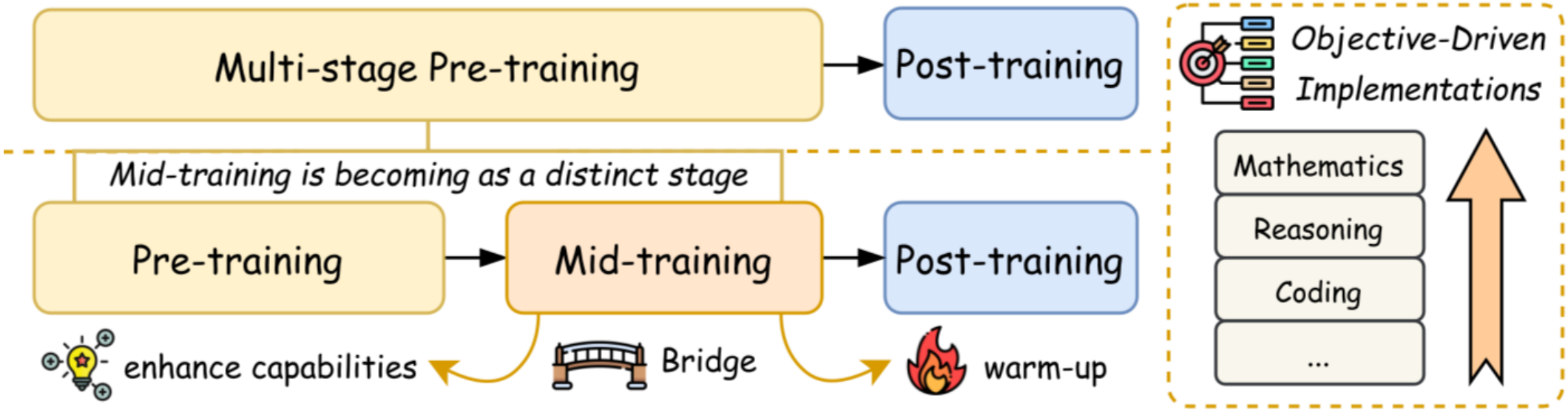
Mid-training



고려대 이정섭

Mid-training

❑ Mid-training 첫 등장

- ❑ 첫 등장 2020년
  - ❑ 대중적으로는 2024년 OpenAI와 Microsoft(Phi)에 의해 유행하기 시작했다고 알려짐.
  - ❑ 실제 첫 등장: 2020년 구글 리서치의 'BLEURT' 논문에서 처음 용어 사용.
  - ❑ BLEURT: 텍스트 생성 품질을 평가하는 인코더 모델(Judge Model).
  - ❑ BERT 모델을 위키피디아 문장으로 워밍업(Warm-up) 시키는 과정을 "Mid-training"이라 명명함.

- ❑ OpenAI 런던 팀: 2024년 7월 "Mid-training" 부서 채용 공고를 내며 용어를 다시 수면 위로 올림.
  - ❑ 흥미로운 연결고리: 이 팀은 DeepMind 출신들(예: Jacob Menick)이 이끌고 있으며, 구글/딥마인드의 연구 유산이 이어진 것으로 추정됨.
  - ❑ 학계의 공식 재도입: Microsoft의 Phi-3.5 (2024) tech report에서 학술적으로 다시 정식 도입됨.

Mid-training

❑ Mid-training의 정의

- ❑ 초기 모호성: 사전 학습(Pre-training)도 아니고, 사후 학습(Post-training)도 아닌 그 중간의 "어떤 것".
  - ❑ OpenAI의 정의: "전통적인 사전 학습과 사후 학습 활동을 모두 포함하는, 모델 개발의 교차(Cross-cutting) 연구 및 엔지니어링."
  - ❑ Yi (01.AI: "데이터 분포의 점진적 변화(Gradual data distribution shifts)를 통해 모델 능력을 향상시키고 컨텍스트 길이를 확장하는 단계."
  - ❑ Olmo 2 (Allen AI, 2025)의 구체화: 학습 후반부의 커리큘럼 러닝(Curriculum Learning) 및 어닐링(Annealing) 단계.

- ❑ 학술적으로 쓰이던 정의
  - ❑ Base Model 훈련과 Instruction Tuning 사이.
  - ❑ 10B ~ 300B 정도 토큰의 “중간 스케일” 데이터셋
  - ❑ 대규모 일반 데이터 학습(Pre-training) 후, 고품질/특수 목적 데이터로 모델을 정제하는 과정.

아직 뚜렷한 정의가 없음 !

Mid-training

Mid-training 목적

- **Domain & Language Extension:**
  - 특정 도메인 지식 주입이나 다국어 능력 강화 (예: Phi-3.5).
- **Long Context Extension:**
  - 긴 문맥을 처리할 수 있도록 위치 임베딩 등을 재조정하며 훈련 (예: Llama 3, Yi).
    - Llama 3 tech report, MiniCPM 논문에서, lr을 줄이고 고품질 데이터를 넣는 순간(어닐링 시작), 정체되어 있던 모델의 성능 그래프가 **수직 상승**하는 현상이 발견
- **Quality & Annealing:**
  - 고품질 데이터(교과서, 과학 논문 등)의 비중을 높여 학습률을 낮추며 마무리하는 단계.
- **Scaling Synthetic Data:**
  - 단순 웹 크롤링 데이터가 아닌, 'Teacher 모델'이 생성하거나 검증(Judge) 과정을 거친 고품질 합성 데이터 집중 학습.

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Introduction

❑ Mid-training

- ❑ 이 용어에 대한 명확하고 널리 합의된 정의가 없기 때문에, 이 연구의 맥락에서 간결하고 엄격한 정의를 제공하고자 함
- ❑ 주요 목적
  - ❑ 이전의 목적 (도메인 & 언어 강화, long context, annealing) 보다도,
  - ❑ 특히, 강화 학습(RL)에 더 잘 적응하고 확장 가능한(RL-scalable) 파운데이션 모델을 만들기 위한 준비 단계로서의 mid-training을 재정의

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Preliminaries

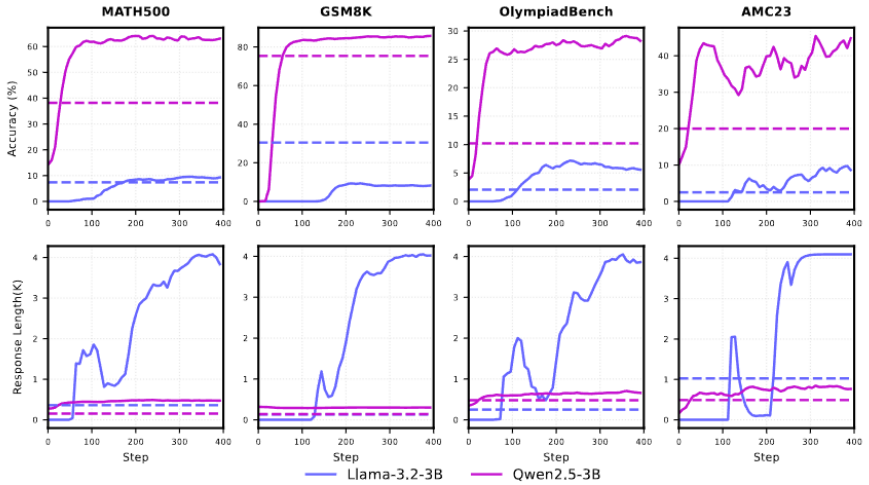


Figure 2 | Training dynamics comparison (downstream performance and the average length of correct responses) between Llama-3.2-3B and Qwen2.5-3B. The dashed line indicates the few-shot evaluation performance and average length of correct responses of the corresponding base models.

Figure 1 | Our strategic mid-training incentivizes Llama's RL scaling, matching Qwen2.5 performance.

Llama-3.2-3B base과 Qwen2.5-3B base를 RL 하였을 때 성능

→ Llama-3.2는 RL 적용 시 Response Length가 굉장히 길어지고 Acc가 낮음, 반면 Qwen2.5는 높은 성능 달성

→ 아마도, PT (혹은 mid-training) 단계에서 무언가 차이가 있을 것이다 !

Llama-3.2 모델을 mid-training을 통해 Qwen2.5 처럼 RL-scalable 하게 만들 수 없을까?

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Key Factors through Controllable Mid-training

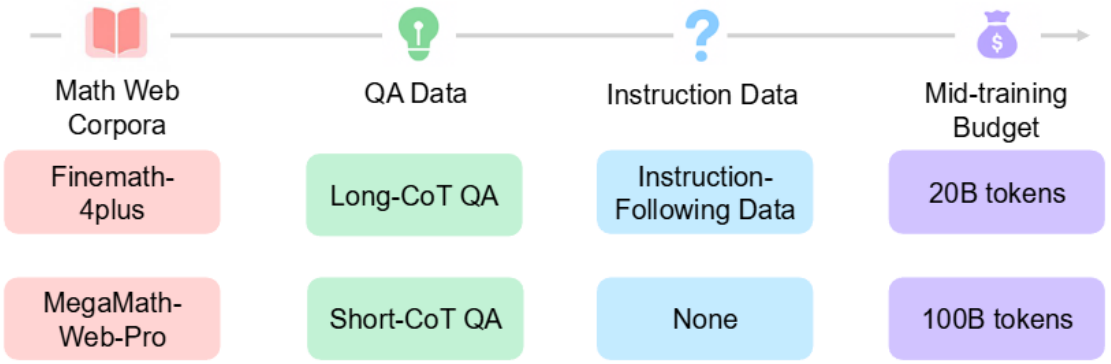


Figure 3 | Potential factors in mid-training that could impact the post-training stage.

20B-토큰 학습 예산 내에서 다양한 데이터 세트 및 학습 구성으로 Llama-3.2-3B-Base를 사용하여 mid-training을 수행

- warmup 없이 코사인 학습률 스케줄러를 사용

- max LR: 3e-5

- min LR: 1.5e-5

- 기본 시퀀스 길이는 8,192이고, 배치 크기는 4백만 토큰

Table 1 | Statistics and Types of different datasets used in our experiments. °We use the `TULU3-sft-personna-instruction-following` subset.

Dataset	Type	# Tokens (B)
FineMath-4plus (Allal et al., 2025)	Math Web Documents	9.57
MegaMath-Web-Pro (Zhou et al., 2025)		13.00
`MegaMath-Web-Pro-Max` (Ours)		73.80
MegaMath-QA (Zhou et al., 2025)	QA (Short-CoT)	5.94
OpenR1-Math-220K (HuggingFace, 2025)	QA (Long-CoT)	1.05
TULU3-sft° (Lambert et al., 2024a)	General Instruction Following	0.01
WildChat (Zhao et al., 2024)		0.29
UltraChat-220K (Ding et al., 2023a)		0.51

IF의 경우, User:{query}₩nAssistant:{content}

QA의 경우, User:{query}₩nAssistant:<think>₩n {사고과정}</think>₩n{content}

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Key Factors through Controllable Mid-training **(Web corpora를 사용했을 때의 성능 분석)**

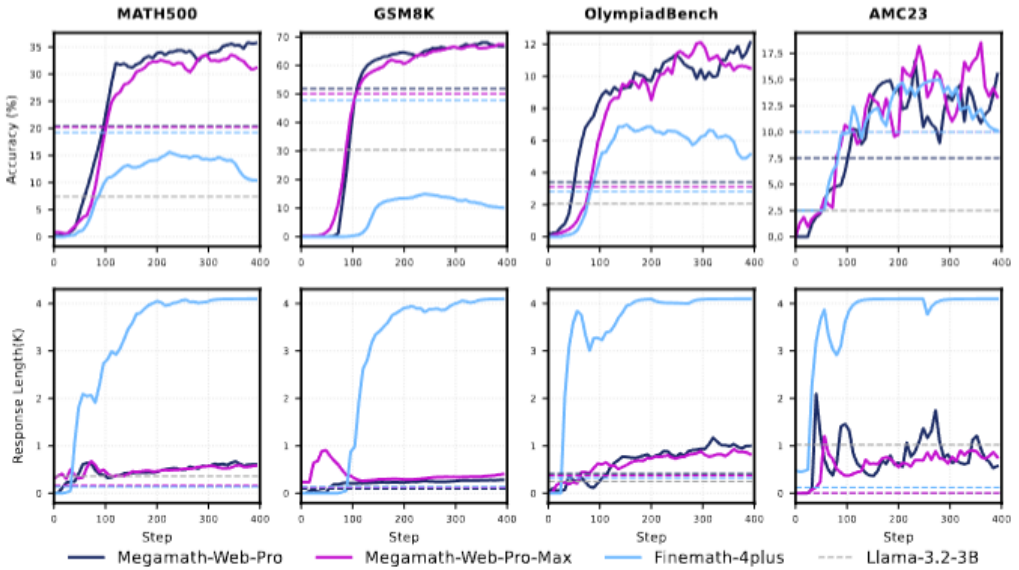


Figure 5 | The effect of different math web corpora during mid-training. We performed mid-training on each corpus with a 20B-token training budget.

정확도 측면 (상단 그래프)

- **MegaMath-Web-Pro** & **MegaMath-Web-Pro-Max**: 두 코퍼스로 미드-트레이닝한 모델은 Llama-3.2-3B 기본 모델(회색 점선)에 비해 모든 벤치마크에서 강화 학습을 통해 상당히 높은 정확도 향상을 보임
- Finemath-4plus: 이 코퍼스로 미드-트레이닝한 모델은 다른 두 MegaMath 코퍼스에 비해 정확도 향상이 미미하거나 오히려 불안정한 모습을 보임

응답 길이 측면 (하단 그래프)

- **MegaMath-Web-Pro** 및 **MegaMath-Web-Pro-Max**: 이 모델들은 강화 학습 훈련 중에도 응답 길이가 안정적으로 유지되었으며, 대부분 1,000토큰(1K) 이하로 합리적인 수준의 증가를 보였습니다. 이는 안정적인 RL 훈련 과정을 나타냄
- **Finemath-4plus**: 강화 학습 초반부터 응답 길이가 급격하게 증가하여 최대 응답 길이(약 4,096토큰)에 도달한 후 유지되는 불안정한 행동을 보임. "Solution" 같은 반복적인 문장이 "boxed{}" 뒤에 오는 비정상적인 출력을 관찰됨.

품질이 낮은 코퍼스(예: Finemath-4plus)는 RL 성능 향상에 거의 기여하지 못하며, 오히려 훈련을 불안정하게 만들고 모델이 비정상적인 긴 응답을 생성하게 할 수 있음

고품질 수학 사전 학습 코퍼스(예: MegaMath-Web-Pro, MegaMath-Web-Pro-Max)는 RL scaling에 매우 중요한 역할을 함

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Key Factors through Controllable Mid-training **(CoT 데이터를 사용했을 때의 성능 분석)**

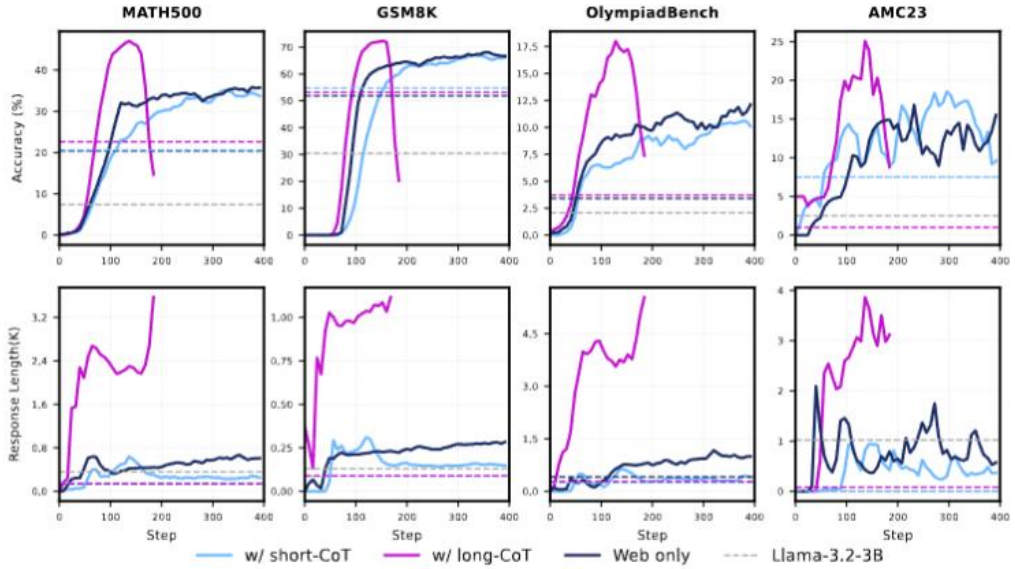


Figure 6 | Impact of incorporating CoT data with varying characteristics during mid-training (9:1 mixture ratio). The figure also illustrates performance and average lengths of correct responses for Llama-3.2-3B-Base and its mid-trained variants for reference (in dashed line with different colors).

- long-CoT 데이터는 모델의 추론 깊이를 향상시킬 수 있지만, 이로 인해 모델 응답이 과도하게 길어지고 RL 훈련이 불안정해질 수 있음

- short-CoT 데이터는 Web only 데이터에 비해 RL 성능 개선에 큰 영향을 미치지 못했음

정확도 측면 (상단 그래프)

- **Web only (남색 선) 및 w/ short-CoT (하늘색 선):** 두 경우 모두 RL 훈련이 진행됨에 따라 꾸준히 정확도가 상승하며 비교적 안정적인 성능을 보임. short-CoT 데이터를 추가한 것이 Web only에 비해 초기 base model의 성능은 약간 높지만, RL 훈련 후에는 큰 차이를 보이지 않거나 오히려 약간 낮은 경우도 있음.
- **w/ long-CoT (자주색 선)**: 초기 훈련 단계(약 150-200 스텝)에서는 다른 두 경우보다 훨씬 빠르게 높은 정확도에 도달하며 우수한 성능을 보임. 하지만 이 시점 이후로 모든 벤치마크에서 정확도가 급격히 하락하여 불안정성을 드러냄. 이는 long-CoT 데이터가 초기에는 강력한 추론 능력을 부여하지만, RL 훈련의 후반부에서는 오히려 성능 저하를 일으킬 수 있음을 시사함.
- 점선 (Llama-3.2-3B 및 mid-trained variants): 각 base model의 초기 성능을 나타내며, RL 훈련을 통해 대부분의 경우 base model보다 성능이 향상됨

응답 길이 측면 (하단 그래프)

- **Web only (남색 선) 및 w/ short-CoT (하늘색 선):** 이 모델들은 RL 훈련이 진행됨에 따라 응답 길이가 완만하게 증가하거나 안정적으로 유지됨
- **w/ long-CoT (자주색 선):** 정확도가 급락하는 시점과 거의 동시에 응답 길이가 급격하게 증가하는 경향을 보임. ➔ 모델이 필요 이상으로 장황한 답변을 생성하거나 반복적인 출력을 하는 등 불안정한 행동을 보인다는 것을 나타냄

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Key Factors through Controllable Mid-training **(IF 데이터 포함 여부 영향)**

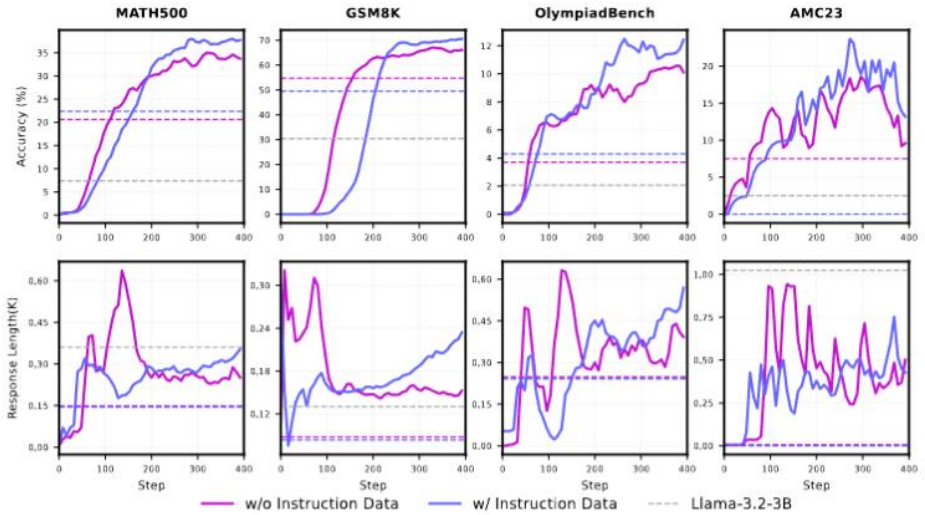


Figure 7 | Impact of incorporating instruction-following data during mid-training with a mixture of web, short-CoT and instruction data in a ratio of 89: 10: 1 . The maximum response length is 4,096. The figure also illustrates performance and average lengths of correct responses for Llama-3.2-3B-Base and its mid-trained variants for reference (in dashed line with different colors).

Web corpus와 short-CoT 데이터에 instruction-following 데이터를 포함하는 것은 RL 훈련 중 응답 길이의 안정성을 높여 모델이 과도하게 길거나 짧은 응답을 내놓는 것을 방지하는 데 도움을 줌

Web corpus와 short-CoT 데이터에 instruction data를 혼합하여 훈련했을 때의 결과 비교 (RL 시 maximum response length 4,096)

정확도 측면 (상단 그래프)

- 파란색 선(instruction data 포함)이 훈련 초기부터 더 빠르게 성능이 향상되며, 최종적으로 자홍색 선(instruction data 미포함)보다 높은 정확도를 보임. 약 200 스텝 이후부터 명확한 성능 차이가 나타남.

➔ instruction-following data를 mid-training에 포함하면, short-CoT 데이터의 잠재력을 발휘하게 하여 RL 훈련 후 성능을 전반적으로 향상시키는 데 기여함.

응답 길이 측면 (하단 그래프)

- 자홍색 선(instruction data 미포함)은 훈련 초중반에 응답 길이가 급격하게 증가했다가 감소하는 불안정한 모습을 보임.

- 반면, 파란색 선(instruction data 포함)은 응답 길이가 더 부드럽고 점진적으로 증가하며, 전체적으로 더 안정적인 패턴을 나타냄.

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Key Factors through Controllable Mid-training **(IF 데이터 포함 여부 영향)**

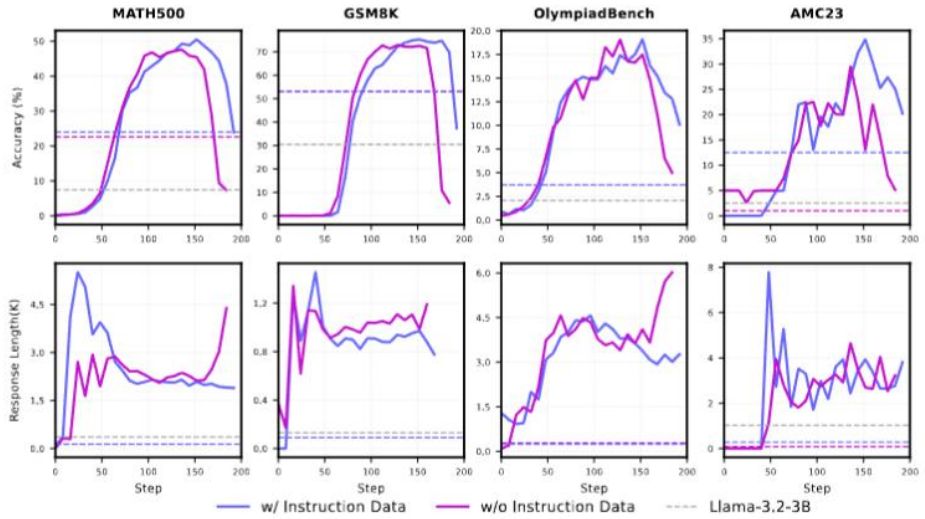


Figure 8 | Impact of incorporating instruction-following data during mid-training with a mixture of web, long-CoT and instruction data in a ratio of 89: 10: 1. The maximum response length is 8,192. The figure also illustrates performance and average lengths of correct responses for Llama-3.2-3B-Base and its mid-trained variants for reference (in dashed line with different colors).

mid-training 시 long-CoT 데이터와 함께 instruction 데이터를 사용하는 것은 RL 훈련 초반에 약간의 성능 이점을 줄 수 있지만, long-CoT로 인한 RL 성능의 전반적인 하락과 응답 길이의 과도한 증가와 같은 불안정성으로 성능이 하락함

➔ long-CoT 데이터 사용 시, RL 훈련의 불안정성을 해결하기 위해 프롬프트 템플릿 수정이나 최대 응답 길이 스케줄러와 같은 다른 전략이 필요함

Web corpus와 long-CoT 데이터에 instruction data를 혼합하여 훈련했을 때의 결과 비교 (RL 시 maximum response length 8,192)

정확도 측면 (상단 그래프)

- instruction 데이터를 포함한 경우(파란색 선)는 대략 150스텝 이후에 성능 개선을 보이는 경향이 있음. (파란색 선이 자홍색 선보다 약간 더 높게 시작)
- 하지만 instruction 데이터를 추가했음에도 불구하고, 훈련 후반(약 150스텝 이후)에는 모든 벤치마크에서 성능이 급격히 하락하는 경향을 보임
- 파란색 선과 자홍색 선 모두 성능이 저하되는 패턴은 유사하게 나타남. ➔ instruction 데이터가 long-CoT 데이터로 인해 발생하는 전반적인 RL 성능 저하를 막지는 못했다는 것을 의미함

응답 길이 측면 (하단 그래프)

- long-CoT 데이터를 사용했을 때 응답 길이가 매우 빠르게, 증가함. 두 경우(instruction 데이터 포함/미포함) 모두 훈련 스텝이 진행됨에 따라 응답 길이가 빠르게 최대치(8K 토큰)에 도달하거나 그에 육박하는 수준으로 급증하는 것을 볼 수 있음
- short-CoT 환경에서는 instruction 데이터가 응답 길이를 안정화하는 데 기여했지만, long-CoT 환경인 Figure 8에서는 instruction 데이터를 추가하더라도 이러한 급격한 응답 길이 증가를 방지하지 못함. ➔ long-CoT 데이터가 RL 훈련에 가져오는 불안정성이 instruction 데이터만으로는 해결하기 어려운 더 근본적인 문제임을 시사

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Key Factors through Controllable Mid-training **(Maximum Response Length 영향)**

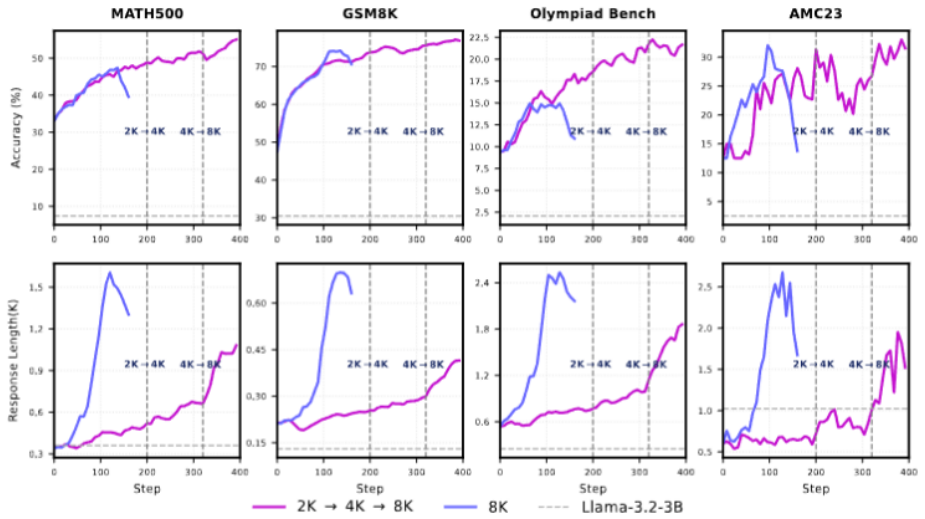


Figure 10 | Impact of the maximum length scheduler on the model response. The figure also illustrates performance and average lengths of correct responses for Llama-3.2-3B-Base in a dashed line.

RL 학습 시 Response Length를 점진적으로 늘려야 함.

짧고 간단한 추론부터 시작하여 점진적으로 더 긴 추론 구조를 학습하도록 유도. 학습 초기에 모델이 "길이 폭주"로 인한 불안정성(repetitive outputs, early collapse)에 빠지는 것을 방지하고, 안정적인 학습 경로를 제공함

미드-트레이닝에서 Long CoT 데이터를 사용했을 때 발생할 수 있는 RL 훈련 불안정성을 완화하는 중요한 요소

고정 8K (파란색): RL 훈련 내내 최대 응답 길이를 8192 토큰(8K)으로 고정했을 때의 성능 및 응답 길이 변화

점진적 스케줄(2K → 4K → 8K) (자홍색): 훈련 초기 200 스텝까지는 2048 토큰(2K), 이후 320 스텝까지 4096 토큰(4K), 그 이후부터는 8192 토큰(8K)으로 최대 응답 길이를 점진적으로 늘려감

정확도 측면 (상단 그래프):

- 고정 8K 방식 (파란색 선): 훈련 초반에는 정확도가 빠르게 상승하지만, 특정 스텝(대략 100~200 스텝) 이후 급격히 하락하거나 매우 불안정한 양상을 보임. ➔ 이는 모델이 과도하게 긴 응답을 생성하면서 훈련이 붕괴되는 현상
- 점진적 스케줄 방식 (자홍색 선): 모든 벤치마크에서 훈련 스텝이 진행됨에 따라 꾸준하고 안정적으로 정확도가 향상됨. 최대 길이가 점진적으로 증가함에 따라 모델이 새로운 길이 제약에 적응하며 성능을 계속 개선함

응답 길이 측면 (하단 그래프)

- 고정 8K 방식 (파란색 선): 훈련 초반에 평균 응답 길이가 급격하게 증가하여 최대 길이(8K)에 도달한 후, 오히려 응답 길이가 줄어들거나 불안정하게 변동하는 패턴을 보임 ➔ 이는 모델이 유의미한 추론 없이 단순히 최대 길이를 채우는 반복적인 출력을 생성하는 문제(verbosity 및 instability)를 겪었음을 시사함.
- 점진적 스케줄 방식 (자홍색 선): 평균 응답 길이가 훈련 스텝에 따라 점진적으로 그리고 안정적으로 증가함. 최대 응답 길이가 2K에서 4K, 4K에서 8K로 늘어나는 시점(200 스텝, 320 스텝)에서 응답 길이가 조금 더 빠르게 증가하는 경향을 보이지만, 전반적으로 제어된 방식으로 길이가 늘어남

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Key Factors through Controllable Mid-training **(Token budget 영향)**

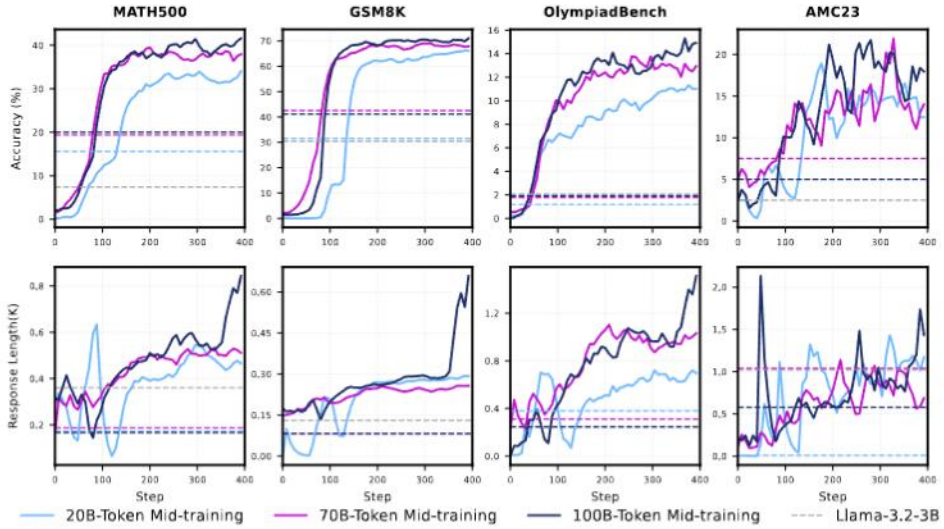


Figure 11 | Impact of scaling up the mid-training budget. The figure also illustrates performance and average lengths of correct responses for Llama-3.2-3B-Base and its mid-trained variants for reference (in dashed lines with different colors).

미드 트레이닝 예산을 늘리는 것은 다운스트림 강화 학습 성능을 일관되게 향상시킴

미드 트레이닝 예산을 늘렸을 때 베이스 모델 평가에서 즉각적인 성능 향상이 나타나지 않더라도, 강화 학습 단계에서는 그 이점이 발현될 수 있음. 즉, 베이스 모델 평가 지표와 RL 단계의 역량 사이에 간극이 있을 수 있다.

하늘색 선은 20B 토큰으로 미드 트레이닝된 모델

자홍색 선은 70B 토큰으로 미드 트레이닝된 모델

짙은 파란색 선은 100B 토큰으로 미드 트레이닝된 모델

정확도 측면 (상단 그래프):

- 미드 트레이닝 예산을 늘릴수록(20B → 70B → 100B 토큰) 강화 학습을 통한 최종 정확도가 전반적으로 더 높아짐

응답 길이 측면 (하단 그래프)

- 미드 트레이닝 예산을 늘릴수록(특히 70B 및 100B 토큰 모델) 강화 학습 훈련 중 응답 길이가 더 안정적이고 점진적으로 증가함.
- 20B 토큰 미드 트레이닝 모델은 훈련 초기에 응답 길이가 급격하게 변동하는 불안정한 모습을 보임. 하지만, 100B 토큰 모델은 안정적으로 증가함

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Recipe: OctoThinker-Base - Branching Reasoning Foundations via 2-Stage Mid-training

First Stage

- Stable Stage: Building Strong Reasoning Foundation

목적: RL 훈련을 위한 견고하고 안정적인 기초를 다지는 단계

방법:

- 데이터: 주로 MegaMath-Web-Pro-Max (고품질 수학 웹 코퍼스)와 DCLM-Baselines (고품질 사전 학습 코퍼스) 등 고품질 웹 코퍼스에 의존함. 여기에 소량의 합성 데이터(synthetic data)가 보충됨.

- 200억 토큰(200B tokens) 이라는 대규모 토큰 예산으로 학습

- 학습률 스케줄러: constant learning rate로 학습

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- Recipe: OctoThinker-Base - Branching Reasoning Foundations via 2-Stage Mid-training

Second Stage

- Decay Stage: Seeking Perfect Blend for RL Scaling

목적: 학습률 감소를 통해 주입되는 데이터의 효과를 증폭시키고, 모델 행동을 다양화하는 데 중점

방법:

- **20억 토큰(20B tokens) 규모로 학습**
- **학습률 스케줄러: 학습률을 점진적으로 감소시키는 cosine decay 사용.**
- **해당 논문에서 mid-training을 세 가지 별개의 data mixtures로 진행하여 실험**
  - **OctoThinker-Long:** long-reasoning 데이터(긴 추론)에 중점을 둠
  - **OctoThinker-Short:** short-reasoning 데이터(짧은 추론)에 중점을 둠
  - **OctoThinker-Hybrid:** long-reasoning과 short-reasoning 데이터를 혼합하여 사용

Table 5 | Specific data mixture for each branch in the decay stage

(a) Long Branch Mixture

Dataset	Weight
DCLM-Baseline	0.05
Instruction Following	0.10
MegaMath-Web-Pro	0.55
Open R1	0.15
AM-DeepSeek-Distilled-40M	0.15

(b) Short Branch Mixture

Dataset	Weight
DCLM-Baseline	0.05
Instruction Following	0.10
MegaMath-Web-Pro	0.55
MegaMath-QA	0.025
OpenMathInstruct2	0.175
NuminaMath1.5	0.10

(c) Hybrid Branch Mixture

Dataset	Weight
DCLM-Baseline	0.05
Instruction Following	0.10
MegaMath-Web-Pro	0.55
OpenMathInstruct2	0.10
NuminaMath1.5	0.10
Open R1	0.10

공통 데이터:
DCLM-Baseline (고품질 사전학습 코퍼스), Instruction Following (IF 데이터), MegaMath-Web-Pro (고품질 수학 웹 코퍼스)

브랜치별 데이터:
Long CoT: Open R1, AM-DeepSeek-Distilled-40M
Shot CoT: MegaMath-QA, OpenMathInstruct2, NuminaMath1.5

OctoThinker: Mid-training Incentivizes Reinforcement Learning Scaling

- OctoThinker-Long vs OctoThinker-shorts vs OctoThinker-hybrid

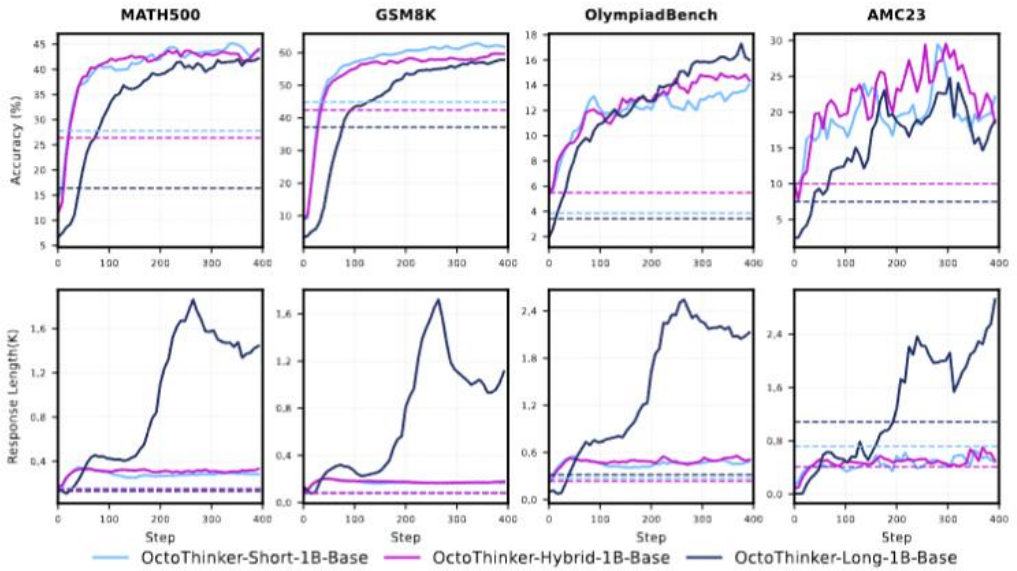


Figure 12 | The RL training dynamics across different branches for OctoThinker-1B series

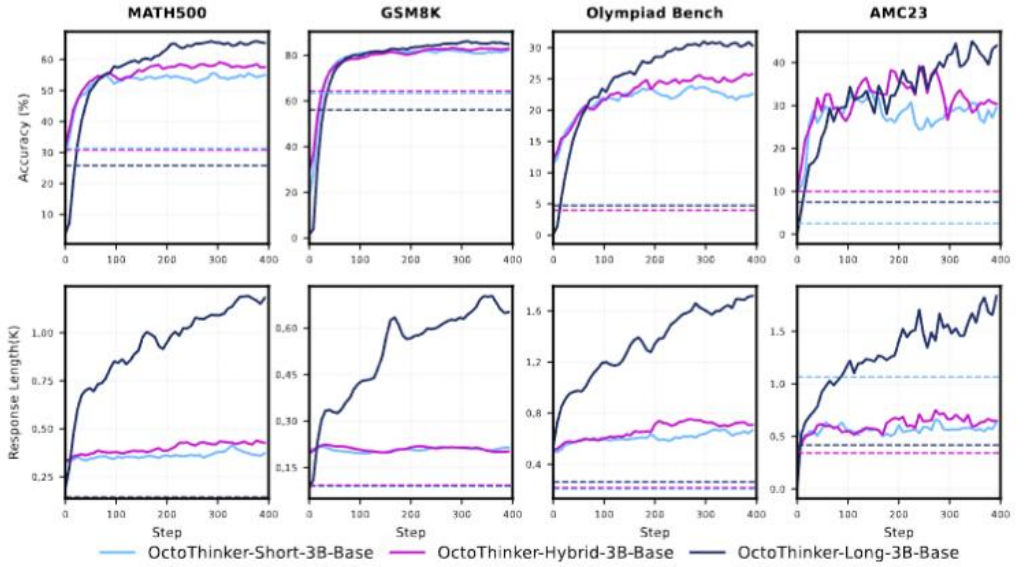


Figure 13 | The RL training dynamics across different branches for OctoThinker-3B series

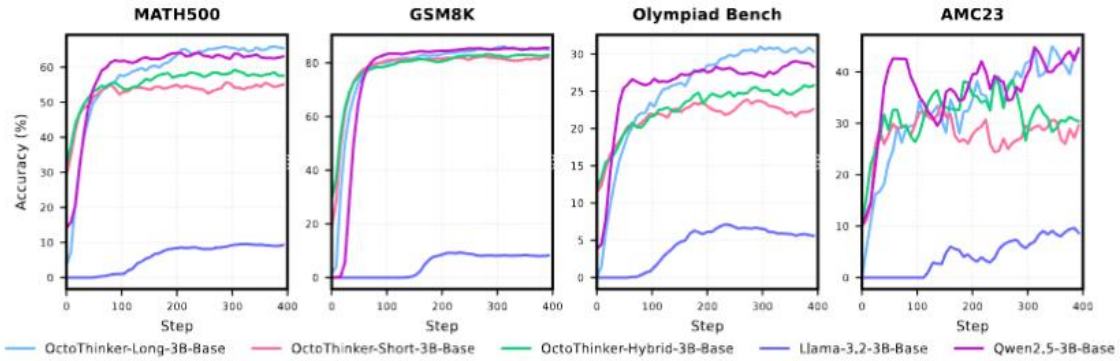


Figure 14 | RL training dynamics among Llama-3.2-3B-Base, OctoThinker series and Qwen2.5-Base.