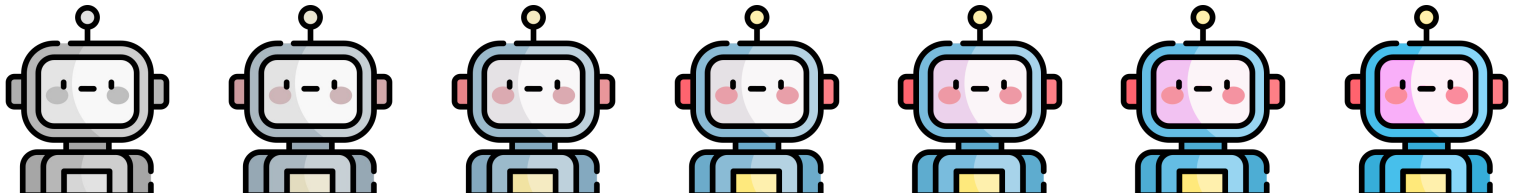
Scaling Multi-Agent LLM Systems

: Performance Gains and Diversity Collapse

260716 하계 세미나
정지민

Natural Language Processing
& Artificial Intelligence

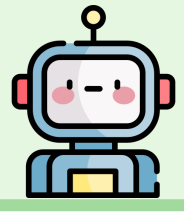


Background

Why Multi-Agent Collaboration?

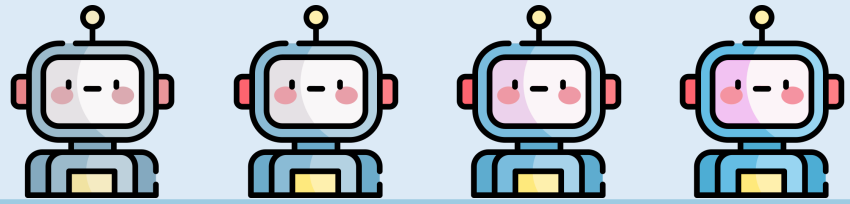
Limitations of **Single LLM**

- 복잡한 과제에서 자신의 오류를 발견하지 못할 가능성
- 하나의 추론 경로나 초기 판단에 고착될 가능성
- 문제를 한 번에 처리하면서 일부 하위 요구사항을 놓칠 가능성



Common **Multi-Agent** Designs

- Planner·Critic·Solver 등 역할 분담
- 서로의 출력을 검토·수정
- 복잡한 문제를 하위 작업으로 나누어 수행



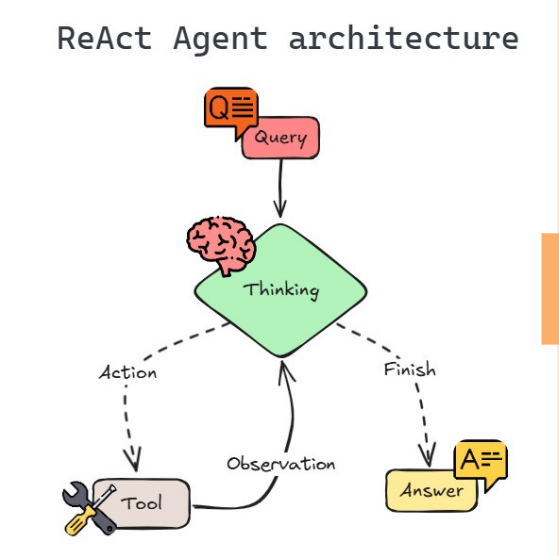
Expected **benefits** of using multi-agent systems

Better Answers & Broader Exploration

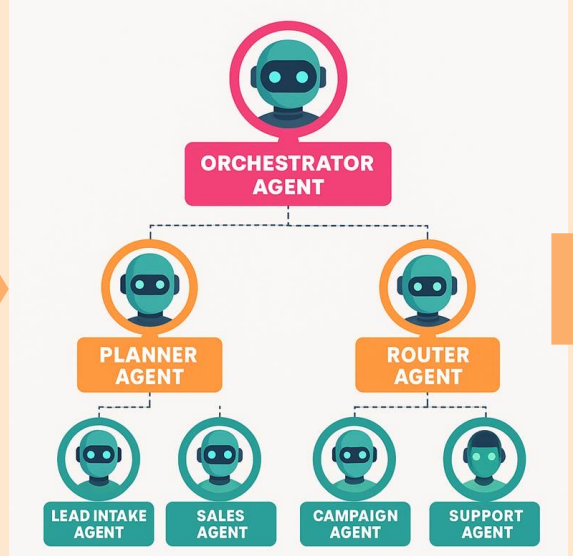
Background

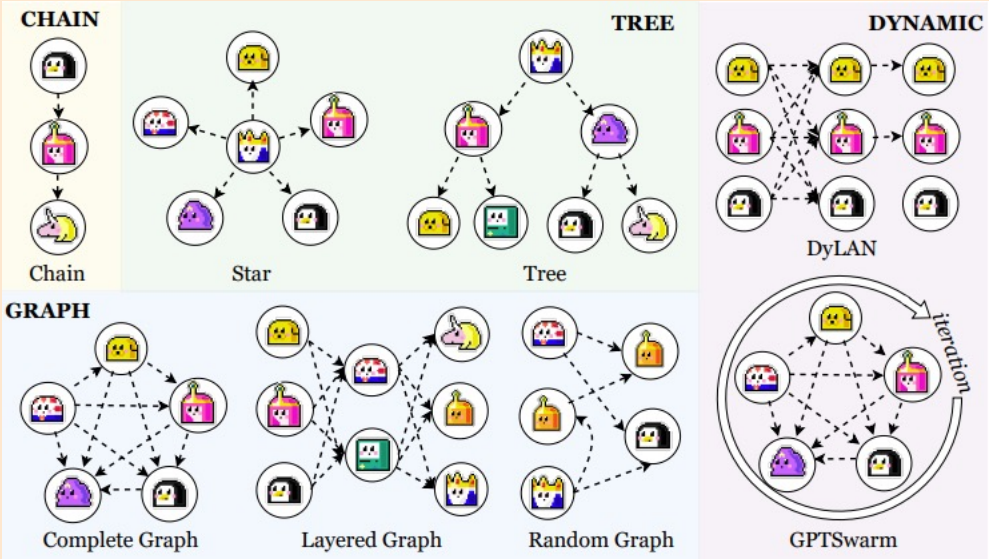
Why Multi-Agent Collaboration?

Limitations of Single LLM



Common Multi-Agent Designs





Expected benefits of using multi-agent systems

Better Answers & Broader Exploration

Background

Why Multi-Agent Collaboration?

그렇다면 에이전트 수를 늘리고, 통신 구조를 확장하면
Better Answers과 Broader Exploration을 얻을 수 있을까?

SCALING LARGE LANGUAGE MODEL-BASED MULTI-AGENT COLLABORATION

Chen Qian*†, Zihao Xie*†, YiFei Wang*, Wei Liu*, Kunlun Zhu*, Hanchen Xia*, Yufan Dang*, Zhuoyun Du*, Weize Chen*, Cheng Yang♣, Zhiyuan Liu*, Maosong Sun*✉
*Tsinghua University ♣Peng Cheng Laboratory
qianc62@gmail.com xie-zh22@mails.tsinghua.edu.cn sms@tsinghua.edu.cn

RQ: 에이전트의 수와 토폴로지를 확장했을 때

최종 task 성능이 어떻게 변화하는가?

Diversity Collapse in Multi-Agent LLM Systems: Structural Coupling and Collective Failure in Open-Ended Idea Generation

Nuo Chen[1] Yicheng Tong[1] Yuzhe Yang[2] Yufei He[1]
Xueyi Zhang[2] Qian Wang[1] Qingyun Zou[1] Bingsheng He[1]
[1]National University of Singapore [2]The Chinese University of Hong Kong, Shenzhen

RQ: 집단 규모와 상호작용 구조를 확장했을 때

아이디어의 폭과 독립적인 탐색이 어떻게 변화하는가?

SCALING LARGE LANGUAGE MODEL-BASED MULTI-AGENT COLLABORATION

Chen Qian[⋆†], **Zihao Xie**[⋆†], **YiFei Wang**[⋆], **Wei Liu**[⋆], **Kunlun Zhu**[⋆], **Hanchen Xia**[⋆], **Yufan Dang**[⋆], **Zhuoyun Du**[⋆], **Weize Chen**[⋆], **Cheng Yang**[♣], **Zhiyuan Liu**[⋆], **Maosong Sun**[⋆✉]
⋆Tsinghua University ♣Peng Cheng Laboratory
qianc62@gmail.com xie-zh22@mails.tsinghua.edu.cn sms@tsinghua.edu.cn

ICLR 2025

Introduction

Limitations of Existing Multi-Agent Collaboration

- 기존 Multi-Agent 연구는 대부분 10개 미만의 에이전트만 사용함
- 에이전트 수 증가와 협업 성능 간 관계는 충분히 분석되지 않음 -> Does multi-agent collaboration exhibit a scaling law?
- Majority voting 및 전체 메시지 공유는 context explosion을 유발함

Contributions of MACNET

(1) **Scalable Multi-Agent Network 제안:** Actor-Critic 역할을 DAG로 연결하여 대규모 협업이 가능한 MACNET 설계

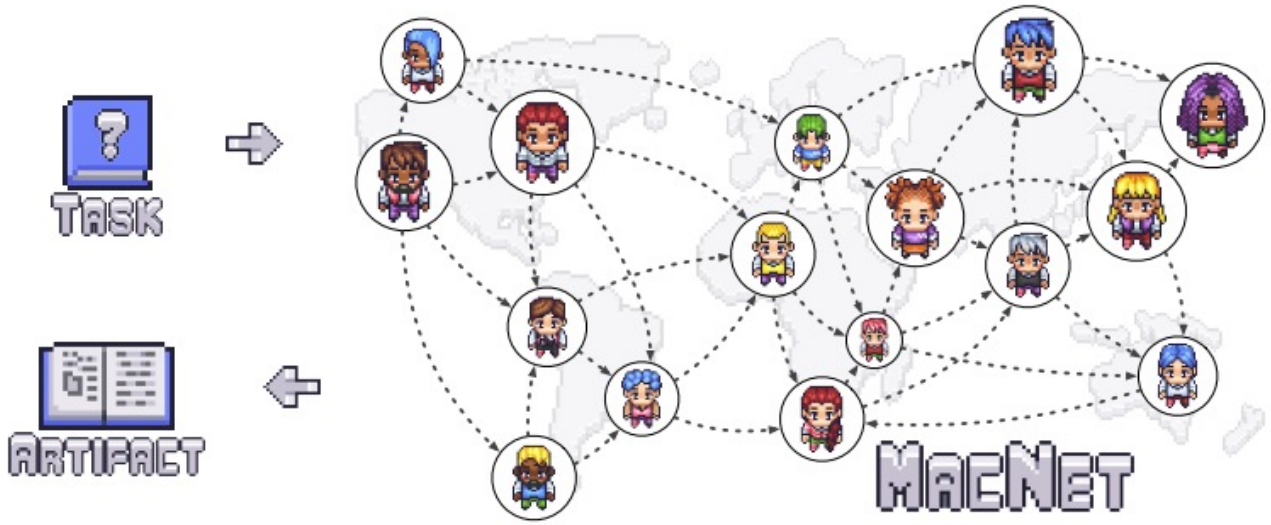
(2) **효율적인 정보 전달 방식 제안**: 전체 대화 대신 정제된 artifact만 전파하여 Global Broadcasting과 Context Explosion 방지

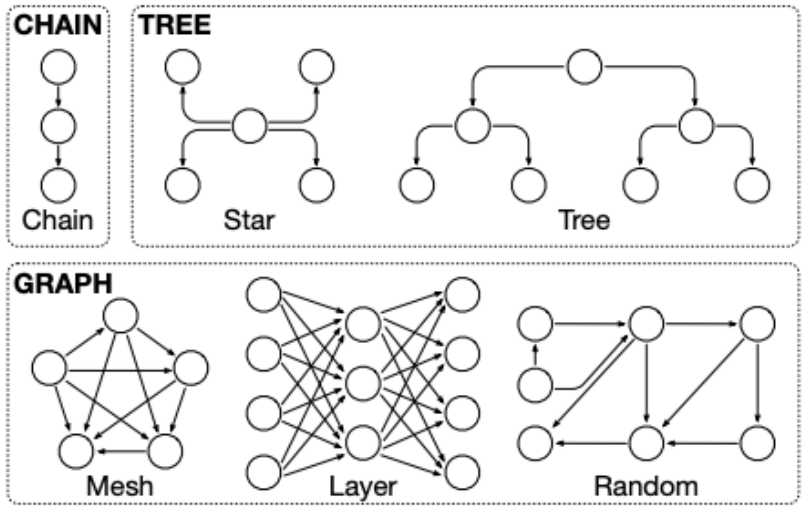
(3) **1000+ Agents 규모의 토폴로지 실험:** Chain, Tree, Graph를 비교하고 Irregular Topology의 성능 우위 발견

(4) **Collaborative Scaling Law 규명**: 에이전트 수 증가에 따라 협업 성능이 Logistic Growth 형태로 향상됨을 확인

Multi-Agent Collaboration Network

Network Construction





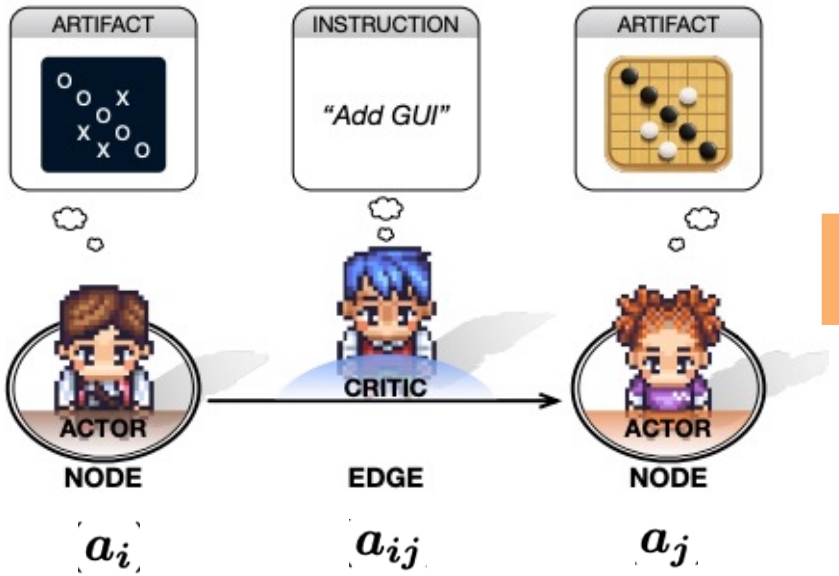
- MACNET은 여러 에이전트의 상호작용을 DAG (방향성 비순환 그래프)로 구성 => 비순환 구조를 통해 정보의 역류와 반복적 상호작용을 방지

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \quad \mathcal{V} = \{v_i | i \in I\} \quad \mathcal{E} = \{\langle v_i, v_j \rangle | i, j \in I \wedge i \neq j\}$$

- 각 노드 V는 에이전트, 엣지 E는 에이전트 간 정보 전달을 의미

Multi-Agent Collaboration Network

Interaction Reasoning



$$\forall \langle v_i, v_j \rangle \in \mathcal{E},\ \mathbb{I}(\boldsymbol{a_i}) < \mathbb{I}(\boldsymbol{a_{ij}}) < \mathbb{I}(\boldsymbol{a_j})$$

: 화살표 방향을 거스르지 않는 control flow

$$\tau(a_i, a_{ij}, a_j) = \tau(a_i, a_{ij}), \tau(a_{ij}, a_j)$$
$$\tau(a_i, a_{ij}) = (a_i \rightarrow a_{ij}, a_{ij}; a_i) \circlearrowleft$$
$$\tau(a_{ij}, a_j) = (a_{ij} \rightarrow a_j, a_j; a_{ij}) \circlearrowleft$$

: 각 연결선 내에서는 critic과 actor과 반복 상호작용

$$\boldsymbol{a_i} = \rho(v_i),\ \forall v_i \in \mathcal{V} \quad \boldsymbol{a_{ij}} = \rho(\langle v_i, v_j \rangle),\ \forall \langle v_i, v_j \rangle \in \mathcal{E}$$

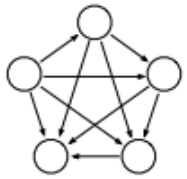
*ρ: 에이전트화 연산 (context aware memory, external tools, professional roles 부여)

Multi-Agent Collaboration Network

Memory Control

각 에이전트의 context를 관리하기 위해 단기 및 장기 메모리를 모두 채택함

- **단기 메모리**: 각 상호작용 내의 작업 메모리를 포착하여 컨텍스트를 인식하는 의사결정을 보장
- **장기 메모리**: 현대 대화에서 도출된 최종 아티팩트만을 유지함으로써 context 연속성을 유지 + 아티팩트 이외의 context가 후속 에이전트에게 접근 불가능하도록 보장



Mesh 구조의 경우 o(n^2)

$$\mathcal{O}(n)_{\text{w/o}} = t + p + s + \boxed{(2m-1)(i+s)(n(n-1)/2 + 2(n-2)) \overset{n\gg 1}{\approx} Cn^2 \propto n^2}$$ 메모리 제어가 없는 경우

$$\mathcal{O}(n)_{\text{w/}} = t + p + s + \boxed{m(i+s)((n-1) + 2(n-2)) \overset{n\gg 1}{\approx} \bar{C}n \propto n}$$ 메모리 제어가 있는 경우

$$\text{where} \quad C \equiv (2m-1)(i+s)/2 \quad C \equiv 3m(i+s)$$

*n: 에이전트의 수, i+s: 메시지의 길이, m: 상호작용의 횟수

Evaluation

Baselines

- CoT: 중간 추론 단계를 생성하여 논리력을 향상하는 기법
- AUTOGPT: 도구 활용과 계획 수립을 통해 작업을 수행하는 에이전트
- GPTSWARM: 에이전트 군집을 최적화 가능한 그래프로 운영
- AGENTVERSE: 에이전트 간 토론으로 결과물을 다듬는 프레임워크

Datasets and Metrics

- MMLU: 모델의 상식과 일반 지식을 묻는 객관식 시험 (정확도)
- HumanEval: 파이썬 함수를 올바르게 짜는지 평가하는 코딩 테스트 (정확도 – pass@k)
- SRDD: 실제 앱 수준의 복잡한 소프트웨어를 만드는 태스크 (완전성, 실행 가능성 및 일관성을 포함하는 종합 지표)
- CommonGen-Hard: 제시된 단어로 논리적인 문장을 만드는 작문 시험 (문법, 유창성, 문맥 관련성 및 논리적 일관성을 포함한 요소들의 종합 지표)

Evaluation

Does Our Method Lead to Improved Performance?

Method	Paradigm	MMLU	HumanEval	SRDD	CommonGen	Quality
CoT		0.3544†	0.6098†	0.7222†	0.6165†	0.5757†
AutoGPT		0.4485†	0.4809†	0.7353†	0.5972	0.5655†
GPTSwarm		0.2368†	0.4969†	0.7096†	0.6222†	0.5163†
AgentVerse		0.2977†	**0.7256**†	0.7587†	0.5399†	0.5805
MacNet-Chain		0.6632	0.3720	**0.8056**	0.5903	0.6078
MacNet-Star		0.4456†	0.5549†	0.7679†	0.7382†	0.6267
MacNet-Tree		0.3421†	0.4878†	0.8044	**0.7718**†	0.6015
MacNet-Mesh		0.6825	0.5122†	0.7792†	0.5525†	0.6316†
MacNet-Layer		0.2780†	0.4939†	0.7623†	0.7176†	0.5629†
MacNet-Random		**0.6877**	0.5244†	0.8054	0.5912	**0.6522**†

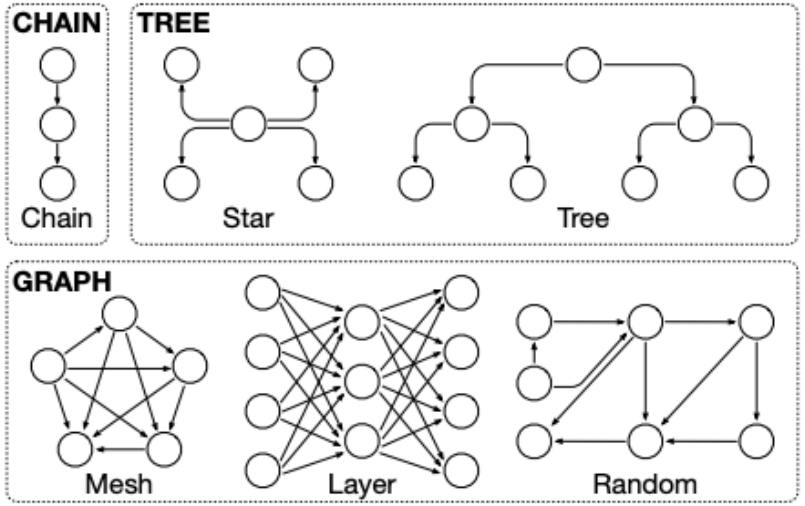
1. **토폴로지별 적합성 존재**: SW 개발은 Chain, 창의적 작문은 Tree에서 최적의 성능을 기록함

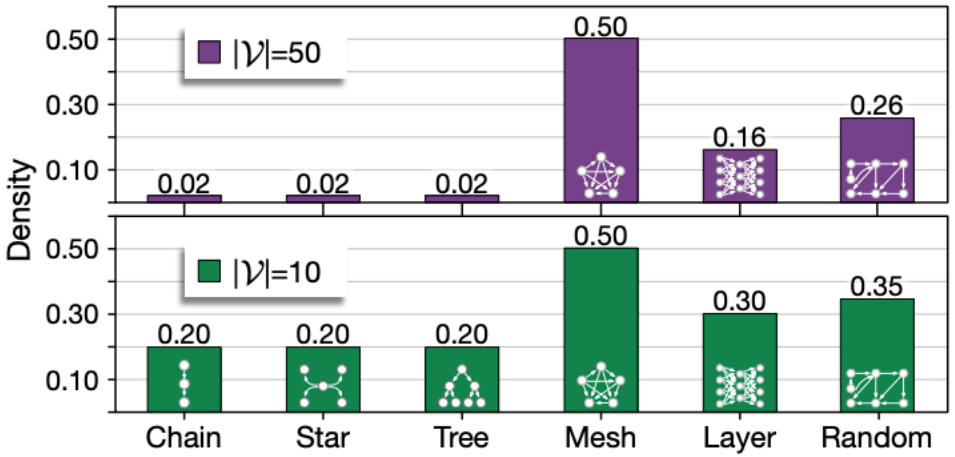
2. **Random 구조의 우수성**: 불규칙한 Random 토폴로지가 평균 Quality 0.6522로 실험 설정 중 가장 높은 성능을 달성함

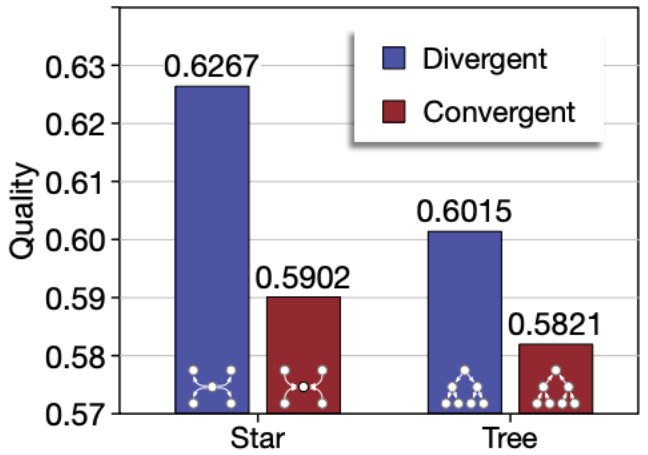
Evaluation

How Do Different Topologies Perform Against Each Other?

Method	Paradigm	MMLU	HumanEval	SRDD	CommonGen	Quality
CoT		0.3544†	0.6098†	0.7222†	0.6165†	0.5757†
AutoGPT		0.4485†	0.4809†	0.7353†	0.5972	0.5655†
GPTSwarm		0.2368†	0.4969†	0.7096†	0.6222†	0.5163†
AgentVerse		0.2977†	**0.7256**†	0.7587†	0.5399†	0.5805
MacNet-Chain		0.6632	0.3720	**0.8056**	0.5903	0.6078
MacNet-Star		0.4456†	0.5549†	0.7679†	0.7382†	0.6267
MacNet-Tree		0.3421†	0.4878†	0.8044	**0.7718**†	0.6015
MacNet-Mesh		0.6825	0.5122†	0.7792†	0.5525†	0.6316†
MacNet-Layer		0.2780†	0.4939†	0.7623†	0.7176†	0.5629†
MacNet-Random		**0.6877**	0.5244†	0.8054	0.5912	**0.6522**†

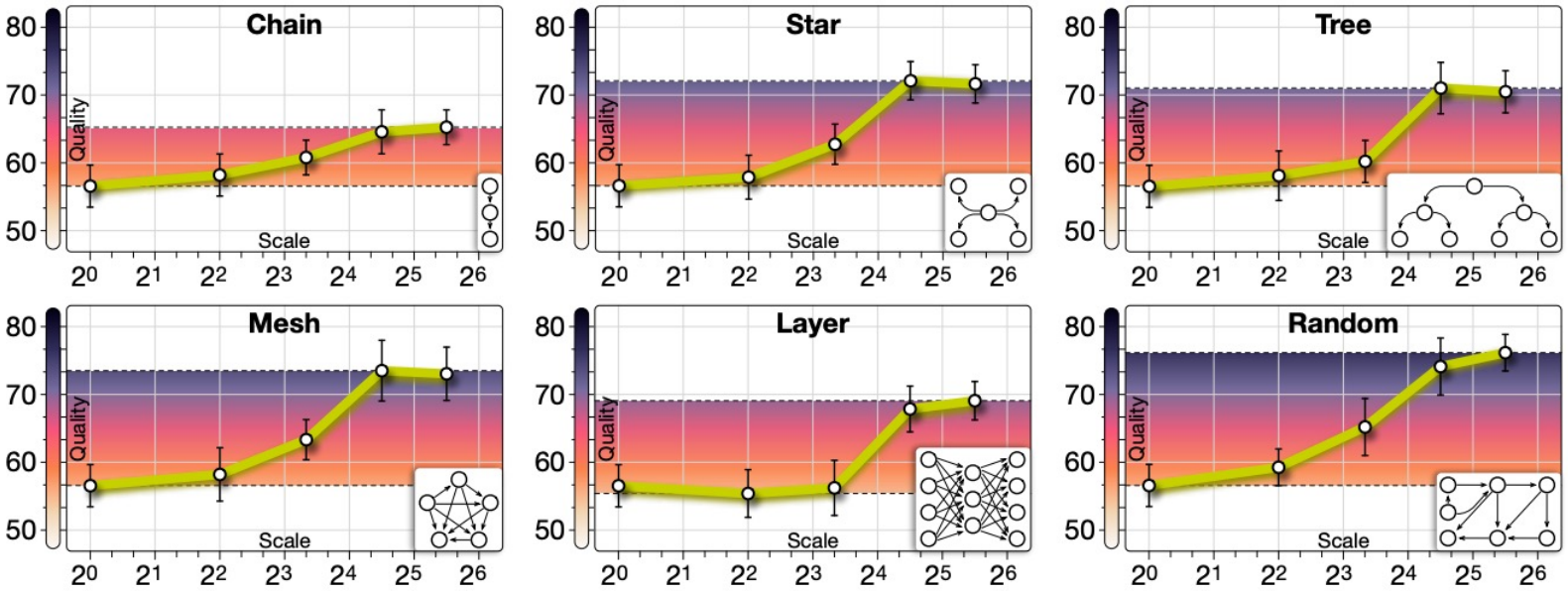






Evaluation

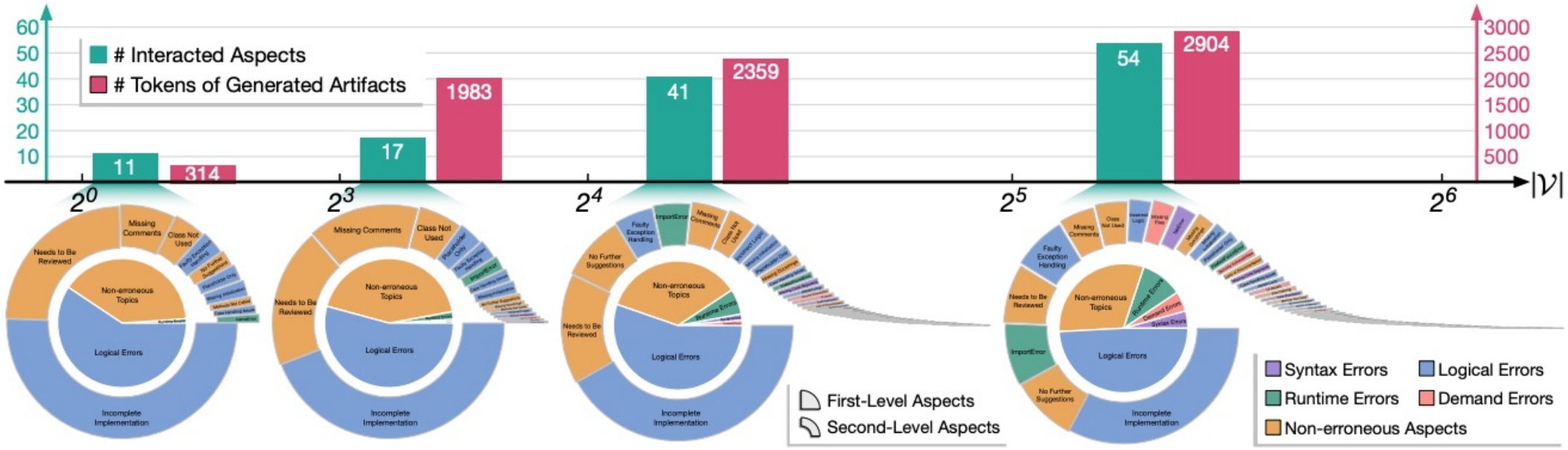
Could a Collaborative Scaling Law be Observed?



- **성능의 sigmoid형 성장**: 에이전트 규모가 커짐에 따라 품질이 완만하게 시작하여 급격히 상승한 후 포화되는 패턴
- **Premature Emergence 발생**: 신경망 모델의 파라미터 확장보다 훨신 적은 규모에서 성능 비약이 나타남
- **토폴로지별 효용성 차이**: 동일 규모에서도 토폴로지에 따라 성능 도달 속도와 임계점이 상이함

Evaluation

What Factors Might Contribute to Collaborative Emergence?



- **검토 항목의 질적 폭발**: 규모 확장 시 대화 주제가 단순 오류 수정을 넘어 수십 개의 세부 품질 항목으로 급증함
- **Long-tail 문제 해결**: 대규모 샘플링을 통해 발생 빈도가 낮은 희귀하고 미세한 결함들까지 포착하여 정제함
- **Critic-Actor 선순환**: Critic의 정교한 지적을 Actor가 93.1%의 높은 확률로 반영하여 결과물의 완성도를 극대화함

Conclusion

- MACNET은 그래프 기반으로 에이전트의 상호작용을 구조화하여 1000개 이상의 에이전트 협업을 지원함
- 실험 결과, 불규칙한 토폴로지가 규칙적인 토폴로지보다 더 높은 성능을 보임
- 에이전트 수 증가에 따라 성능이 로지스틱 형태로 향상되는 Collaborative Scaling Law를 제안함
- 다만 성능 향상에는 확장 한계가 존재하며, 에이전트 협업은 재학습 없이 추론 단계에서 성능을 높이는 대안이 될 수 있음

Diversity Collapse in Multi-Agent LLM Systems: Structural Coupling and Collective Failure in Open-Ended Idea Generation

Nuo Chen[1] **Yicheng Tong**[1] **Yuzhe Yang**[2] **Yufei He**[1]
Xueyi Zhang[2] **Qian Wang**[1] **Qingyun Zou**[1] **Bingsheng He**[1]

[1]National University of Singapore [2]The Chinese University of Hong Kong, Shenzhen

ACL 2026 Findings

Introduction

Limitations of Existing Multi-Agent Ideation

- Open-ended idea generation에서는 하나의 정답보다 다양한 가정과 해결 경로를 탐색하는 능력이 중요함]
- 기존 Multi-Agent 연구는 서로 다른 역할과 관점을 부여하면 아이디어 공간이 확장된다고 가정함 (but, 대부분의 에이전트가 동일한 base 모델과 학습 분포를 공유함 -> 새로운 관점을 만들기보다 공유된 편향과 유사한 아이디어를 반복할 수 있음)
- When does multi-agent collaboration lead to diversity collapse?에 대해 충분히 분석되지 않음

Three Levels of Diversity Collapse

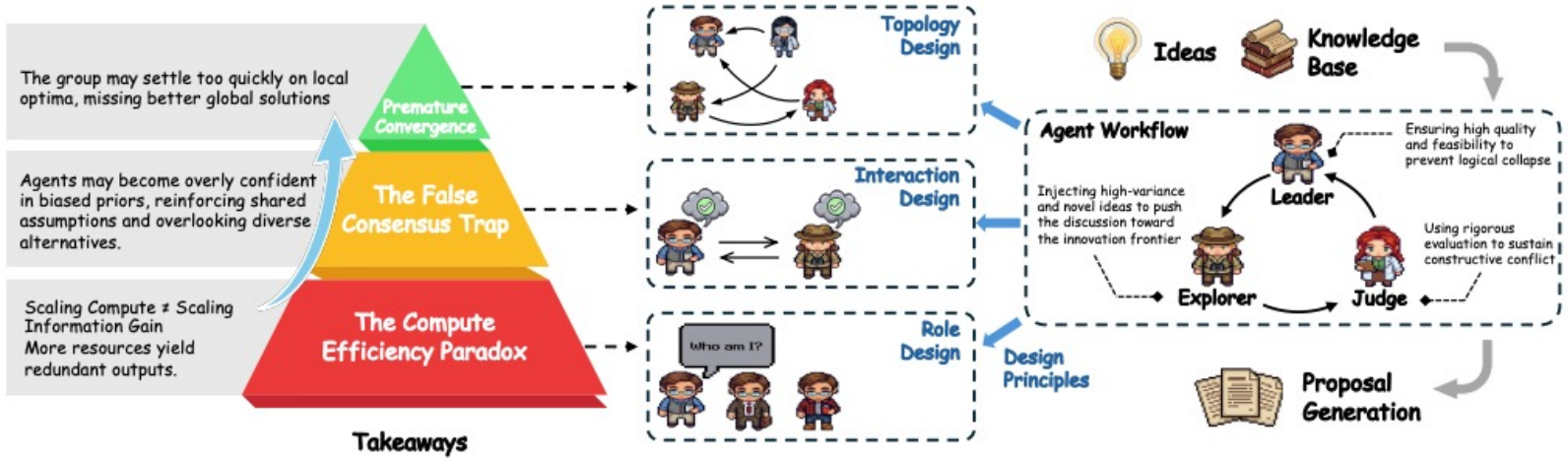
(1) **Model Intelligence:** 모델 성능과 표현의 완성도는 향상되지만, 생성된 아이디어의 의미적 유사도도 함께 증가

(2) **Agent Cognition**: 서로 다른 역할을 부여해도 공유된 편향과 권위 중심 상호작용으로 인해 독립적 비판보다 합의가 강화

(3) **System Dynamics:** 에이전트 수와 연결 밀도가 증가할수록 탐색 범위가 넓어지기보다 유사한 경로에 조기 수렴

Methodology

The Multi-Agent Ideation Pipeline



- **Role Instantiation**: 에이전트들은 시스템 프롬프트를 통해 과학 위원회를 모방하는 고유한 페르소나를 부여받음
- **Iterative Deliberation**: 에이전트들은 다회차 대화에 참여하여 관점의 충돌, 전제에 대한 비판, 개념의 정교화를 이끌어냄
- **Proposal Synthesis**: 편집자 에이전트가 비구조화된 토론 이력을 구체적인 과학적 결과물로 변환함

Methodology

On the Validation of Diversity

- **Effective Diversity (Vendi Score)**: Kernel 행렬의 엔트로피를 활용해 시스템이 semantic space에서 얼마나 많은 유효한 고유 모드들을 효과적으로 탐색하고 있는지 측정함
- **Structural Disorder**: 개별 결과물과 그룹 평균 사이의 유사도를 통해 시스템이 단일 의견으로 쏠리는 Echo Chamber 현상을 겪고 있는지, 아니면 다원적 관점을 유지하고 있는지 진단함
- **Semantic Dispersion (PCD)**: 제안서 간의 평균 쌍별 코사인 거리를 계산하여 semantic mode들이 representation space 상에서 얼마나 넓은 범위로 확산되어 있는지 기하학적 거리를 측정함
- **Lexical Uniqueness**: IDF 가중치 n-gram을 통해 표면적인 어휘 중복을 탐지함으로써, 의미적 다양성이 단순히 표현만 바꾼 중복 아이디어에 의한 것이 아닌지 검증함

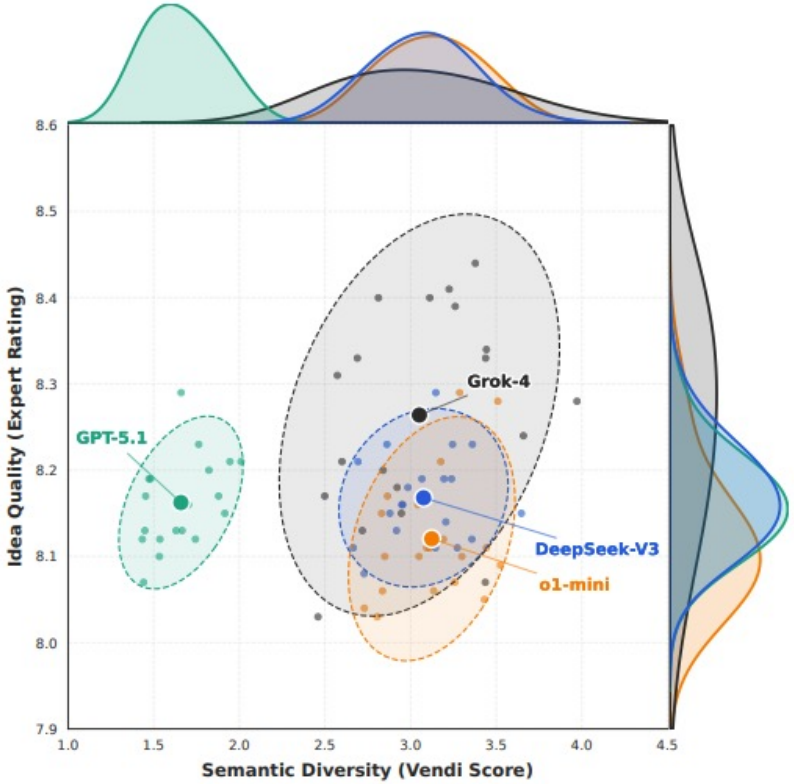
Metric	Human Agreement (%)
Vendi Score	87%
$1 - \phi$	82%
PCD	81%

<- Human Evaluation 기반 일치도 평가 결과

Methodology

The Intelligence Landscape: Quality vs. Diversity

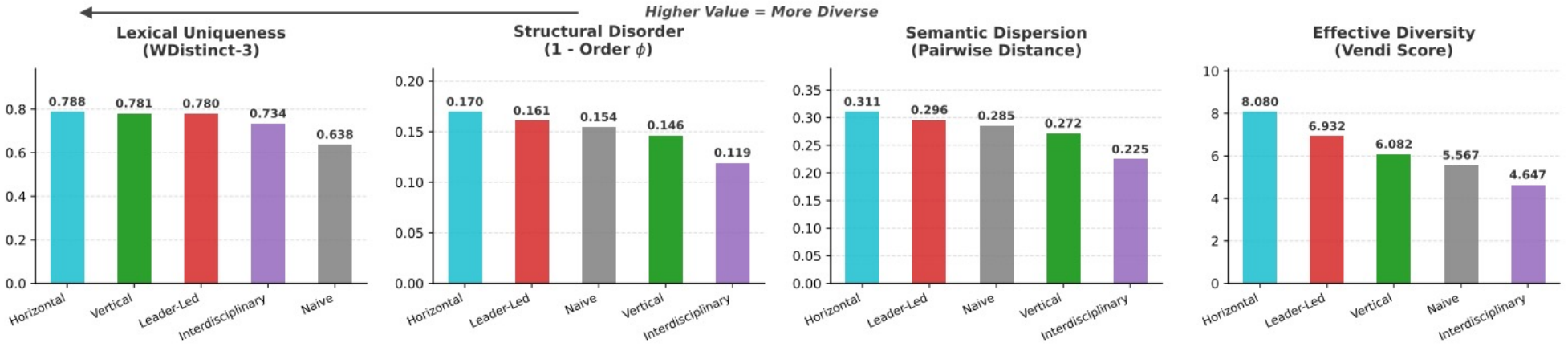
1. Alignment에 의한 다양성 압축: 품질 향상 없이 아이디어 범위만 제한하는 '검역기' 작용으로 다양성이 억제됨
2. 내재적 분산 확대의 한계: 무작위성을 높여 얻은 다양성은 품질의 변동성이 너무 커서 실용성이 낮음
3. 품질 제한 요소의 변화: 개별 모델의 지능은 이미 충분하므로, 내재된 다양성을 상호작용 중에 잃지 않게 지키는 것이 관건임



Cognition: Authority-Induced Collapse

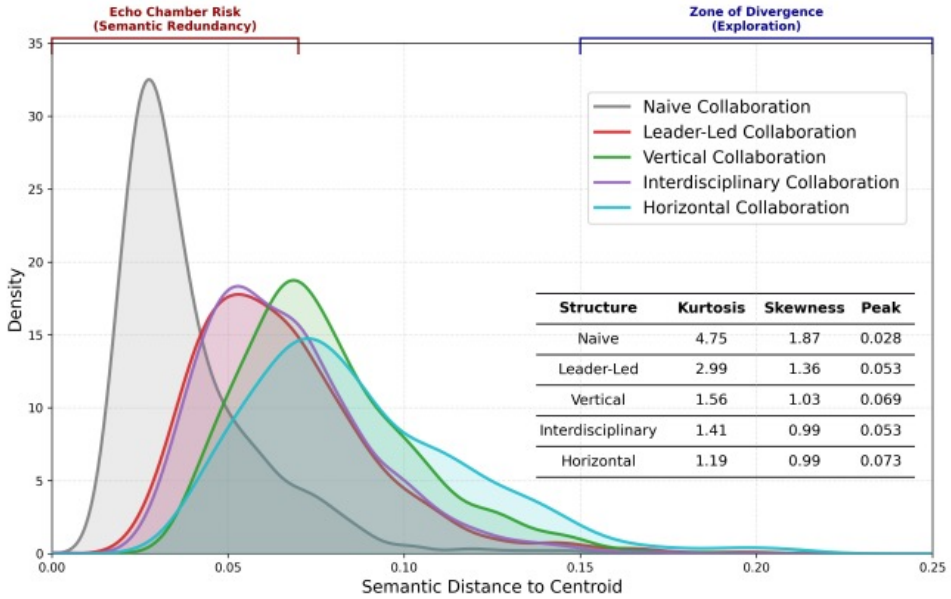
Quantitative Analysis

1. **Naïve Collaboration**: 정의된 역할이나 계층 없이 상호작용
2. **Leader-Led Collaboration**: 시니어 전문가가 토론을 이끌고, 주니어 에이전트들은 지침을 따름
3. **Horizontal Collaboration**: 초기 경력 연구원 그룹이 시니어 감독 없이 평등하게 협업
4. **Interdisciplinary Collaboration**: 서로 다른 분야의 전문가들이 협업하여 cross-domain 아이디어 합성
5. **Vertical Collaboration**: 시니어 전문가, 중견 연구원, 초기 경력 학자들의 계층적 혼합



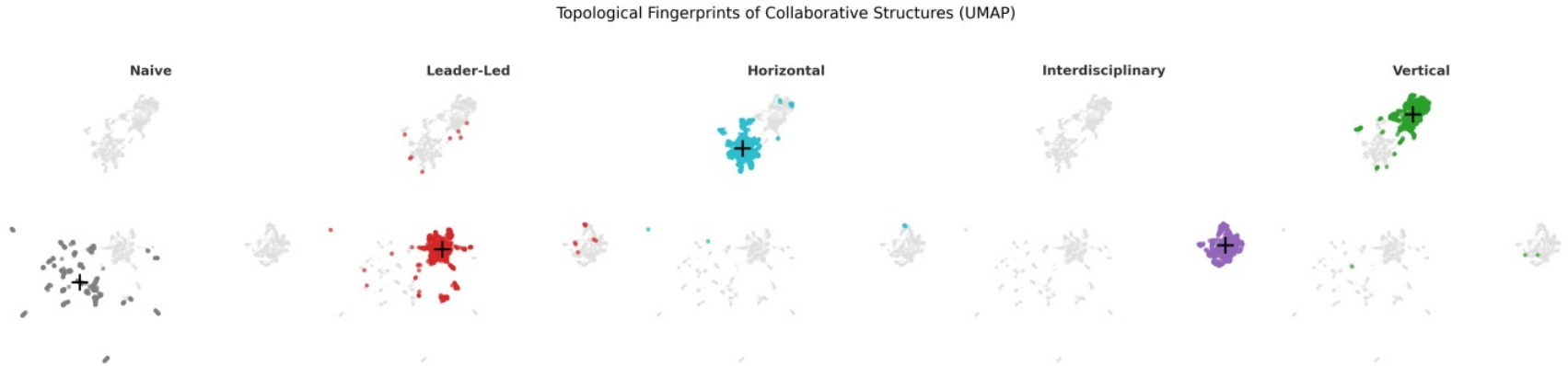
Cognition: Authority-Induced Collapse

Distribution Dynamics & Topological Segregation: Two Semantic Regimes



Structure	Kurtosis	Skewness	Peak
Naive	4.75	1.87	0.028
Leader-Led	2.99	1.36	0.053
Vertical	1.56	1.03	0.069
Interdisciplinary	1.41	0.99	0.053
Horizontal	1.19	0.99	0.073

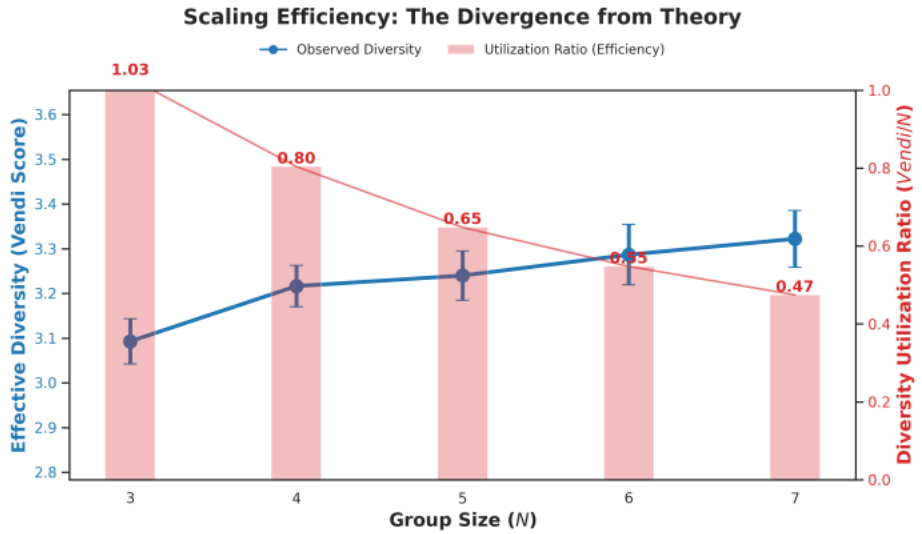
- Gravitational Collapse: Leader-Led / Naïve Collaboration
- Zone of Divergence: Horizontal / Vertical Collaboration
- Vertical 구조는 시니어의 체계성과 주니어의 창의성 사이에서 절묘한 균형을 잡은 성공적인 타협점



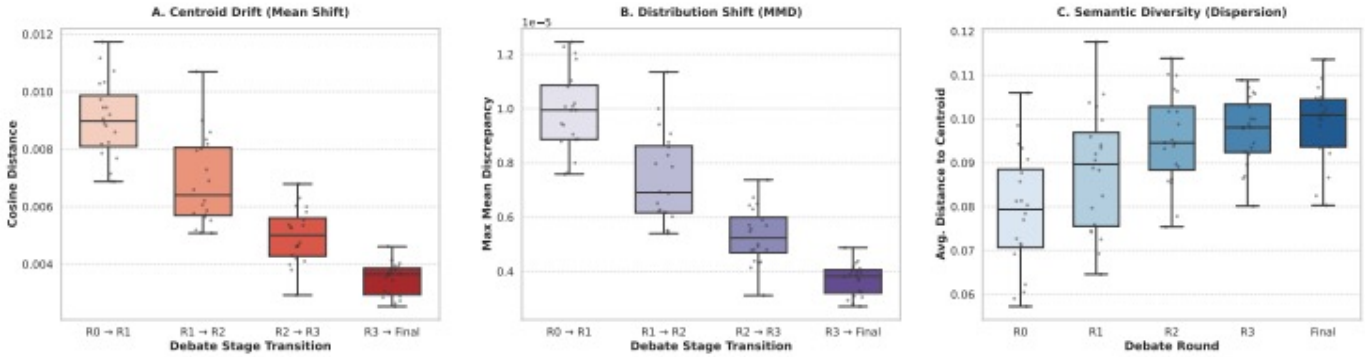
- 전문가 집단은 보기 좋은 결과물은 잘 만들지 몰라도, 그 대가로 혁신에 필요한 수많은 가능성을 다 죽여버림
- 즉 미미한 품질 이득을 얻음

Group Dynamics: Scaling, Evolution, and Topology

Temporal Evolution: Rounds and Trajectories



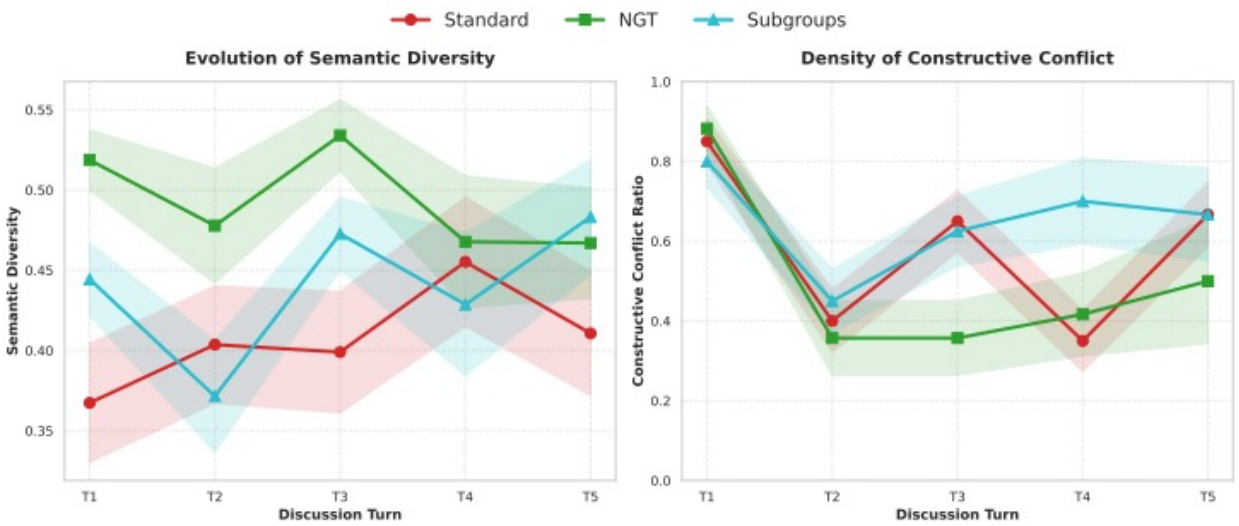
- 절대적 다양성은 단조 증가하지만, 에이전트당 활용도는 급락함
- 새로운 에이전트들은 기존 에이전트들과 점점 더 중복되며, 정보 획득이 빠르게 포화됨



- 전역적 합의는 안정화되는 반면, 국소적 탐색은 확장됨
- 궤적은 무작위적인 환각적 도약이 아닌 일관된 표류를 보여주며, 이는 진정한 세션 내 개선을 의미함

Group Dynamics: Scaling, Evolution, and Topology

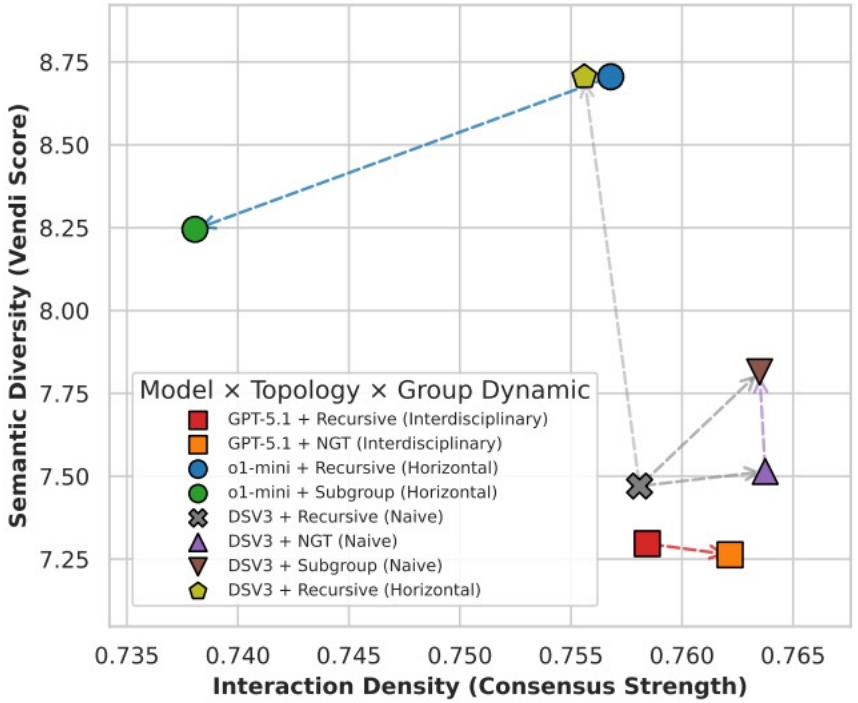
Topology: The Impact of Communication Structure



- NGT는 가장 높은 다양성으로 시작함 – blind writing은 production blocking과 anchoring을 완화함
- Subgroups은 토론 중반에 회복력 스파이크를 보이며 가장 높은 건설적 갈등을 유지함
- Subgroups을 작은 그룹들로 분할하면 locally하게 divergence가 지속적으로 관찰됨

Discussion

Synthesizing the Hierarchical Interplay



- **지능 X 토폴로지**: 표준 모델 (DeepSeek-V3)은 NGT scaffolding의 이점을 얻는 반면, 추론 위주의 모델 (o1-mini)은 방해를 받음
- **정렬 가중치**: 강하게 align된 모델 (GPT-5.1)은 하나의 좁은 합의 영역에 머무르며, 개입을 통해서도 돌파 불가능함
- **붕괴 = 모델의 결함이 아닌 구조의 문제**: 조정을 강제하는 것은 궤적을 동기화시키고 그룹을 단일 경로에 갇히게 만듦

Conclusion

- 단순히 에이전트를 추가한다고 해서 더 다양한 아이디어가 보장되는 것은 아님

- 다양성 붕괴 = model alignment, 계층/역할, 밀집된 커뮤니케이션에 걸친 구조적 결합 <- 에이전트들을 조기 합의로 이끎

- 독립성을 보존하는 설계는 약간의 품질 비용으로 더 높은 다양성을 산출함

Thank You