관계 기반 방법
3. Causal Attention Suppression: 특정 문장에 대한 어텐션을 마스킹하여 모델의 computation을 직접 조작하고, 이를 통해 reasoning chain의 단계들 사이의 정확한 인과적 의존 관계를 파악하는 방법

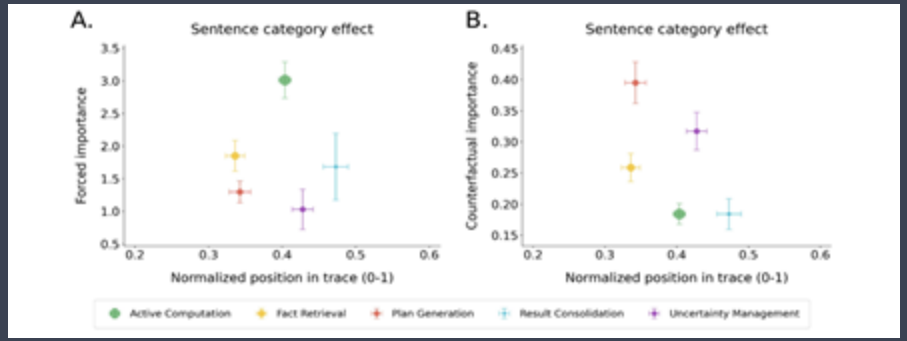
# Sentence Categorization

1. Problem Setup: 초기 문제 서술을 분석하거나 다시 표현함.
2. Plan Generation: 행동 전략이나 계획을 명확히 설명함; 메타 추론의 한 형태.
3. Fact Retrieval: 계산을 수행하지 않고 공식, 사실, 또는 문제의 세부사항을 상기함.
4. Active Computation: 답을 도출하기 위해 대수적 조작, 계산 또는 기타 단계를 수행함.
5. Uncertainty Management: 혼란을 표현하거나, 이전 단계를 재평가하거나, 특정 추론 과정에서 되돌아감.
6. Result Consolidation: 중간 결과를 종합하거나 진행 상황을 요약함.
7. Self Checking: 이전의 계산이나 단계를 명시적으로 검증함.
8. Final Answer Emission: 최종 답변을 서술함.

# Methodology 1: Black-Box Counterfactual Resampling

- Counterfactual Importance: 문장을 의미적으로 다른 문장으로 대체했을 때의 Final Answer Probability의 KL Divergence를 측정하여, planning과 같은 상위 수준 문장의 중요도를 Forced Importance 보다 정확히 포착
  - 특정 문장($S_i$)이 있을 때와 없을 때의 최종 답변 분포를 비교하여 문장의 중요도를 측정
- 단순히 추론을 중단하는 forced-answering 방식과 달리,
  문장 제거 후 전체 추론 과정을 100회 resampling하여 보다 robust한 인과관계를 파악함

- 결과:
  - Plan Generation과 Uncertainty Management 문장이 다른 유형의 문장보다 일관되게 높은 중요도를 보임
  - 이는 모델의 추론이 저수준의 계산 단계가 아닌, 고수준의 계획 및 방향 설정 단계에 의해 좌우됨을 보여줌

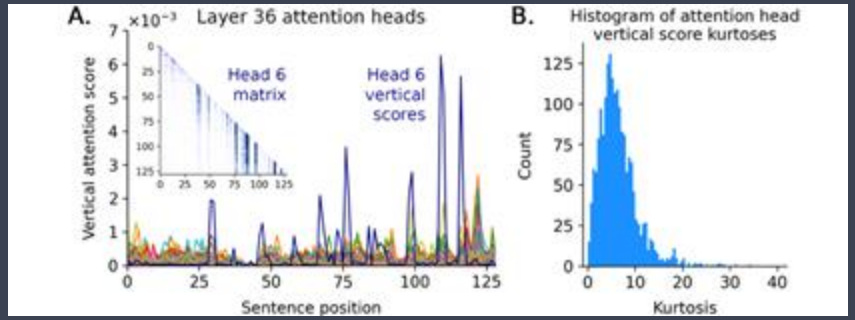


# Methodology 2: White-Box Attention Aggregation

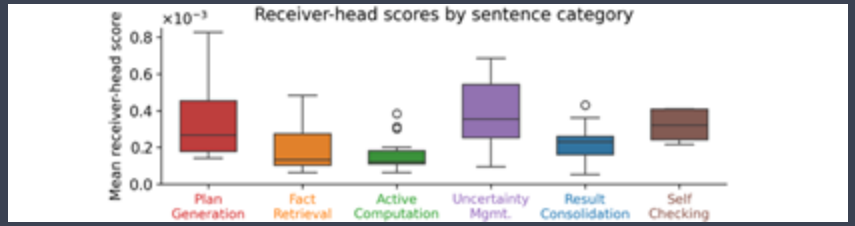
- Receiver Heads: 특정 과거 문장에 어텐션을 집중시키는 특수화된 어텐션 헤드. 어텐션 분포의 높은 첨도(Kurtosis)를 통해 식별
- Broadcasting Sentences: Receiver Head들이 집중적으로 주목하는, 중요 정보를 '방송'하는 핵심 문장들
- 모델 내부의 어텐션 가중치를 직접 분석하여 문장 간의 관계를 파악하는 White-Box 접근법
- 중요한 문장은 후속 문장들로부터 더 많은 어텐션을 받을 것이라는 가설에 기반

# Methodology 2: White-Box Attention Aggregation

- 결과:
  - 'Plan Generation', 'Uncertainty Management'과 같은 high level 문장들이 가장 많은 어텐션을 받음.
  - Receiver Head를 제거할 경우, 무작위 헤드를 제거했을 때보다 모델의 정확도가 더 크게 하락하여 기능적 중요성을 입증함.
  - Reasoning 모델은 Base 모델에 비해 특정 중요 문장에 어텐션을 더 집중시키는 경향을 보임.




High Kurtosis => Attention is concentrated on a few specific sentences

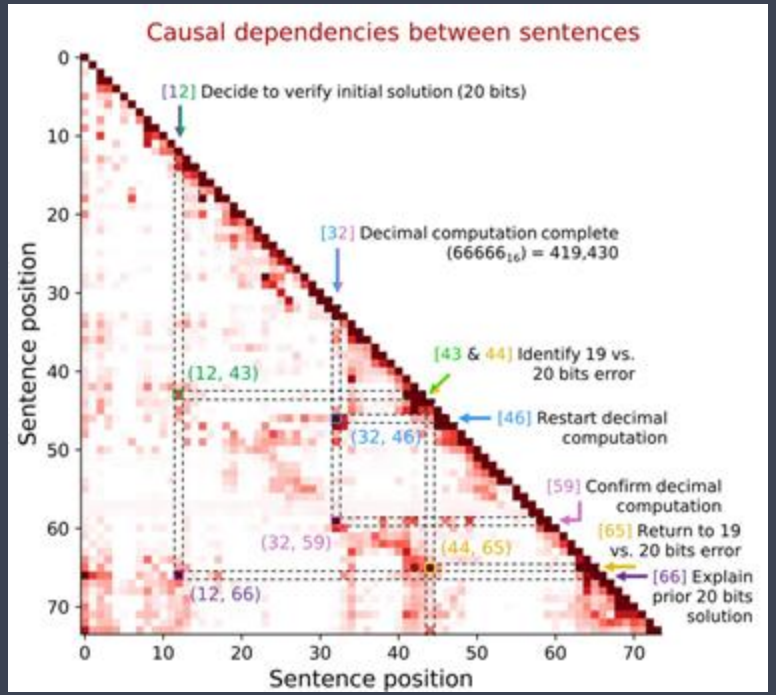


# Methodology 3: Causal Attention Suppression

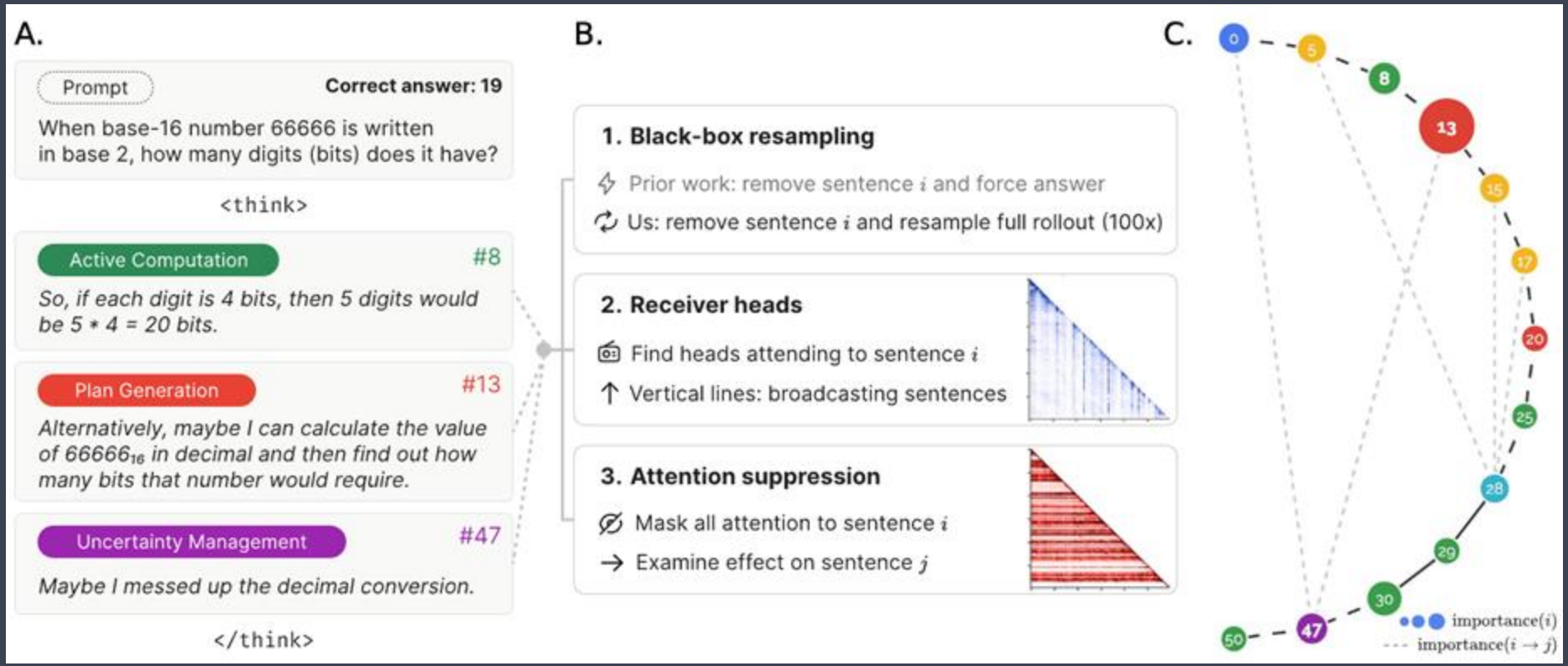
- 특정 문장으로 향하는 모든 어텐션을 인위적으로 masking하여, 후속 문장 생성에 미치는 영향을 측정
  - Black-Box 방식이 total effect를 측정하는 반면, 이 방법은 direct effect를 분리하여 더 정밀한 논리적 연결망을 파악

- 어텐션 차단 전후 logits 간 KL Divergence를 계산하여, 문장 $S_i$가 문장 $S_j$에 미치는 직접적인 영향력을 정량화

- 결과:
  - 문장들 사이의 정보 흐름과 인과적 의존성을 정밀하게 매핑하여, 추론 과정의 내부 circuit을 발견함.
  - Case Study에서 잘못된 초기 답변 제시 → 불일치 발견 → 최종 오류 수정으로 이어지는 일련의 self-correction 패턴을 성공적으로 포착
  - 다른 두 방법론의 결과와 높은 상관관계를 보여주며 접근법의 타당성을 입증함.



# https://www.thought-anchors.com/



# Conclusion

- 두 연구의 통합적 관점
  - 'Thought Anchors': LLM의 긴 추론 과정 속에서 최종 답변에 가장 큰 영향을 미치는 핵심적인 '사고의 앵커' 단계를 식별하는 방법을 제시함
  - 'Steering Vectors': 모델의 내부 활성화를 직접 조작하여 '백트래킹'이나 '불확실성 표현'과 같은 특정 추론 행동을 강화하거나 억제하는 방법을 보여줌
- Precise Control through Identification
  - 두 연구를 결합하여, Thought Anchors로 식별된 핵심 추론 단계에 Steering Vectors를 적용함으로써 보다 효율적이고 정밀한 추론 제어가 가능해짐
  - 예시로, 모델의 Plan Generation 앵커 단계에 back tracking 벡터를 적용하여, 초기 계획의 오류 발생 시 더 효과적인 대안 탐색을 유도할 수 있음
- Future Work
  - Trustworthy AI: 추론 과정의 핵심 지점을 파악하고 제어함으로써, 모델이 잘못된 방향으로 추론할 때 이를 수정하고 보다 신뢰할 수 있는 답변 생성을 유도함
  - Transparency: LLM이 단순한 블랙박스를 넘어, 내부 작동 방식을 이해하고 분석하며 원하는 방향으로 이끌 수 있는 해석 가능한 모델로 발전할 가능성을 확인