Closed-Loop Scaling

Autonomous Improvement of LLM and LVLM Reasoning

Yuxi Xie

National University of Singapore, Singapore

Advisor: Associate Professor Min-Yen Kan

Content

- **PART 1**: Introduction
- **PART 2**: Layer One – Inference-time Scaling
- **PART 3**: Layer Two – Training-time Scaling
- **PART 4**: Layer Three – Architectural & Contextual Grounding
- **PART 5**: Reflections & Diagnosis
- **PART 6**: Future Directions & Closing

Content

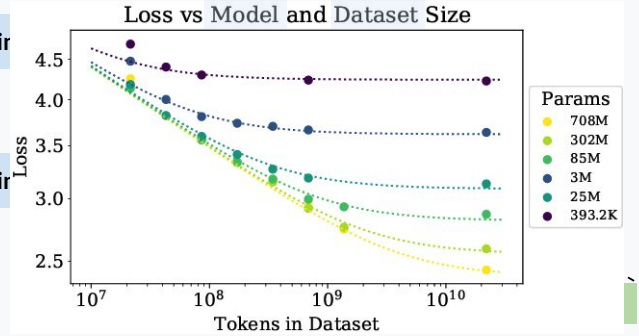
- ❑ **PART 1**: Introduction
- ❑ **PART 2**: Layer One – Inference-time Scaling
- ❑ **PART 3**: Layer Two – Training-time Scaling
- ❑ **PART 4**: Layer Three – Architectural & Contextual Grounding
- ❑ **PART 5**: Reflections & Diagnosis
- ❑ **PART 6**: Future Directions & Closing

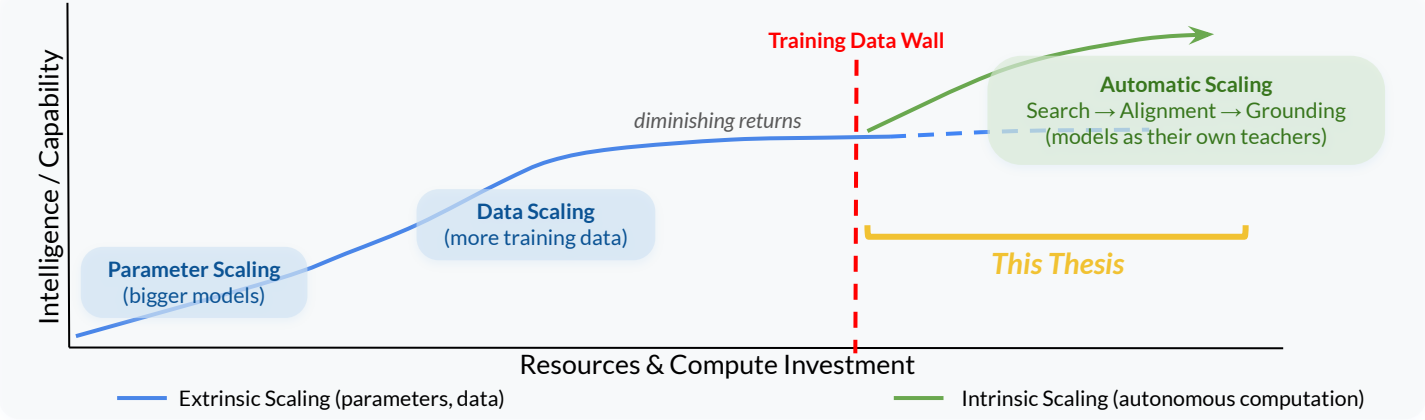
The Scaling Frontier: Towards General Automatic Scaling

- *Brute Force* Scaling Laws: L(N, D)

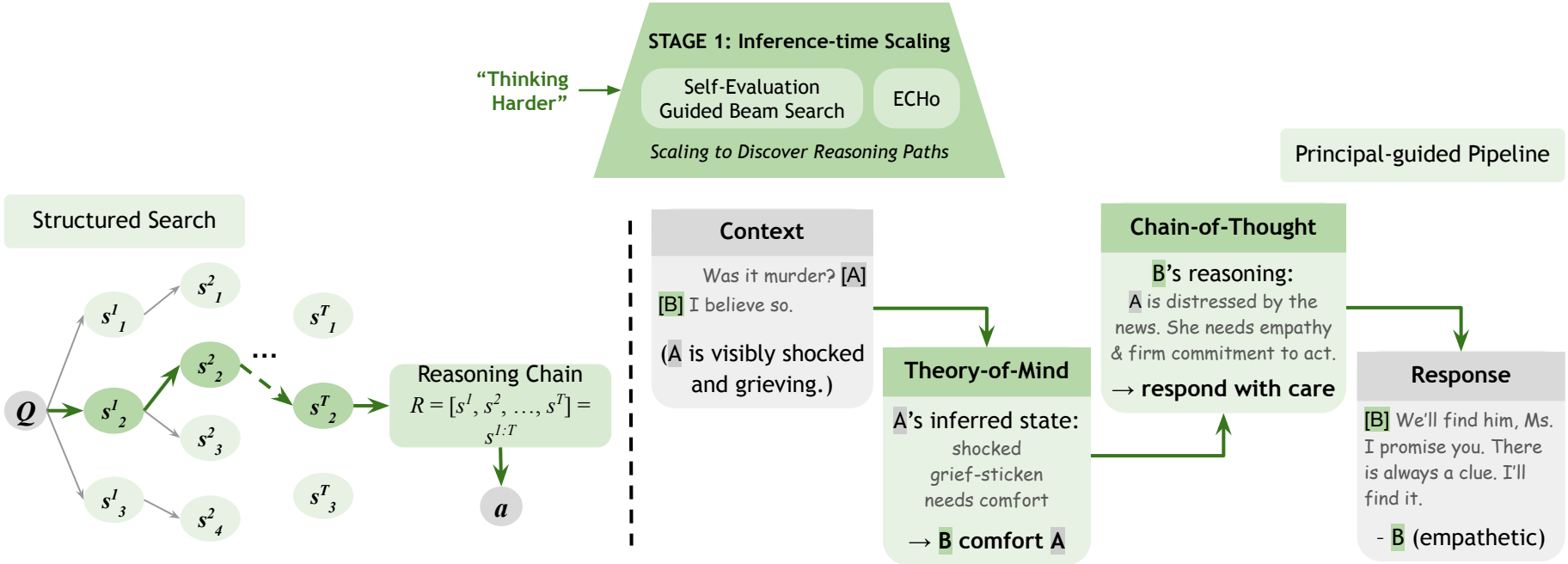
  → Test-time Compute: L(N, D, T)

  → Autonomous Conversion of Computation into Intelligence

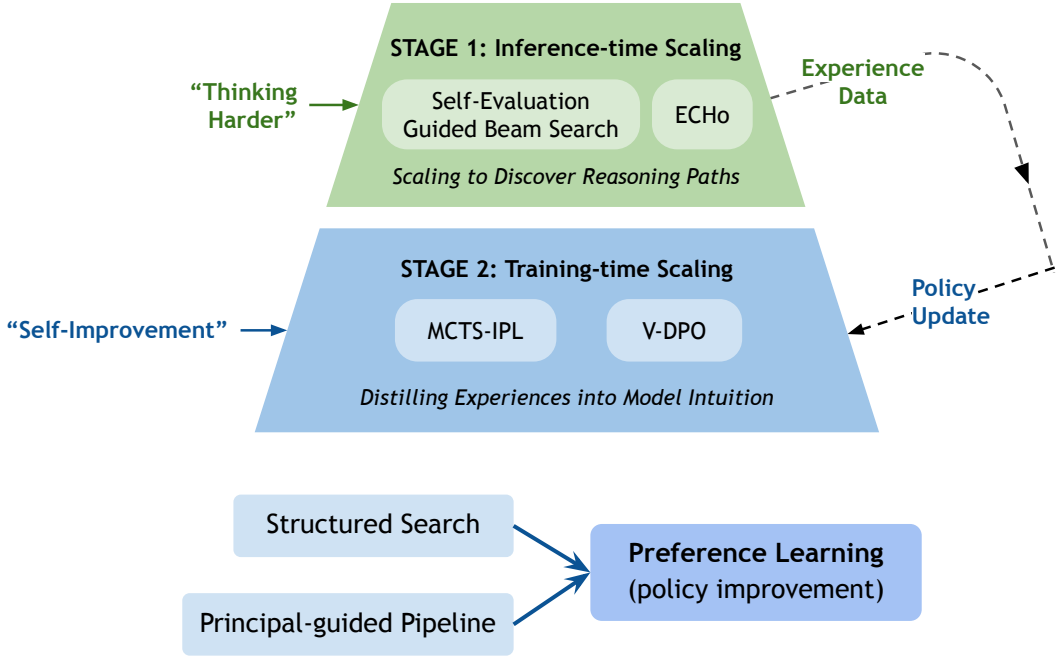




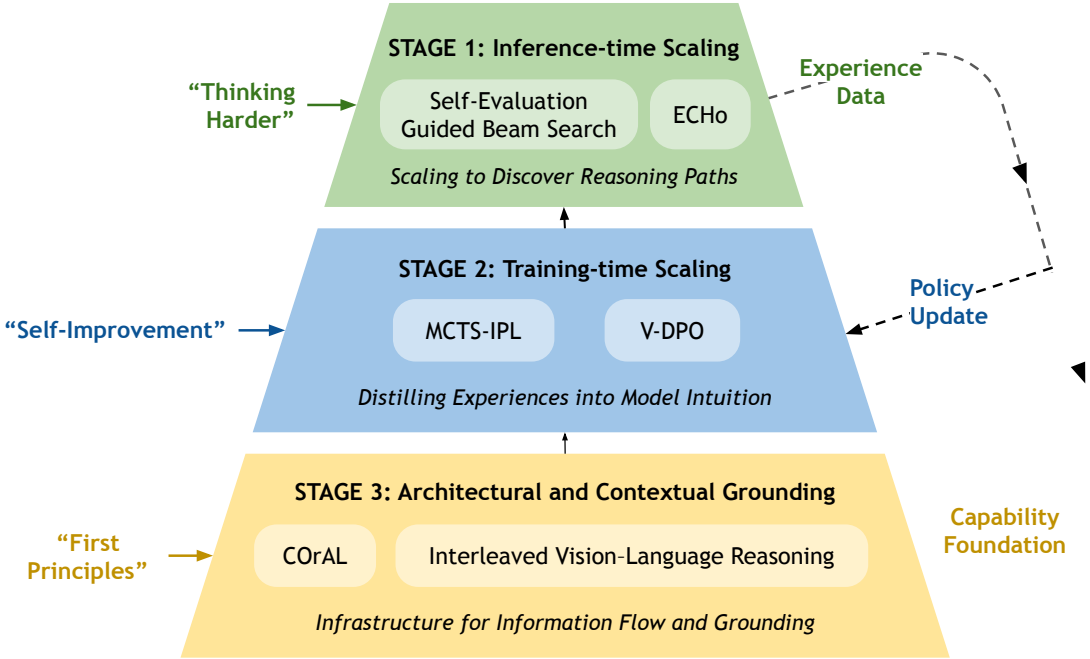
Three Layers: Search



Three Layers: Search → Alignment



Three Layers: Search → Alignment → Grounding

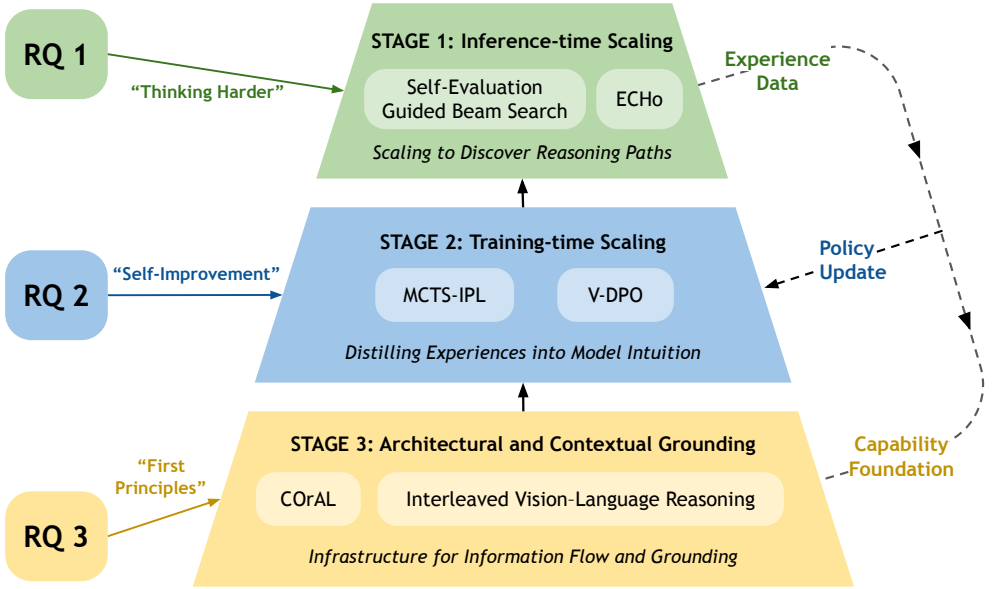


Research Questions

- **RQ1**: Scaling through **Exploration**
  → *inference-time*
  How can we autonomously expand the model capabilities during inference?

- **RQ2**: Scaling through **Distillation**
  → *training-time*
  How can autonomous inference-time gains be internalized into model parameters?

- **RQ3**: Scaling through **Grounding**
  → *architectural & contextual foundation*
  What architectural primitives are required to sustain long-term autonomous evolution?

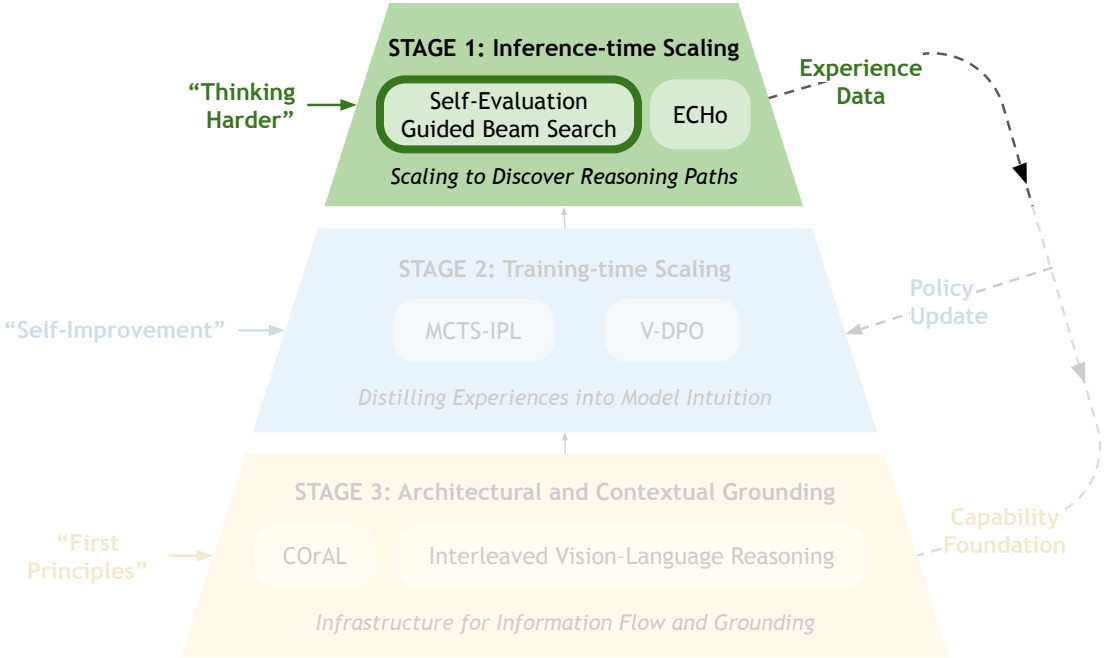


Content

- ❏ **PART 1**: Introduction
- ❏ **PART 2**: Layer One – Inference-time Scaling
- ❏ **PART 3**: Layer Two – Training-time Scaling
- ❏ **PART 4**: Layer Three – Architectural & Contextual Grounding
- ❏ **PART 5**: Reflections & Diagnosis
- ❏ **PART 6**: Future Directions & Closing

From Intuition to Deliberation

Inference-time Reasoning Scaling



Challenge in Test-time Scaling for Reasoning

LLMs struggle with error accumulation across multiple steps

- Breaking down a problem solution into **intermediate steps** facilitates reasoning.

$$P(a \mid x) = \mathbb{E}_{R \sim P(R \mid x)} P(a \mid R, x)$$

decompose answer generation into multiple reasoning steps

$$P(R = s^{1:T} \mid x) = \prod_t P(s^t \mid x, s^{1:t-1})$$

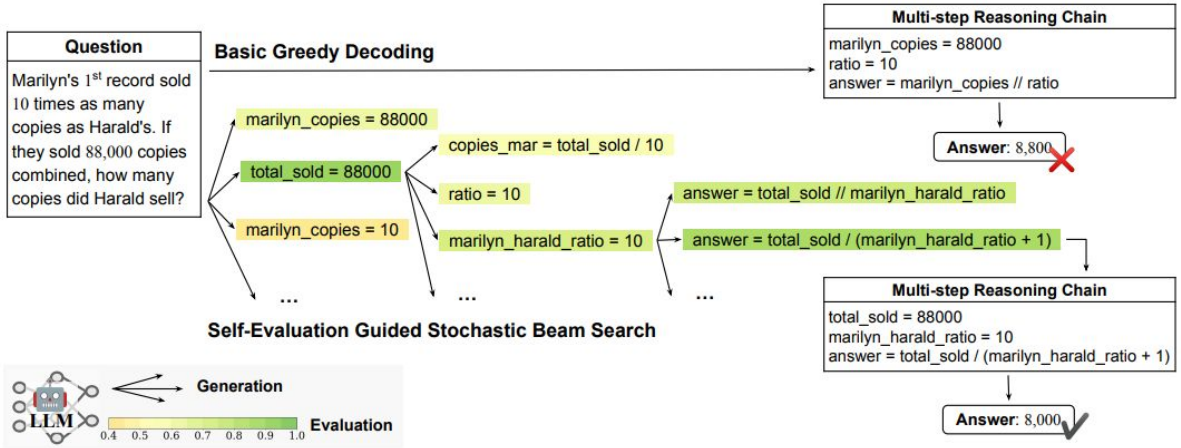
factorize reasoning in an autoregressive manner

- The growing number of steps leads to an **exponential growth in the search space** for generation.

Towards "Self-Balancing" of the Reasoning Process

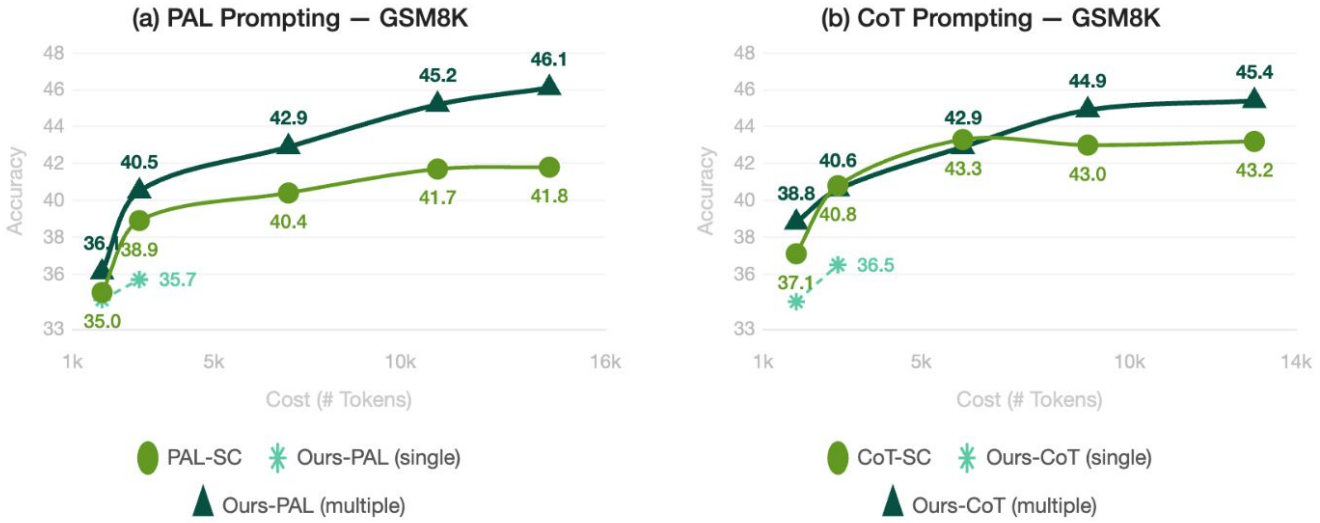
Core Hypothesis: The Calibration Gap – Evaluation *vs.* Generation

- Models can often *recognize* a correct step even when they cannot *generate* it spontaneously.
- We integrate **stepwise self-evaluation** to guide the reasoning process.



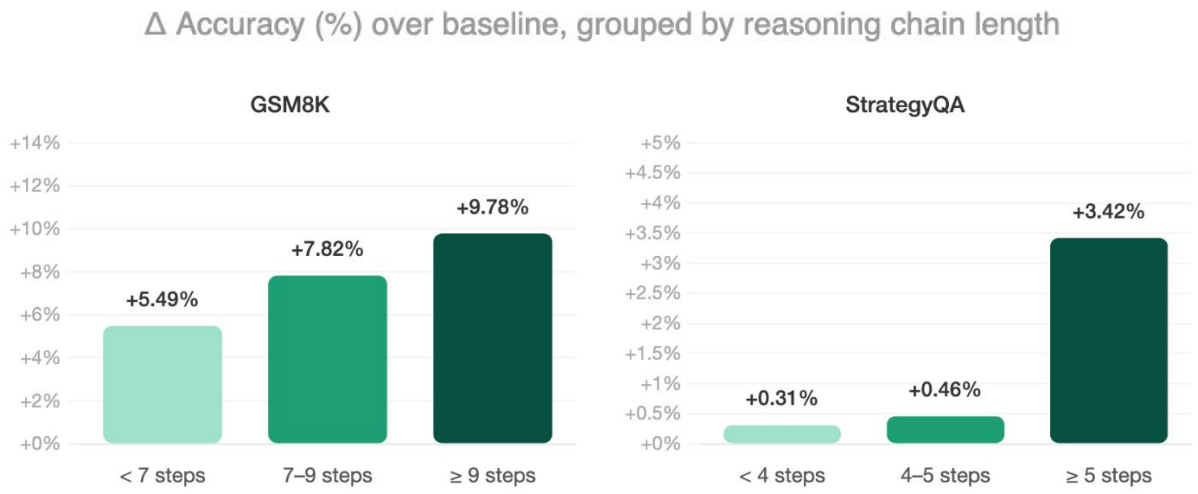
Performance & Cost: Exploitation v.s. Exploration

- *Self-Evaluation Guided Beam Search* can **exploit** to find a better reasoning chain for higher accuracy.



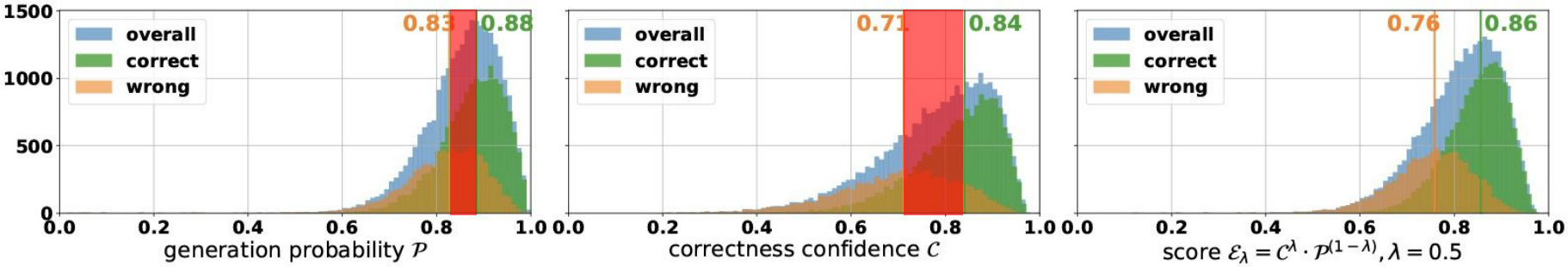
Performance & Cost: Exploitation v.s. Exploration

- ❏ *Self-Evaluation Guided Beam Search* can **exploit** to find a better reasoning chain for higher accuracy.
- ❏ Performance gains mainly come from longer reasoning chains.

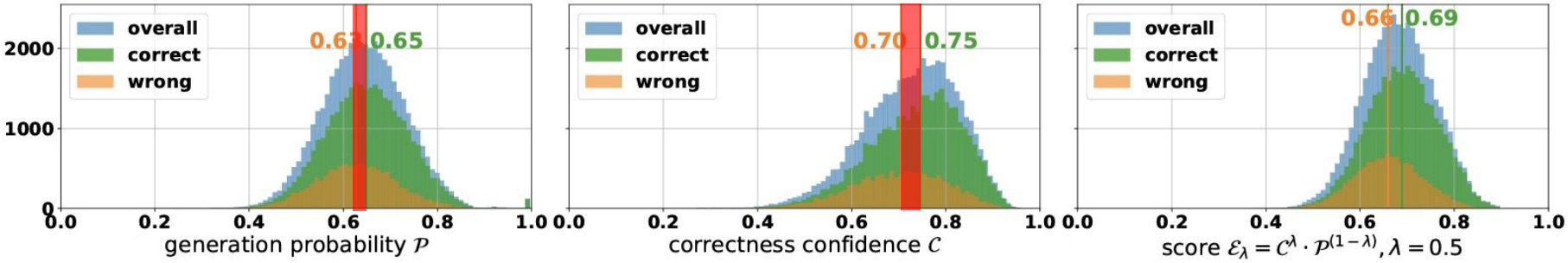


Key Insight: When Does Self-Evaluation Work?

- **Calibration gap**: GSM8K vs. StrategyQA
  - task-dependent ⇔ structure of the underlying reasoning space



(a) Score distribution of PAL baseline predictions on GSM8K.



(b) Score distribution of CoT baseline predictions on StrategyQA.

PS: The correct/wrong label is determined by the correctness of final answer. The samples analyzed are all complete responses.

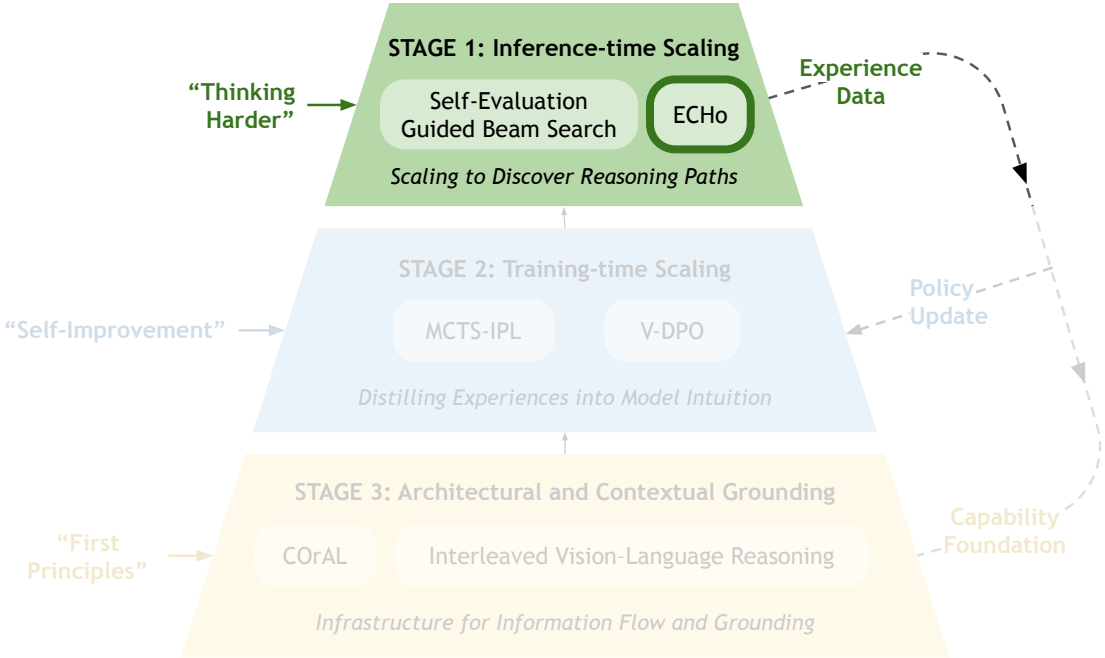
Key Insight: When Does Self-Evaluation Work?

- **Calibration gap**: GSM8K vs. StrategyQA
  - task-dependent ⇔ structure of the underlying reasoning space

- The **coherence–correctness gap**: Self-evaluation tends to optimize for *coherent* reasoning rather than *correct* reasoning.
  - Hard to disentangle *correctness* and *fluency* in free-text & open-ended tasks.

From Intuition to Deliberation

Inference-time Reasoning Scaling

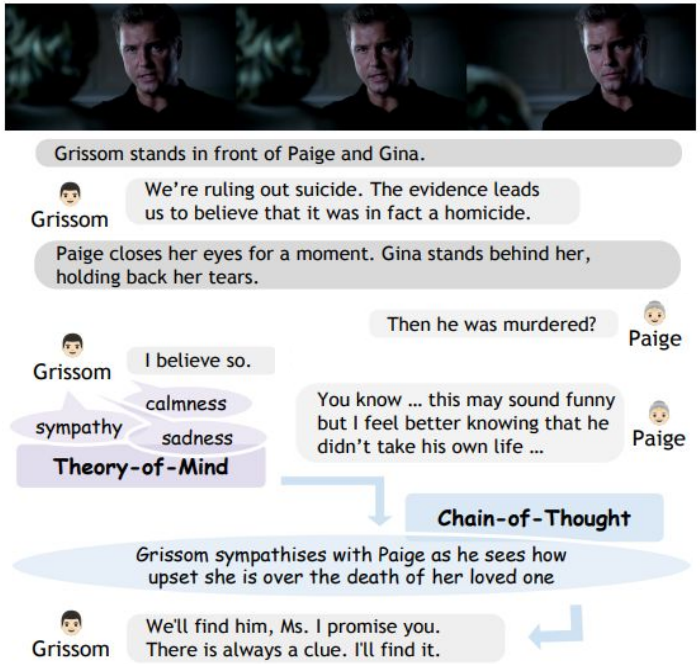


Extending to Social Cognition

Theory of Mind (ToM) as a predetermined principal to guide Chain-of-Thought reasoning

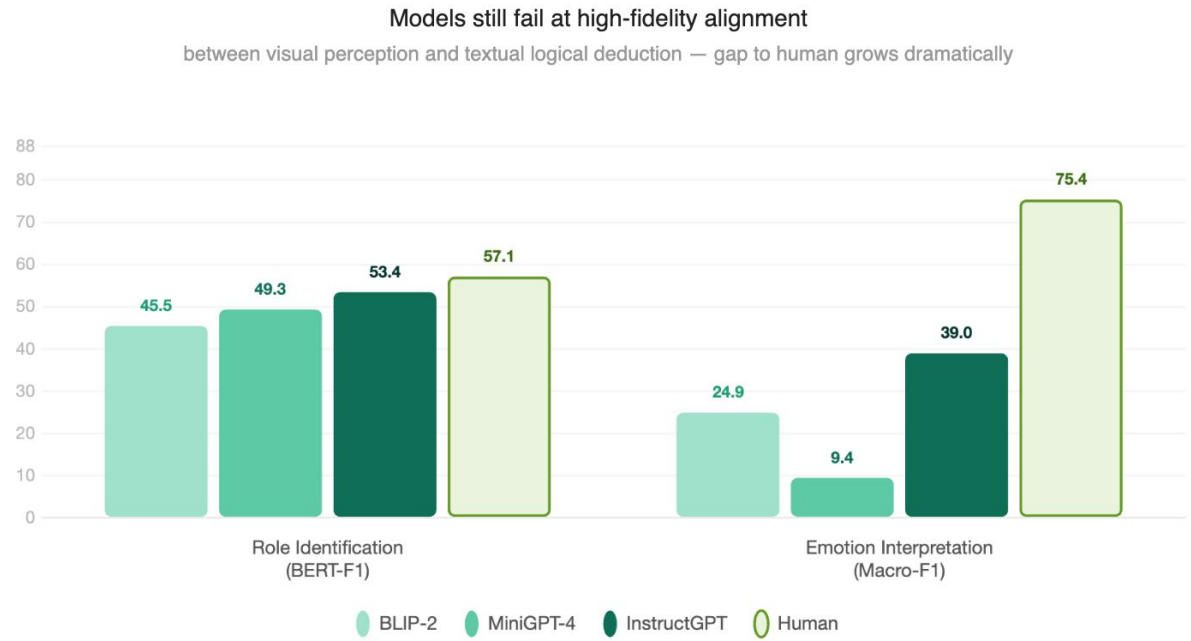
- Theory of Mind (ToM):

  The ability to attribute mental states to others and understand that their beliefs may differ from reality.



Key Insight: When Does Principle-guided Framework Work?

- Models still fail at high-fidelity alignment between visual perception and logical deduction.

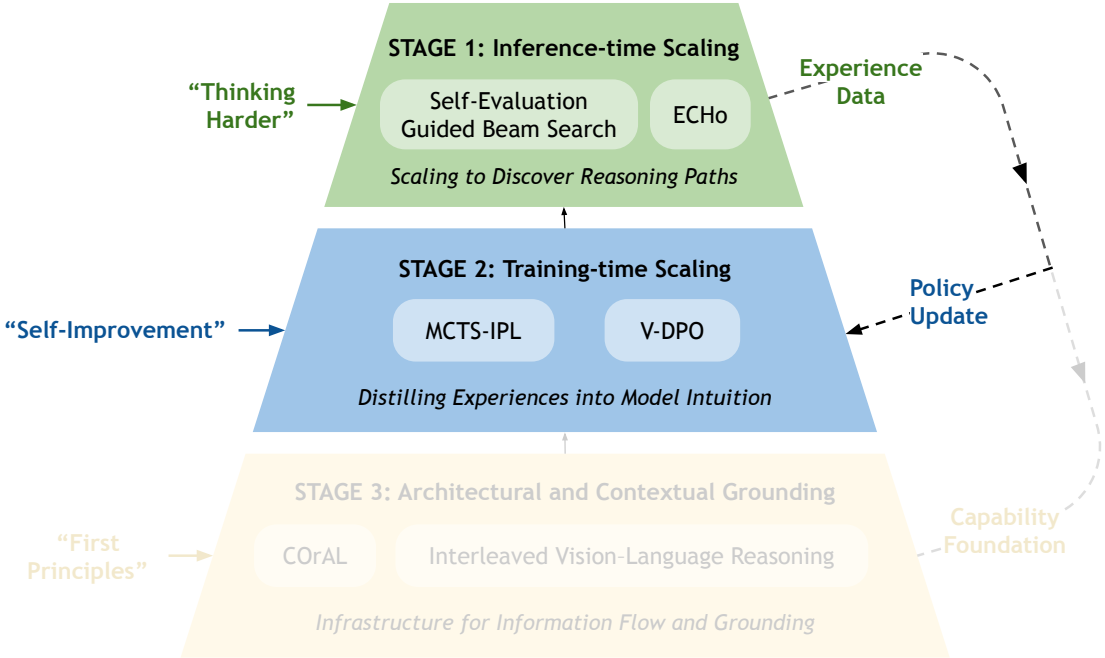


Key Insight: When Does Principle-guided Framework Work?

- ❑ Models still fail at high-fidelity alignment between visual perception and logical deduction.
- ❑ Principle-guided formulations have task-specific gains but limited generalizability.
  - limited dimensions << humans' real emotions

Transition: Exploration is Powerful but Expensive

Can we internalize this into model parameters?

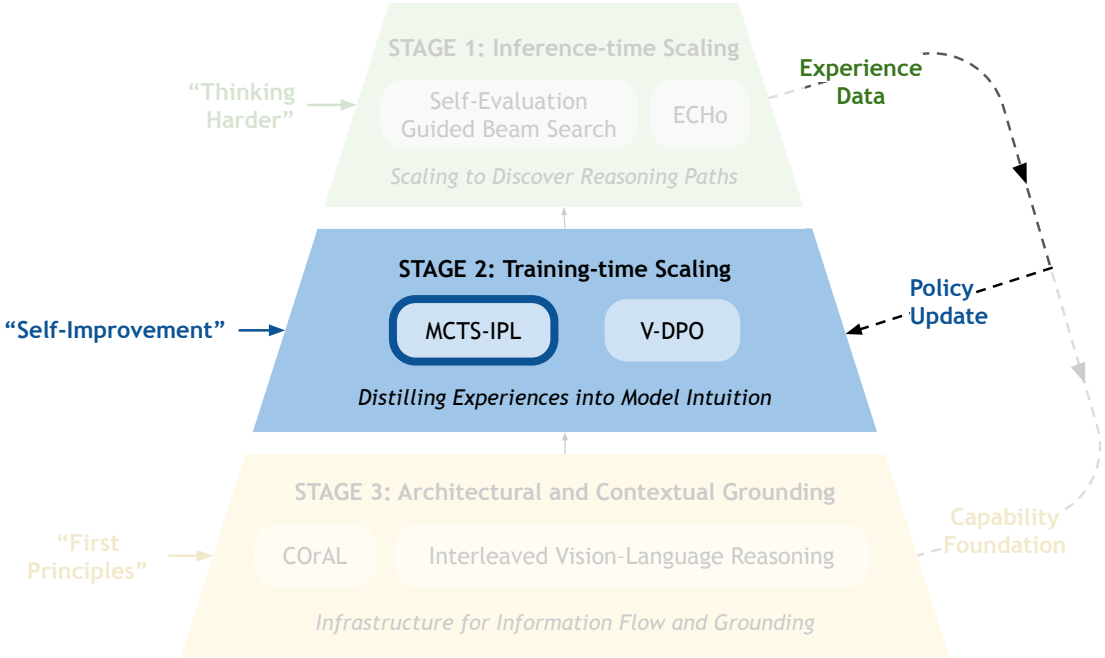


Content

- ❏ **PART 1**: Introduction
- ❏ **PART 2**: Layer One – Inference-time Scaling
- ❏ **PART 3**: Layer Two – Training-time Scaling
- ❏ **PART 4**: Layer Three – Architectural & Contextual Grounding
- ❏ **PART 5**: Reflections & Diagnosis
- ❏ **PART 6**: Future Directions & Closing

From Deliberation to Intuition

Training-time Scaling via Alignment



Challenge in Credit Assignment (Evaluation)

external rewards → sparse; internal signals → noisy
How do we collect supervision signals to enable self-evolution?

- ❑ We lack high-quality *process-level* **data** and **evaluation signals**.
  - *binary outcome labels* (correct or not): **sparse** and **coarse-grained**.
  - LLM-as-the-judge: **biased** and **inaccurate** *process-level* self-evaluation that is not fully reliable.

The “Search–Alignment” Paradigm

Monte Carlo Tree Search (MCTS) as Data Engine + Iterative Alignment

- *stochastic* elements in MCTS

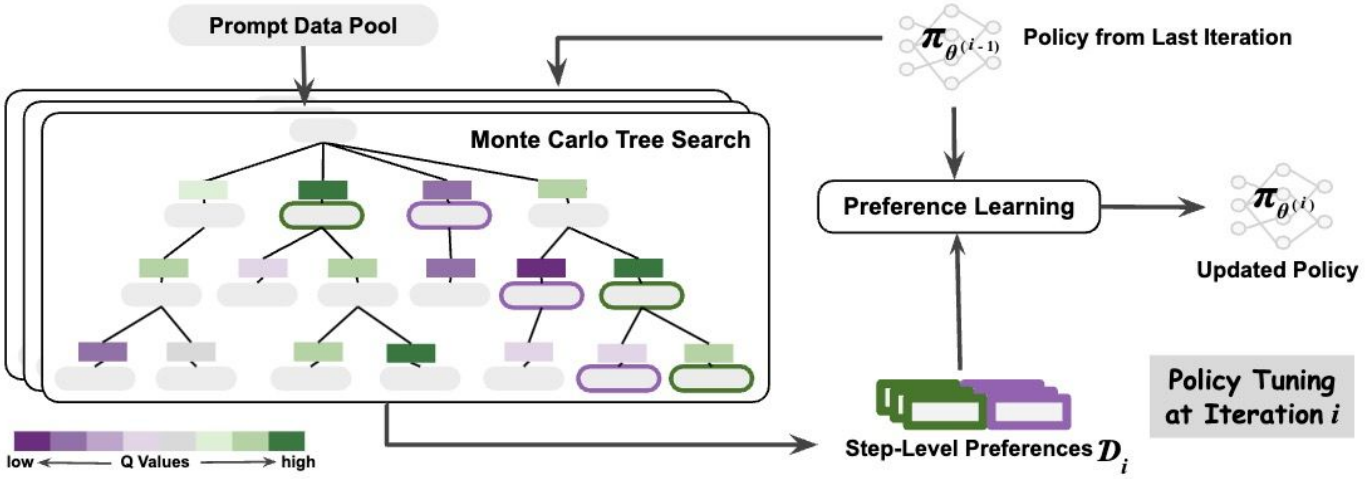
  → estimate the value of intermediate states based on stochastic sampling

- test-time compute from MCTS

  → distilled back into the base LLM (training-time policy improvement)

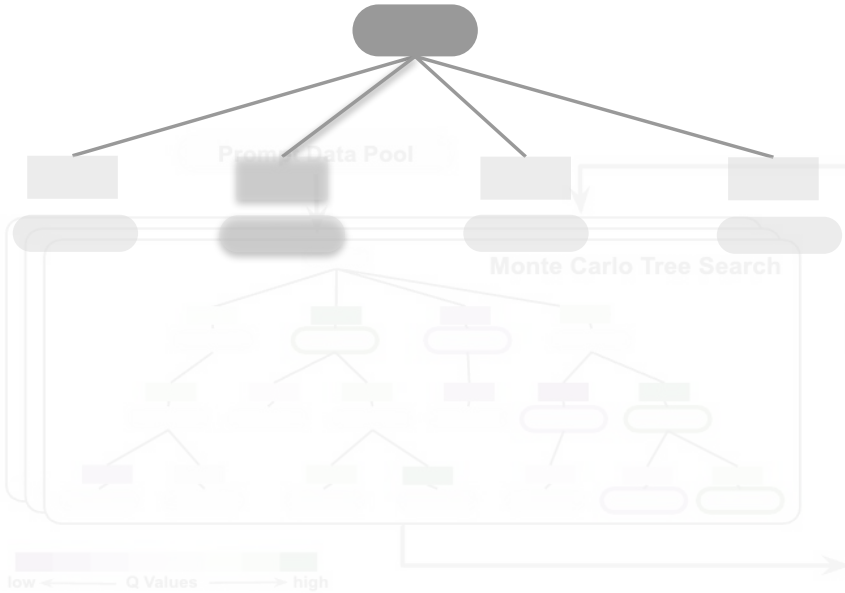
The “Search–Alignment” Paradigm

Better Policy → More efficient Search → Higher quality Data → Stronger Policy



Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning (NeurIPS 2024 System-2 Reasoning Workshop)
Xie et al.

“Search” – MCTS as Data Engine

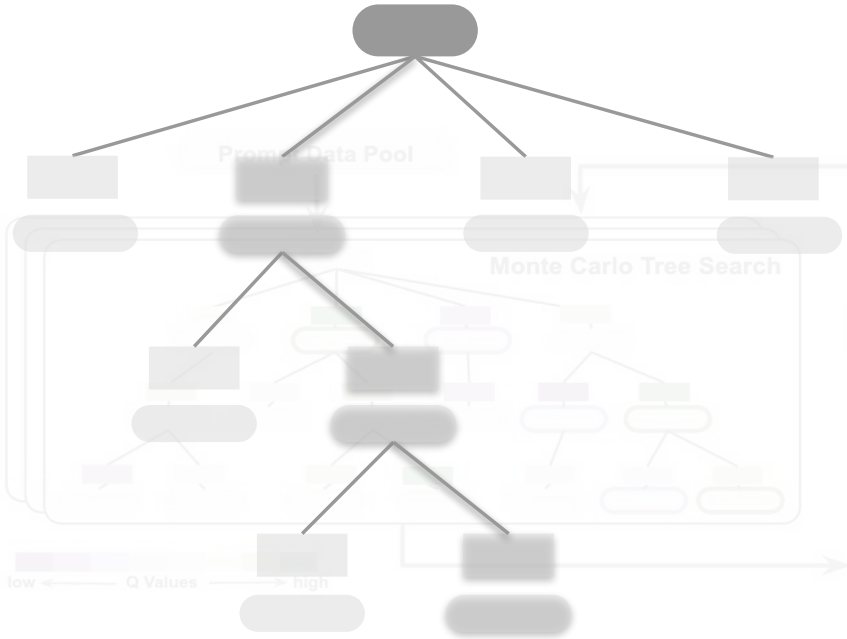


– **Select**. We guide the selection process by the action-value function Q and the number of visited times N.

$$s_{t+1}{}^{*} = \arg\max_{s_t}[Q(s_t, a) + c_{\text{puct}} \cdot p(a \mid s_t)\frac{\sqrt{N(s_t)}}{1 + N(s_{t+1})}]$$

the current policy’s probability distribution for generating a step a at state s_t

“Search” – MCTS as Data Engine



– **Select**. We guide the selection process by the action-value function Q and the number of visited times N.

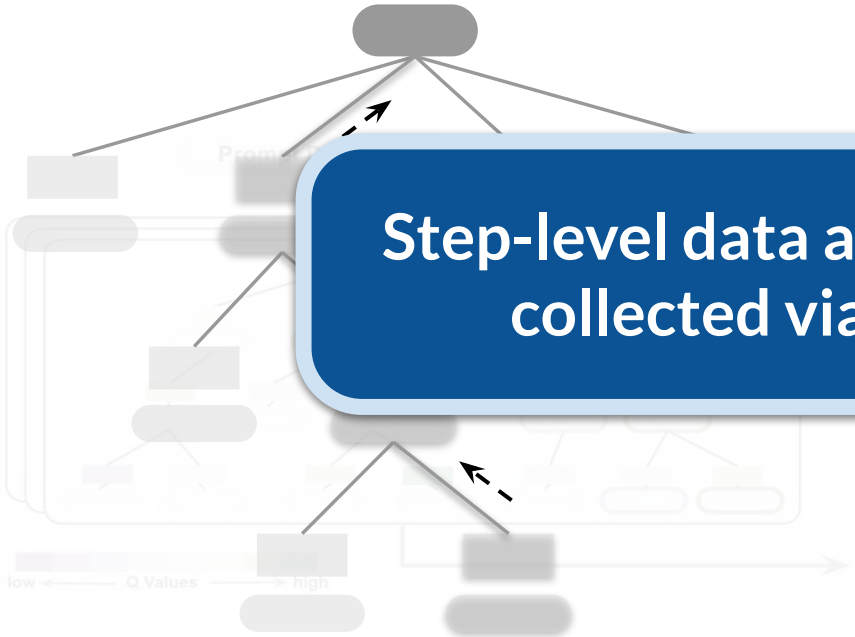
– **Expand**. We expand by generating children nodes with the current policy. We assess each new node based on the advantage of taking the action a at state s_t.

Advantage: reward difference between states s_t and s_{t+1}
Reward: outcome correctness & self-evaluation

1 for correct and -1 for incorrect terminal states

$$\mathcal{C}(s_t) = \pi_\theta(\mathrm{A} \mid \mathrm{prompt}_{\mathrm{eval}}, x, s_t)$$

“Search” – MCTS as Data Engine



– **Select**. We guide the selection process by the action-value function Q and the number of visited times N.

and by
We assess
ng the action a

Step-level data and evaluation signals collected via stepwise MCTS

ed out from the terminal state back to the root.

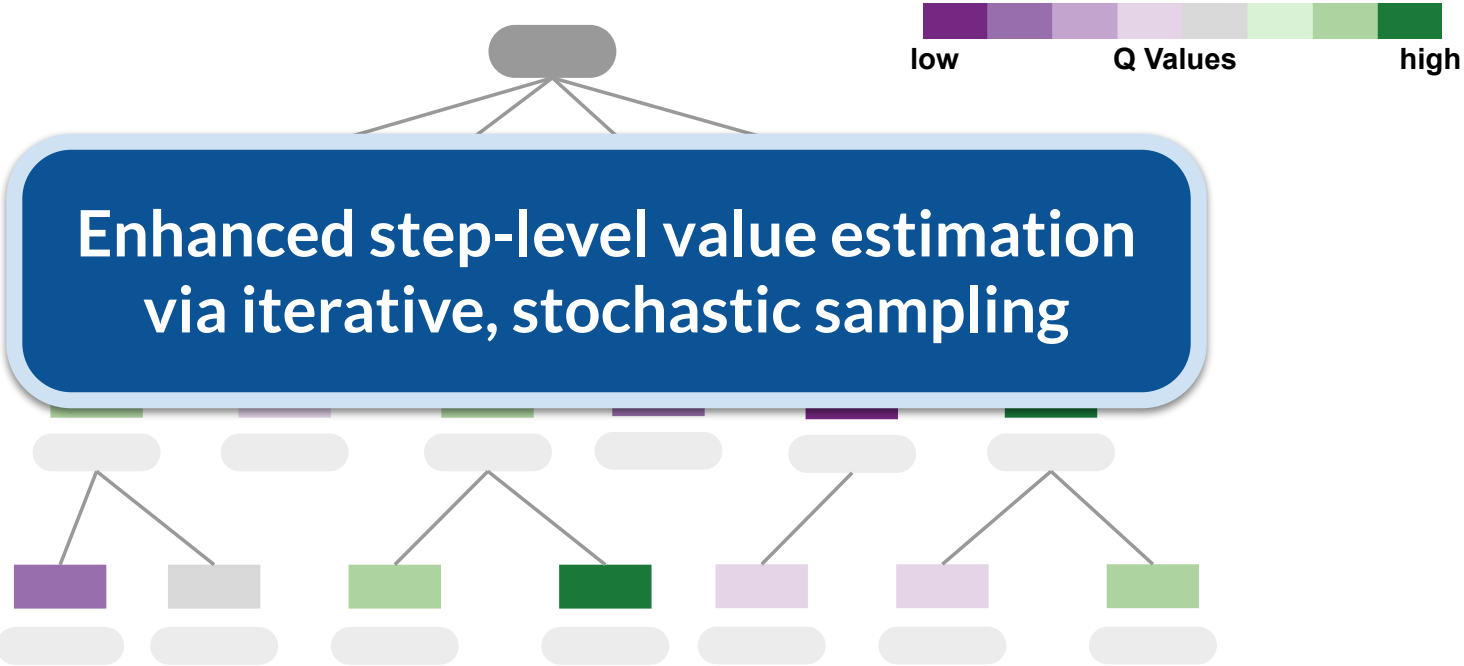
$$Q(s_t, a) \leftarrow r(s_t, a) + \gamma V(s_{t+1})$$

$$V(s_t) \leftarrow \sum_a N(s_{t+1}) Q(s_t, a) / \sum_a N(s_{t+1})$$

$$N(s_t) \leftarrow N(s_t) + 1$$

“Search” – MCTS as Data Engine

After multiple iterations...



"Alignment" – Iterative Step-level Preference Learning

– **Step-level Preference Data**. We leverage the visit-count N and Q values from MCTS to label trajectory preferences (higher → better).

A search tree of depth T yields T pairs of step-level preferences.

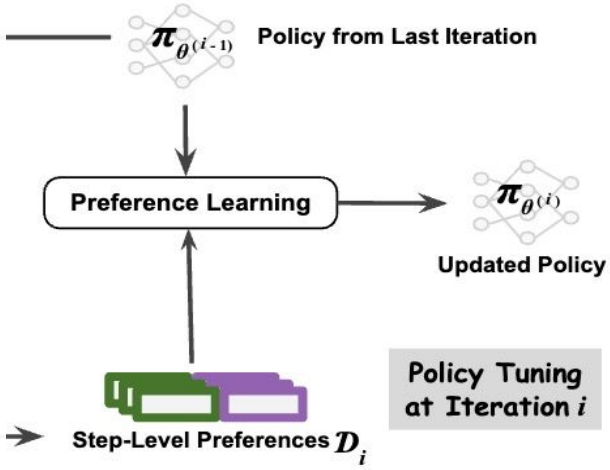
"Alignment" – Iterative Step-level Preference Learning

– **Step-level Preference Data**. We leverage the visit-count N and Q values from MCTS to label trajectory preferences (higher → better).

– **Iterative Preference Learning**. We employ conservative DPO with a data-dependent label smoothing variable, α.

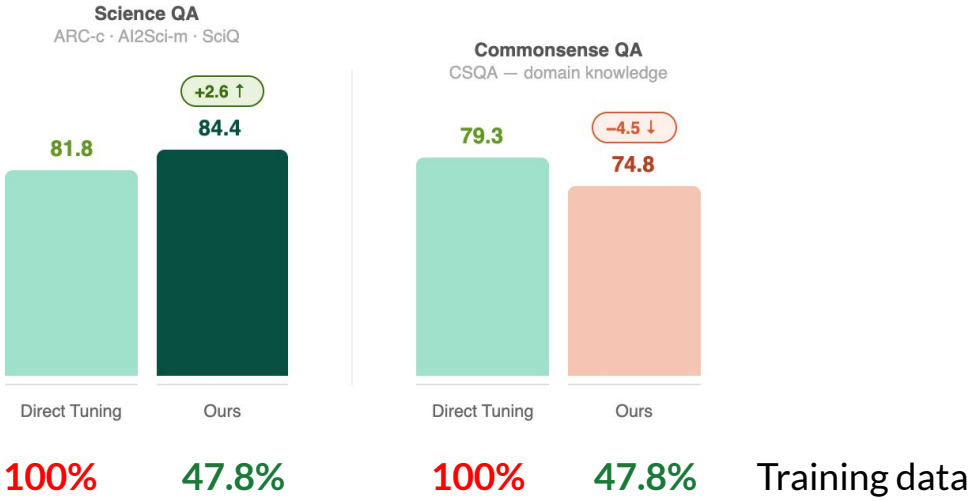
When MCTS is confident about which trajectory is better
→ learn aggressively ($\alpha \to 0$, standard DPO)

When MCTS is uncertain
→ smooth the label, learn conservatively (α grows)



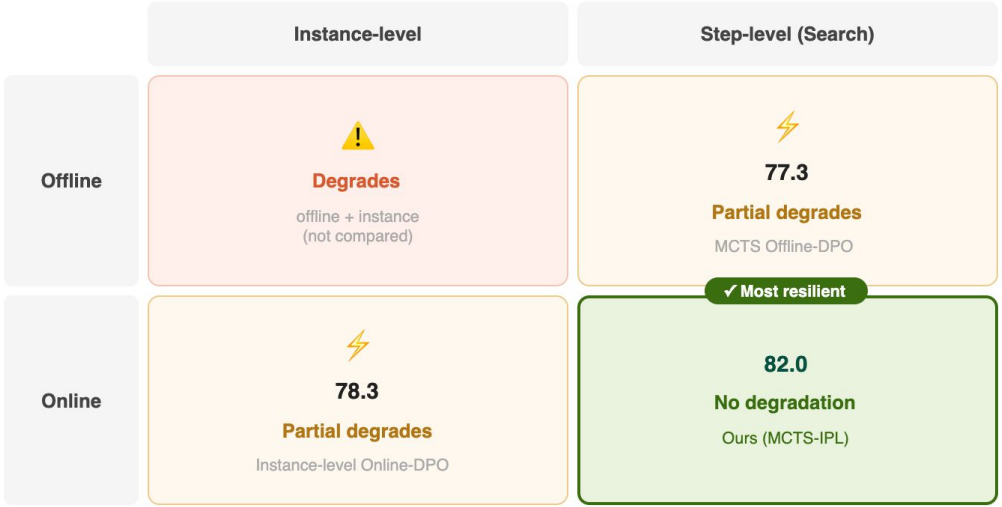
Experiment: Sample Efficiency

- The Self-Improvement Loop is more **sample-efficient** than Direct Tuning.
  - Structural reasoning tasks benefit especially strongly.



Experiment: Online Step-level Learning improves Resilience

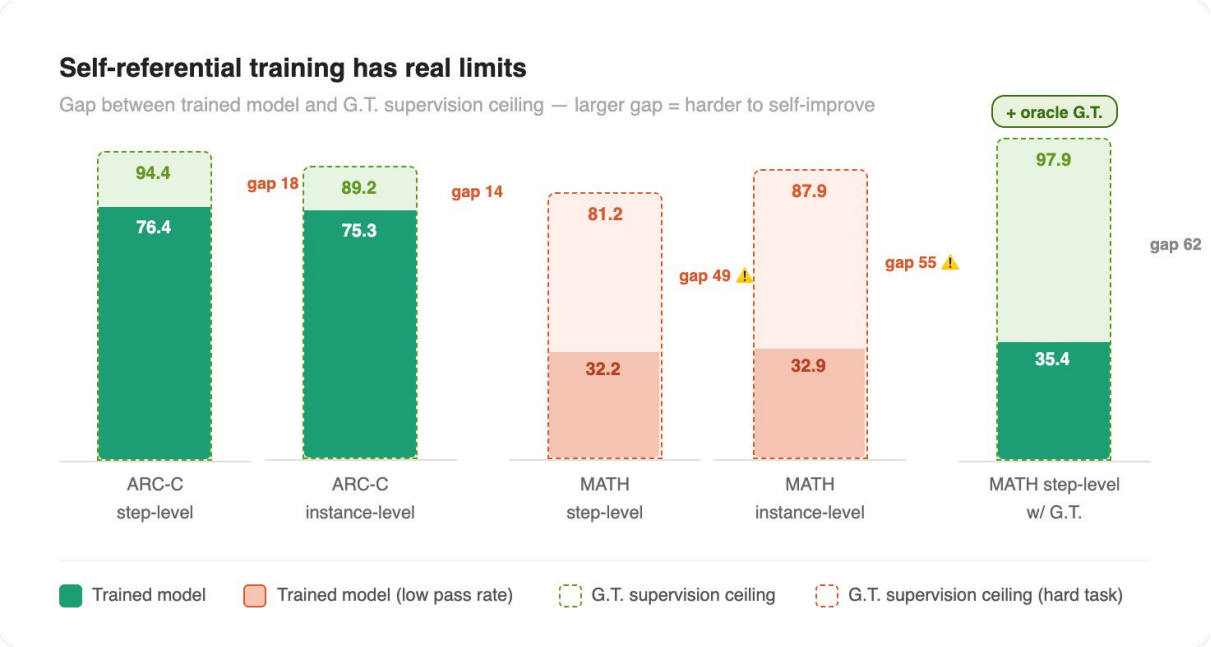
- Online (Iterative) v.s. Offline Learning: *Online* is more **resilient to degradation**.
- Step-level Search v.s. Instance-level Sampling: *Search* is more **resilient to degradation** in iterative DPO.



Average accuracy across
ARC-C, AISci-m, CSQA, SciQ

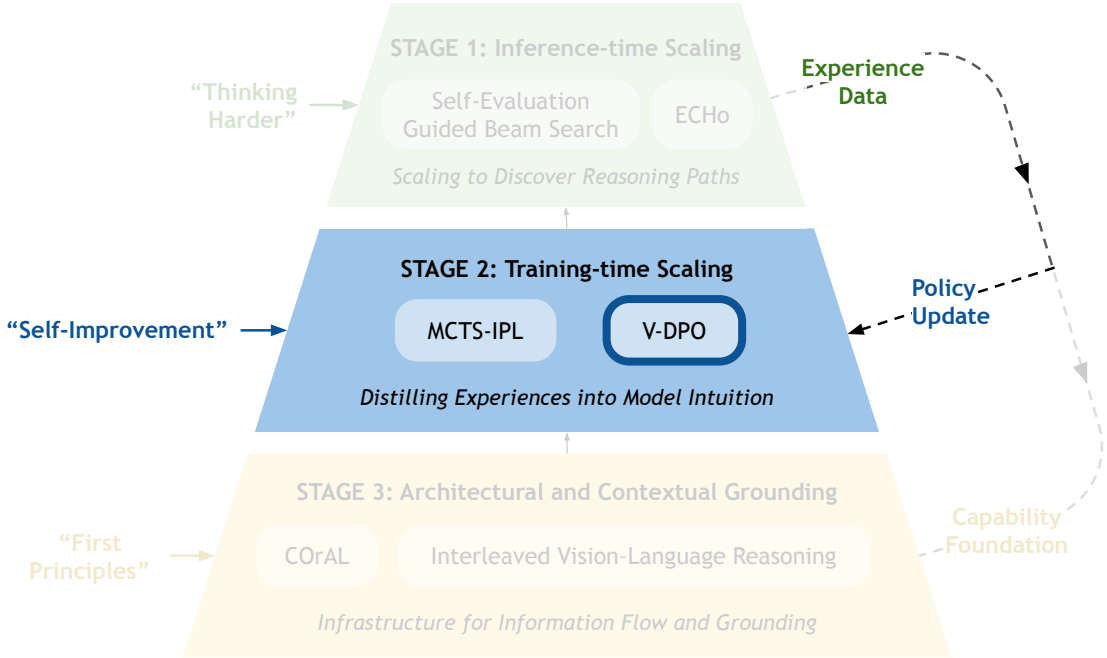
Limitation of Self-referential Training

- With the same compute budget, step-level search (MCTS) can suffer from low pass rate due to constraint on the maximal search breadth.



From Deliberation to Intuition

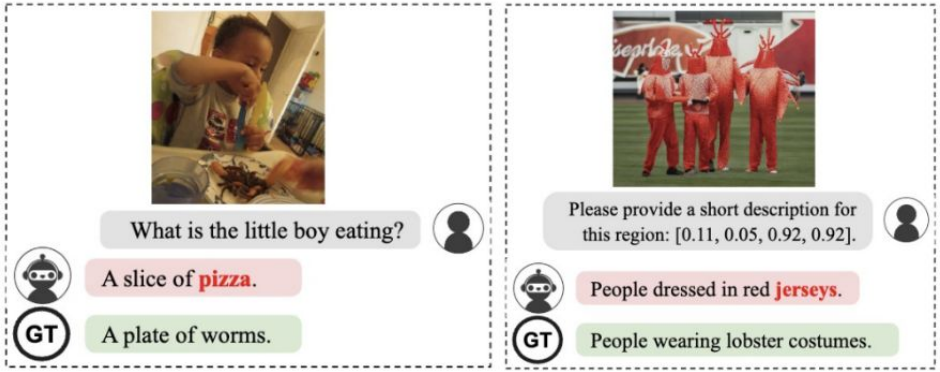
Training-time Scaling via Alignment



Specific Data Pipeline for Hallucination Mitigation

Hallucination in Large Vision Language Models

- **Non-existing or erroneous descriptions of visual contents**, which is especially challenging to understanding unconventional images.



One Important Cause of Hallucination – Language Bias

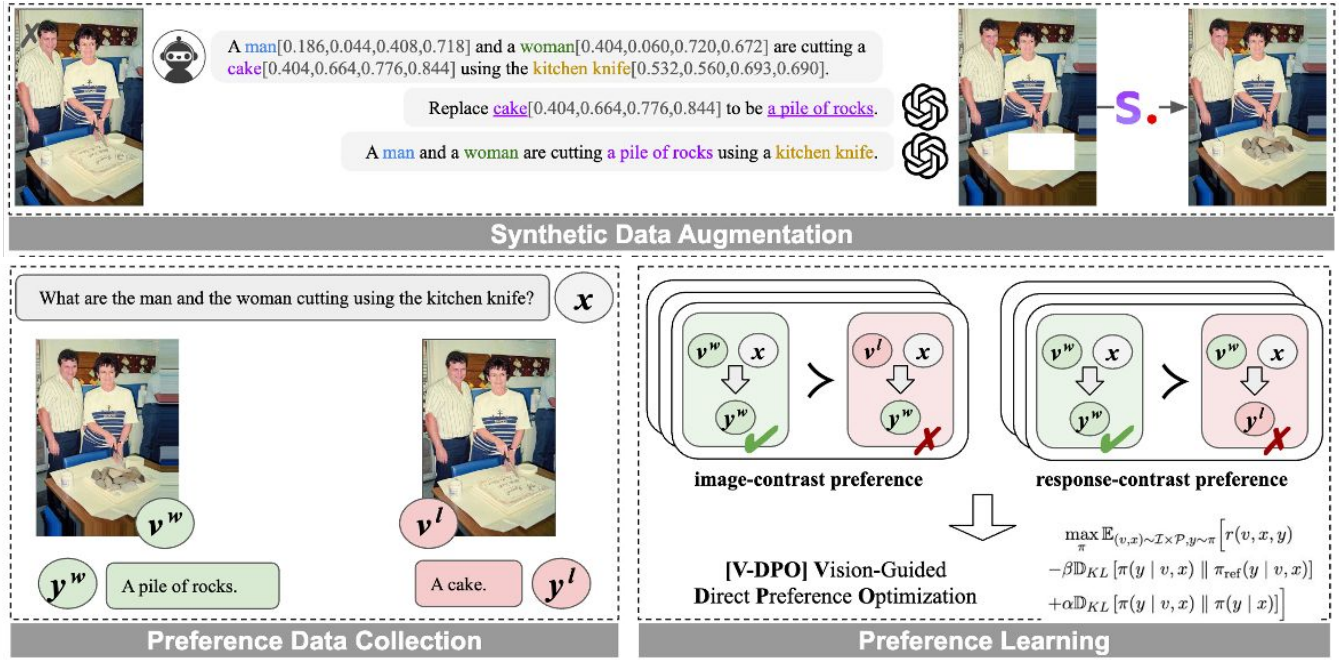
LVLM hallucination can be attributed to the integration of pretrained LLM backbones.

- **Insufficient context attention**: the model prioritizes *language patterns* and focuses on *partial tokens* rather than fully grounding generation in both visual and textual context.

$$\log p_{\text{accurate}} - \log p_{\text{hallucinatory}}$$

Multi-Model Collaborative Pipeline for Preference Data Collection

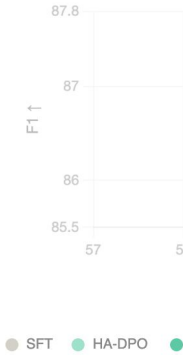
V-DPO with Image- and Response-Contrast Preference Data



Experiment: Hallucination Reduction

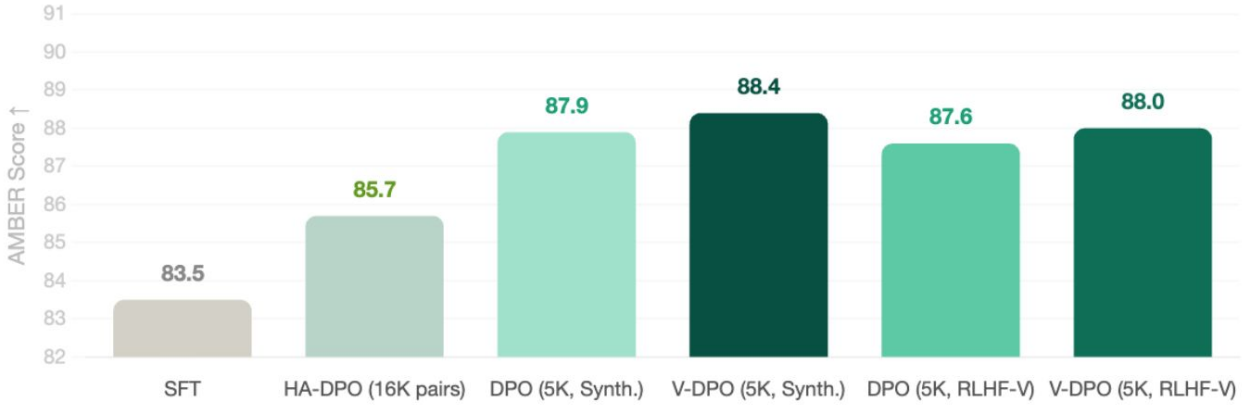
POPE — Better F1

V-DPO lands in the ideal t



AMBER — 5K pairs beats 16K pairs

Visual guidance is more sample-efficient for hallucination mitigation



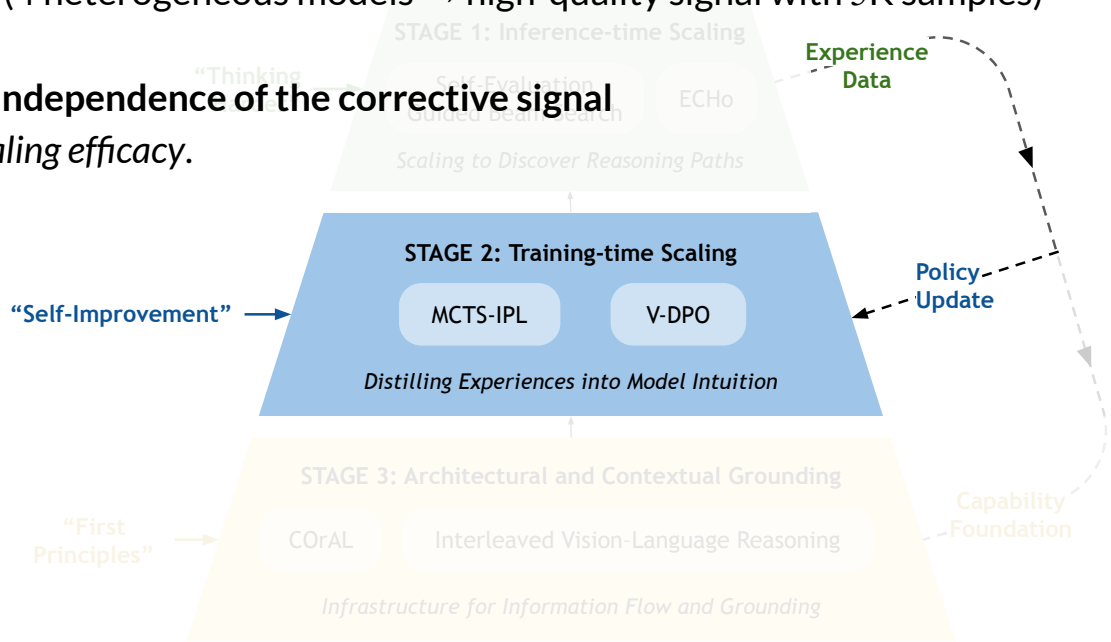
V-DPO (5K pairs) outperforms **HA-DPO (16K pairs)** by **+2.7** — 3× fewer data, better result.

V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided DPO (EMNLP 2024 Findings)

Xie et al.

Summary: The Power of External Verification

- **Contrast**: **MCTS-IPL** (same model generates & evaluates + G.T. outcome-only labels → instability) versus **V-DPO** (4 heterogeneous models → high-quality signal with 5K samples)
- **Diversity** and **independence of the corrective signal** determines *scaling efficacy*.

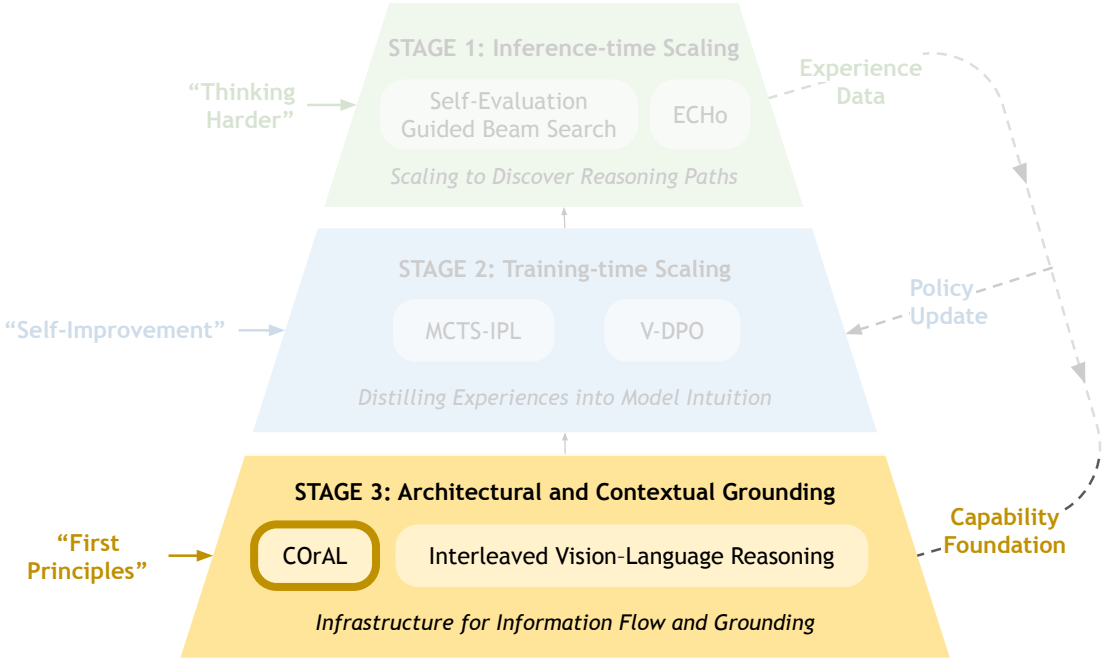


Content

- ❏ **PART 1**: Introduction
- ❏ **PART 2**: Layer One – Inference-time Scaling
- ❏ **PART 3**: Layer Two – Training-time Scaling
- ❏ **PART 4**: Layer Three – Architectural & Contextual Grounding
- ❏ **PART 5**: Reflections & Diagnosis
- ❏ **PART 6**: Future Directions & Closing

Returning to First Principles

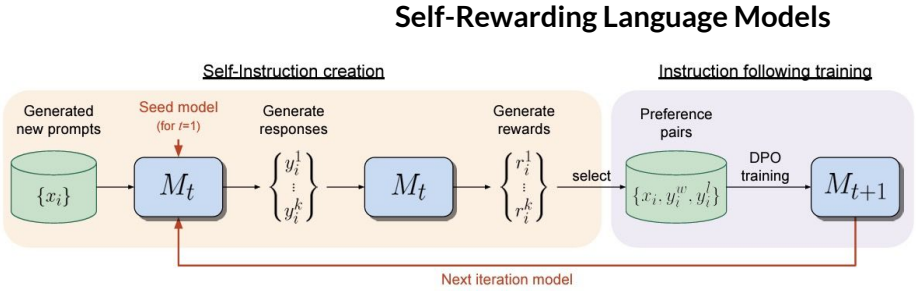
Architectural & Contextual Grounding

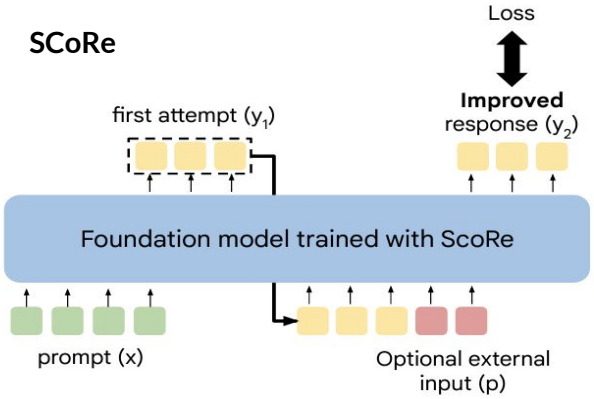


Existing Iterative Refinement Frameworks

Background

- *Iterative Refinement* and *Self-Correction* have emerged as promising paradigms to improve efficacy.
- Existing approaches typically implement at the **application or prompting level**, as a multi-turn process relying on *next-token* prediction based on *autoregressive* (AR) modeling.





Architectural Grounding

Pros and Cons of Next-Token based AR Language Modeling

VL: varying-length generation
BT: backtrack / look-ahead
MV: multi-variable generation
MD:
FS: fi
EF: ir
IT: m

Can we unify the strengths of *denoising* techniques with *order-agnostic modeling* to enhance the capabilities of AR-LLMs?

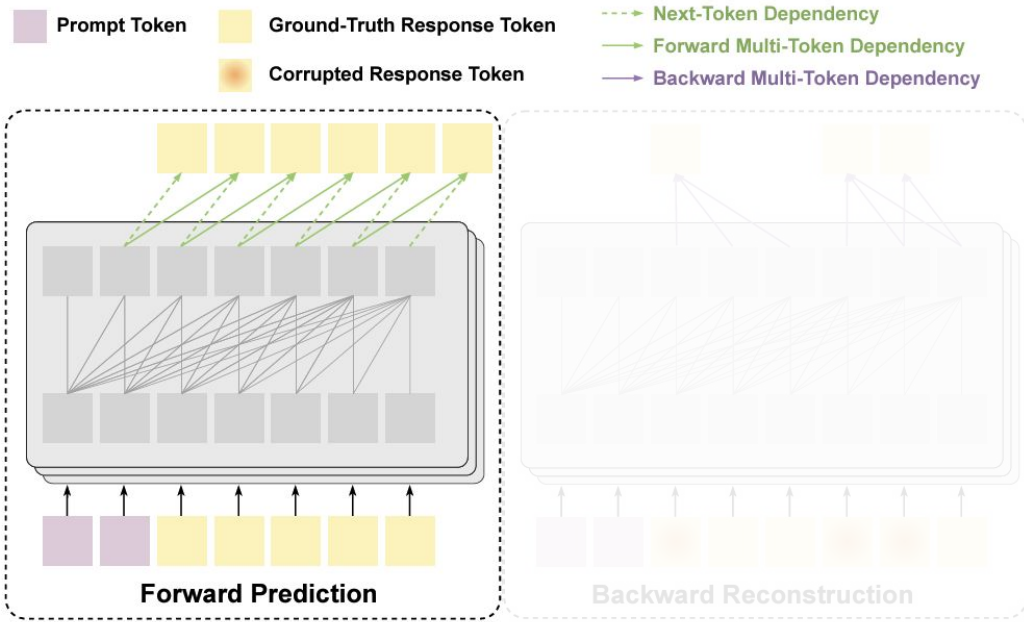
Next-							
Permutation-Based AR (Uria et al., 2014)	✗	✓	✓	✓	✗	✓	✗
NAR (Gu et al., 2018)	✗	✓	✓	✓	✓	✓	✓
Diffusion (Ho et al., 2020)	✗	✓	✓	✓	✓	✗	✓
Consistency Model (Song et al., 2023)	✗	✓	✓	✓	✓	✓	✓
COrAL (Ours)	✓	✓	✓	✓	✓	✓	✓

Sliding Blockwise Order-Agnostic Decoding

Context-wise Order-agnostic Language Modeling (COrAL) performs multi-token prediction and refinement in the sliding block.

Context-Wise Order-Agnostic Language Modeling

We visualize the order-agnostic dependencies within a context window $k = 2$.



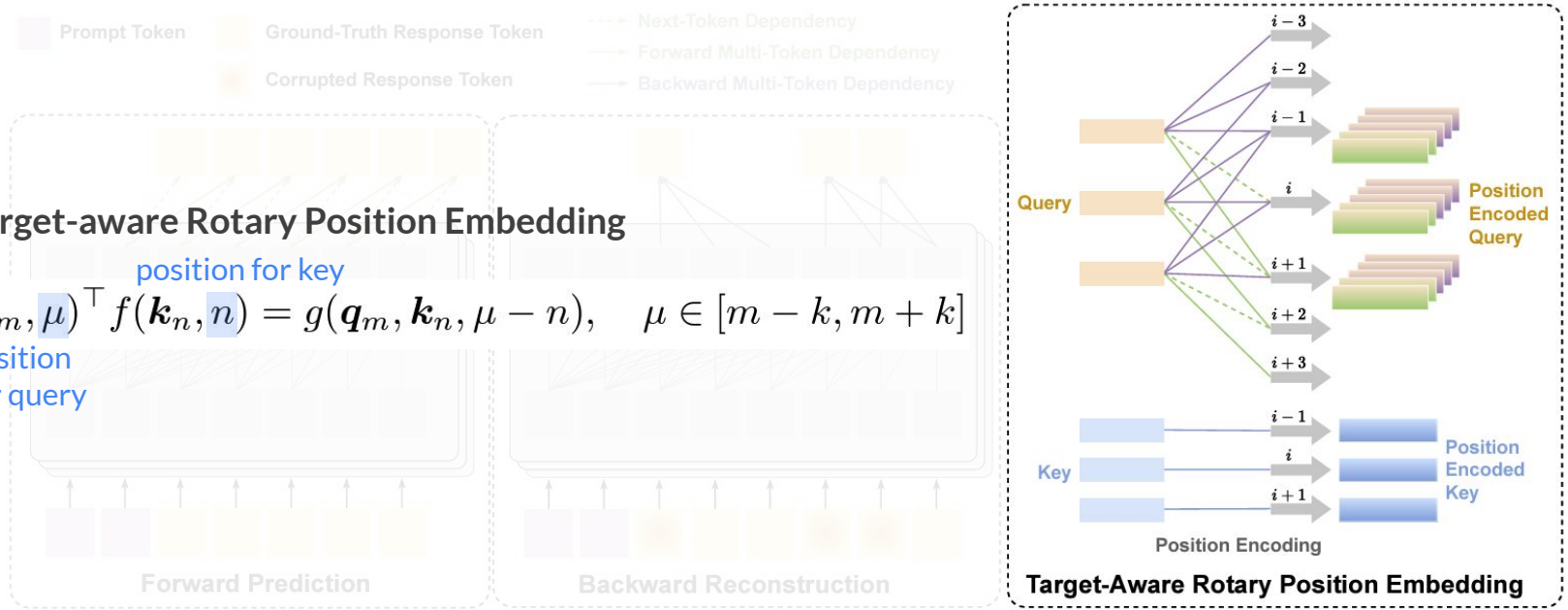
Training Objective

$$\log p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \geq \sum_{t=1}^{T} \mathbb{E}_{i \in [t-k, t+k]} \mathbb{E}_{l \geq 0} \log p_\theta(y_t \mid \boldsymbol{y}_{\leq i}^{(l)}, \boldsymbol{x})$$

- **Forward Prediction**
  Predict multiple future tokens simultaneously, given ground-truth past tokens.
- **Backward Reconstruction**
  Reconstruct randomly-corrupted tokens as intermediate states.

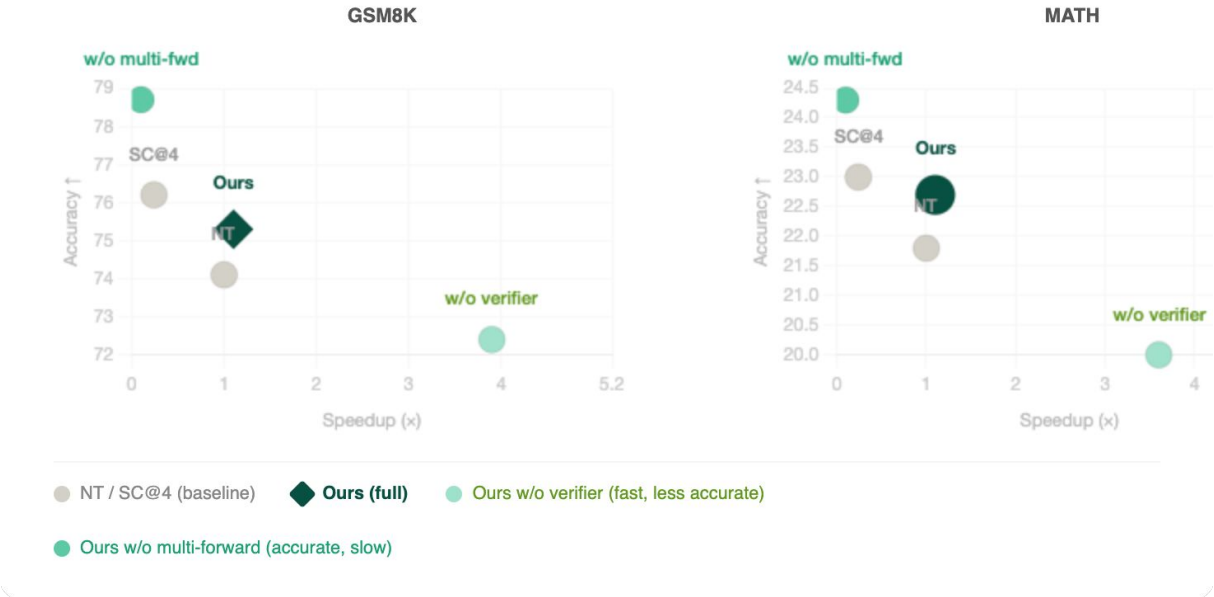
Context-Wise Order-Agnostic Language Modeling

We visualize the order-agnostic dependencies within a context window $\mathrm{k} = 2$.



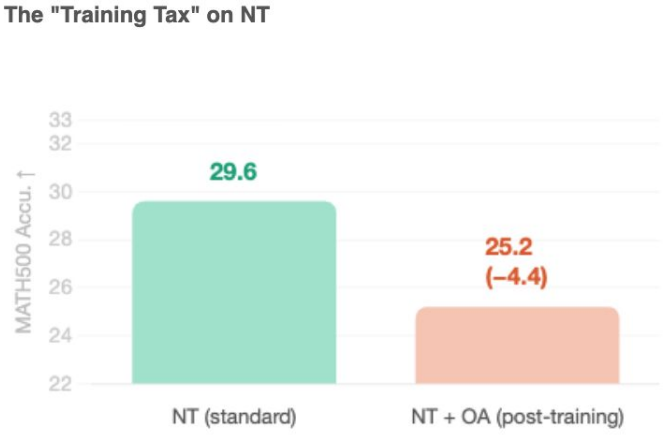
Performance–Speed Trade-off

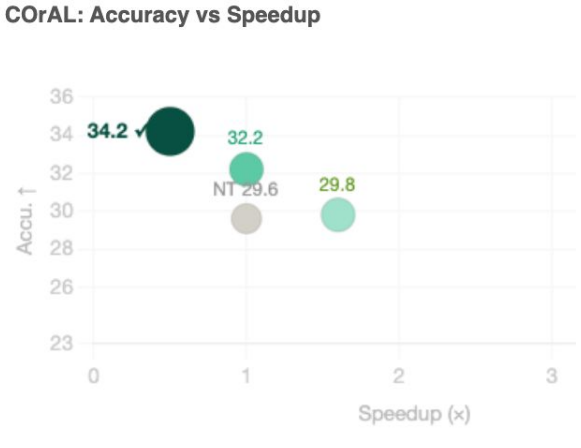
- ❏ Multi-token *forward* prediction → **Speed up** the inference process
- ❏ *Backward* refinement → Iteratively revise the generated content for **better accuracy**.



Key Observation: Training Tax from Post-training Grounding

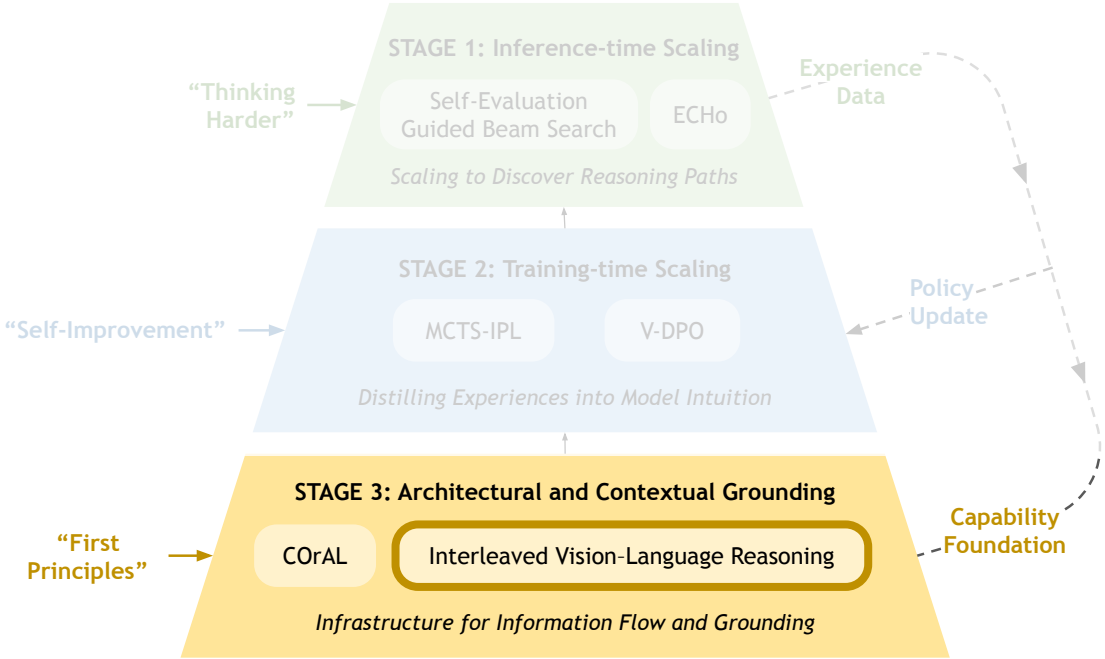
- Order-agnostic post-training incurs a systematic training tax due to misalignment with pretraining objectives.
- The AR pretraining paradigm may be **fundamentally misaligned** with the requirements of grounding-intensive tasks that demand *parallel*, *iterative*, or *order-agnostic* reasoning.





Returning to First Principles

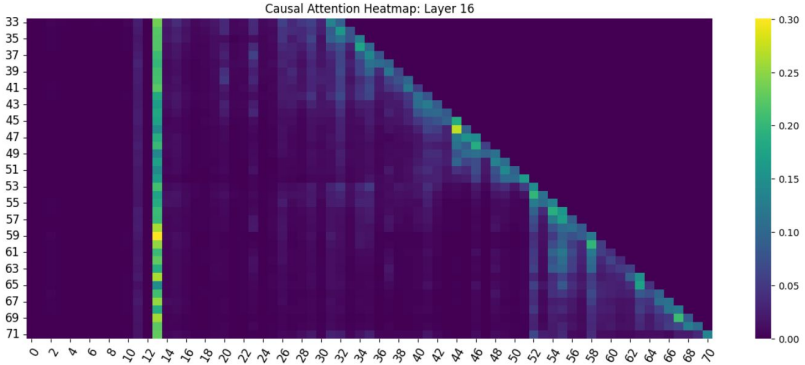
Architectural & Contextual Grounding



Static Visual Embeddings Lack Dynamic Grounding

LVLMs struggle with Visual Understanding due to architectural inductive bias

- **Language Bias**: The model **prioritizes language patterns** and **focuses on partial tokens** rather than fully grounding the generated content in both visual and textual context.
  - <|im_start|>system\nYou are a helpful assistant.<|im_end|>\n<|im_start|>user\n<image>\nWhat are the main objects on the table in the image?<|im_end|>\n<|im_start|>assistant\nThe image shows a table with various items, including a plate with remnants of food and a glass. The lighting suggests that the photo was taken in an indoor setting during the evening or night.<|im_end|>



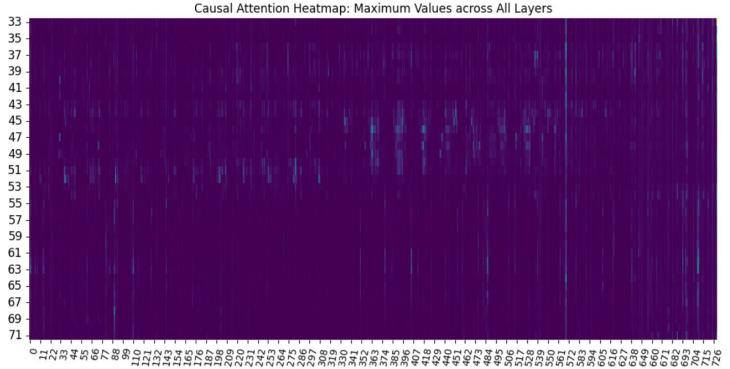
Summarization tokens attract most attention.

Textual tokens attract significantly more attention than visual tokens.

Static Visual Embeddings Lack Dynamic Grounding

LVLMs struggle with Visual Understanding due to architectural inductive bias

- **Language Bias**: The model **prioritizes language patterns** and **focuses on partial tokens** rather than fully grounding the generated content in both visual and textual context.
- **Sparse (and Smooth) Attention (on Visual Tokens)**: As the *Key* set in *Attention* grows, standard softmax tends to **spread mass more evenly**, weakening focus on truly relevant tokens.



<|im_start|>system\nYou are a helpful assistant.<|im_end|>\n<|im_start|>user\n<image>\nWhat are the main objects on the table in the image?<|im_end|>\n<|im_start|>assistant\nThe image shows a table with various items, including a plate with remnants of food and a glass. The lighting suggests that the photo was taken in an indoor setting during the evening or night.<|im_end|>

Visual Attention varies across tokens, seeking visual information when generating visual-specific tokens such as *objects* and *attributes*.

We observe the problem of *attention dilution / fading with length* (similar with long-context LLMs).

Static Visual Embeddings Lack Dynamic Grounding

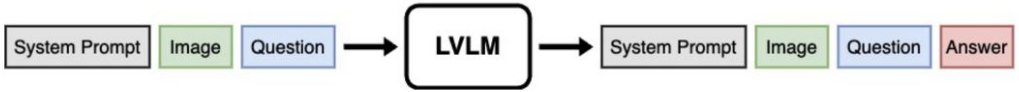
LVLMs struggle with Visual Understanding due to architectural inductive bias

- **Language Bias**: The model **prioritizes language patterns** and **focuses on partial tokens** rather than fully grounding the generated content in both visual and textual context.
- **Sparse (and Smooth) Attention (on Visual Tokens)**: As the *Key* set in *Attention* grows, standard softmax tends to **spread mass more evenly**, weakening focus on truly relevant tokens.
- **Static Visual Embedding**: Vision features are past through **a fixed projection** and then processed into **a fixed set of visual tokens** as the input to LLM backbone.
  - Information Bottleneck or Dilution ← the budget is not adaptable for specific context/queries.
  - Weaker Grounding ← context-agnostic mapping under-emphasizes essential visual evidence.

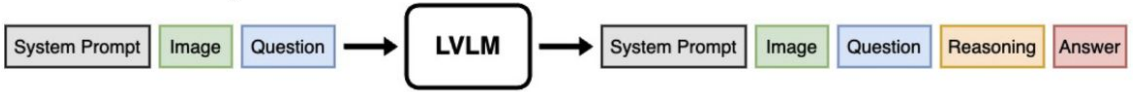
Interleaved Vision Language Reasoning

VL Prediction Objectives: Language Head + Additional Vision Head

- Direct Visual Question Answering



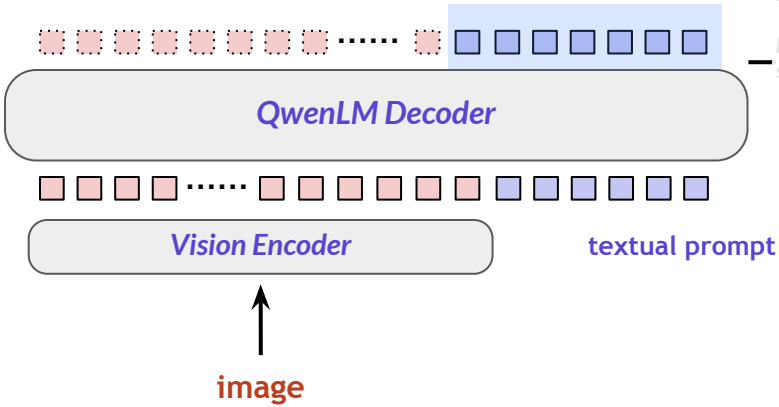
- VL-CoT Reasoning



evidence serves
l deduction.

Interleaved Vision Language Reasoning

VL Prediction Objectives: Language Head + Additional Vision Head



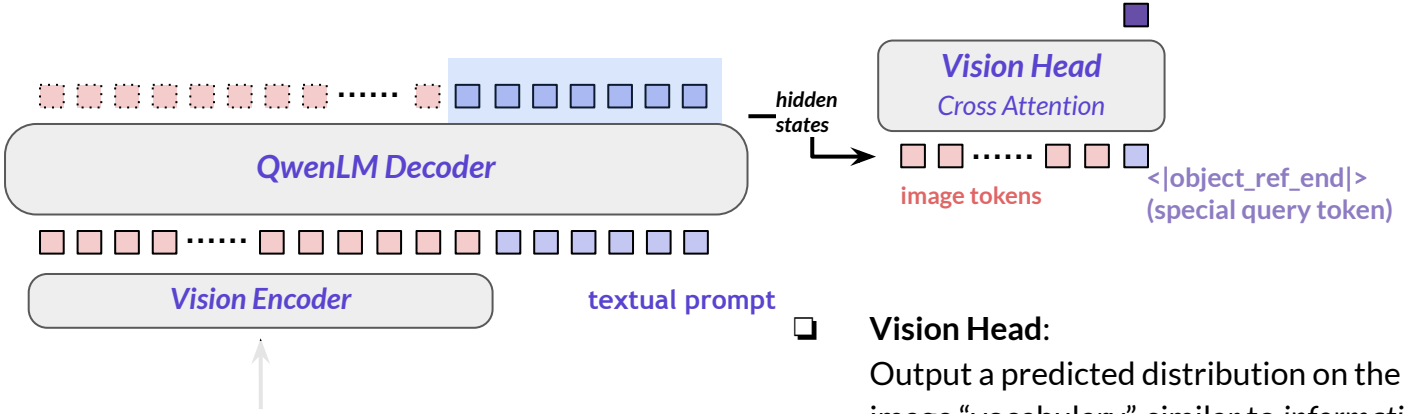
❑ **Language Head**

Each token is determined according to the predicted distribution on the textual vocabulary.

$$\mathcal{L}_{\text{text}} = -\sum_{i \in \mathcal{I}} \log p_\theta(y_i \mid y_{<i}, \text{image})$$

Interleaved Vision Language Reasoning

VL Prediction Objectives: Language Head + Additional Vision Head



$$\mathcal{L}_{\text{BCE}} = \sum_{t,k} \Big[- B_{t,k} \log \hat{S}_{t,k} - (1 - B_{t,k}) \log(1 - \hat{S}_{t,k}) \Big]$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_{t,k} B_{t,k} \hat{S}_{t,k}}{\sum_{t,k} B_{t,k} + \sum_{t,k} \hat{S}_{t,k} + \epsilon}$$

$$\mathcal{R}_{\text{ent}} = \pm \sum_{t} \mathsf{H}(\hat{\mathbf{P}}_t)$$

- **Vision Head**:
  Output a predicted distribution on the input image "vocabulary", similar to *information retrieval*.

$$\mathcal{L}_{\text{vision}} = \mathcal{L}_{\text{VL-basic}} + \alpha \mathcal{R}_{\text{ent}}$$

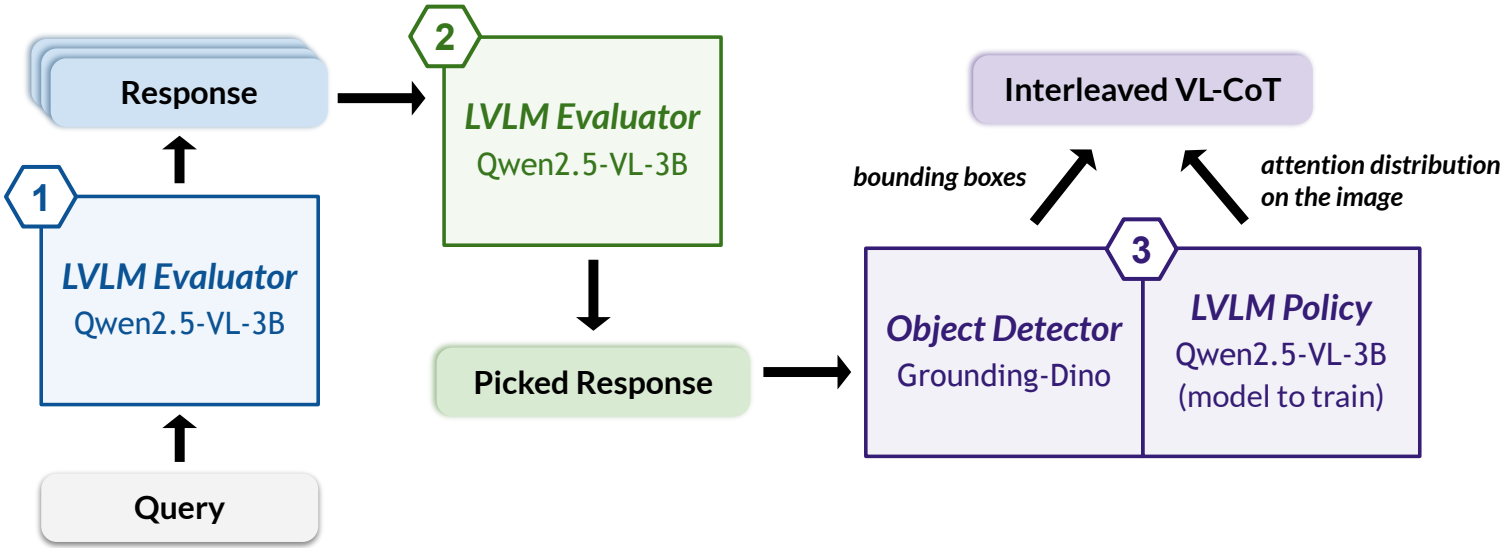
$$\downarrow$$

$$\mathcal{L}_{\text{VL-basic}} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{Dice}}$$

Interleaved Vision Language Reasoning

Rejection Sampling Tuning Guided by Automatic Scores

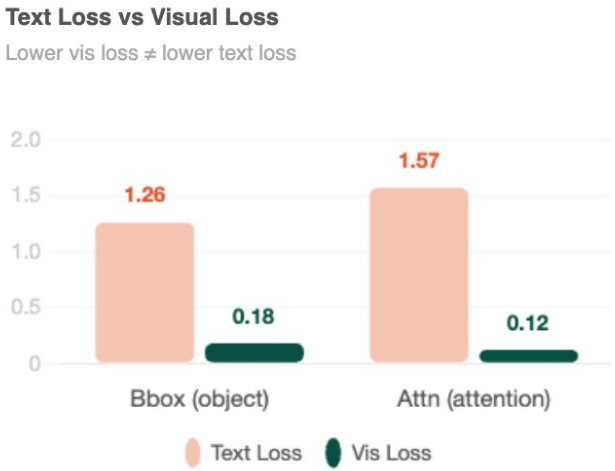
❑ **Data Collection Pipeline**: backboned by the base model itself



VoCoT: Visually Grounded Object-centric Chain-of-Thought Reasoning (Li et al., 2024)

❑ Post-training Misalignment

- While the overall loss could be reduced to a certain level, the model still struggles to confidently identify the **injection point for visual self-reference**.

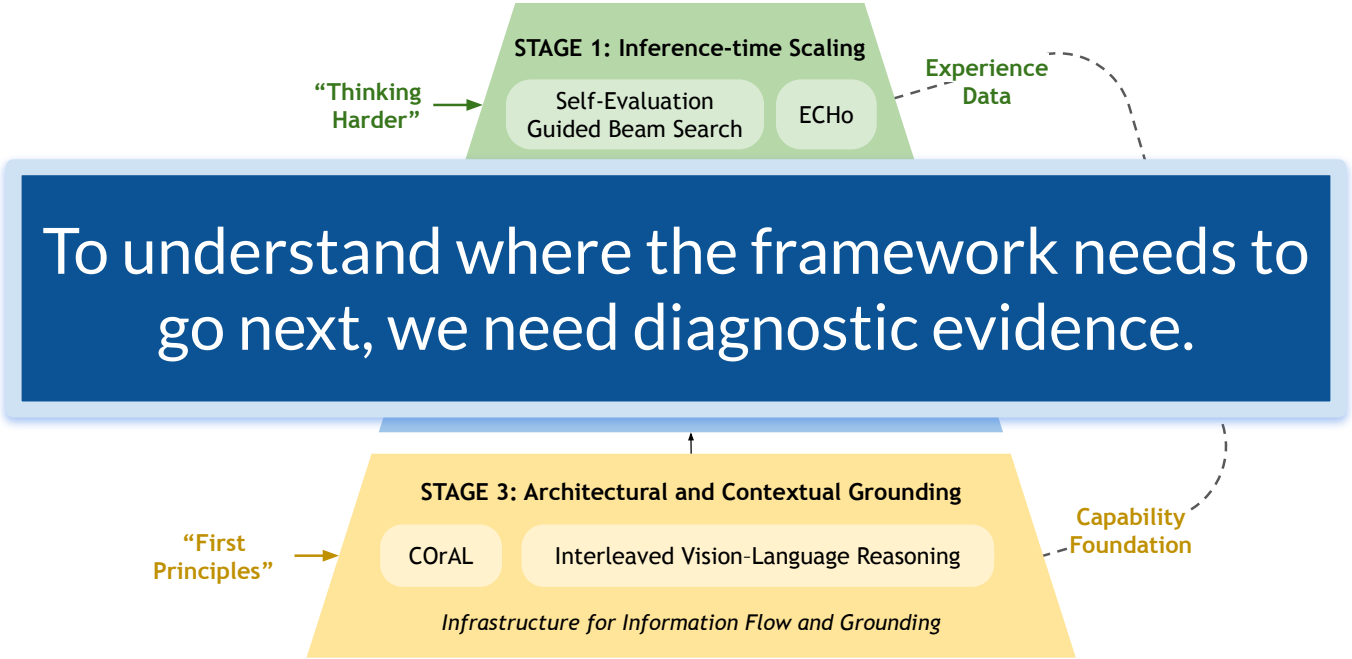


VoCoT: Visually Grounded Object-centric Chain-of-Thought Reasoning (Li et al., 2024)

- **Divergence** between model- vs. function-based scores
  - Current evaluators may reward coherence and the appearance of grounding more than strict rule-based correctness.

Recap: The Closed-Loop Automatic Scaling Framework

Three Layers: Search → Alignment → Grounding



Content

- ❑ **PART 1**: Introduction
- ❑ **PART 2**: Layer One – Inference-time Scaling
- ❑ **PART 3**: Layer Two – Training-time Scaling
- ❑ **PART 4**: Layer Three – Architectural & Contextual Grounding
- ❑ **PART 5**: Reflections & Diagnosis
- ❑ **PART 6**: Future Directions & Closing

Key Learnings

- ❏ Self-evaluation calibration determines the ceiling. ← Self-Evaluation Guided Beam Search
- ❏ Heterogeneous Verification > Self-Referential Signals ← MCTS-IPL & V-DPO
- ❏ Task-specific formulations trade generalizability for accuracy. ← ECHo
- ❏ Inference-time and training-time scaling are complementary. ← MCTS-IPL & V-DPO & Interleaved VL-CoT
- ❏ Post-training grounding incurs a training tax. ← COrAL & Interleaved VL-CoT

Limitations

- **Coherence–correctness gap**
  - systemic, not fixable without external signals

← Self-Evaluation Guided Beam Search, MCTS-IPL, COrAL, & Interleaved VL-CoT

- **Computational cost**
  - Iterative evaluation & search are expensive.

← Self-Evaluation Guided Beam Search & MCTS-IPL

- **Domain sensitivity**
  - Reasoning process transfers, knowledge doesn't.

← MCTS-IPL & V-DPO

- **Trustworthiness**
  - Structured domains: reliable; Open-ended: risky

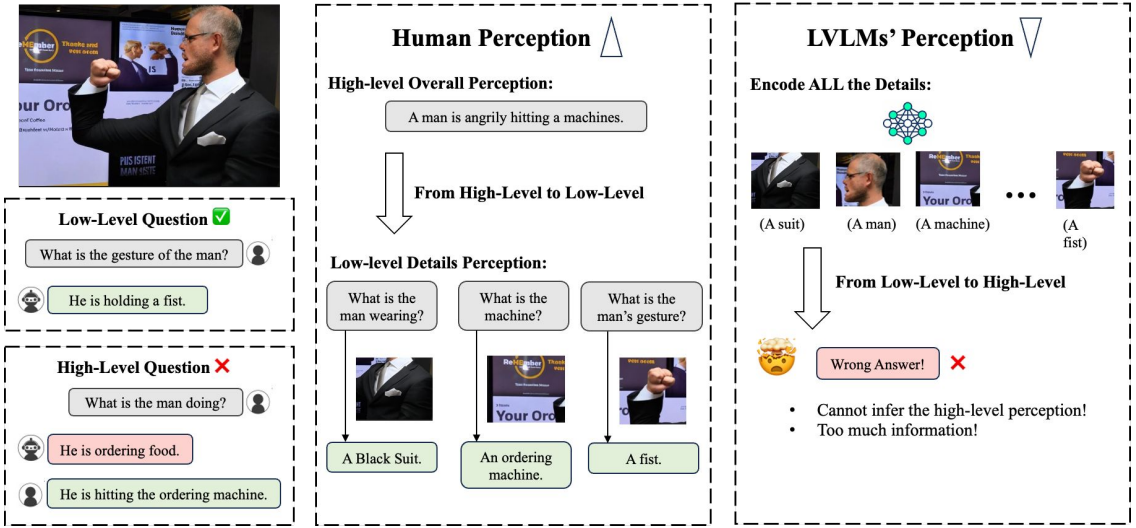
← Self-Evaluation Guided Beam Search, ECHo, & MCTS-IPL

What Could Have Been Done Differently?

- ❑ Earlier external verification. ← Self-Evaluation Guided Beam Search **&** MCTS-IPL
- ❑ Tree Search vs. Instance-level Sampling ← MCTS-IPL **vs.** DeepSeek-R1
- ❑ Pretraining COrAL with order-agnostic modeling ← COrAL
- ❑ Continual pretraining for interleaved VL ← Interleaved VL-CoT

Can LVLMs perceive at multiple levels like humans?

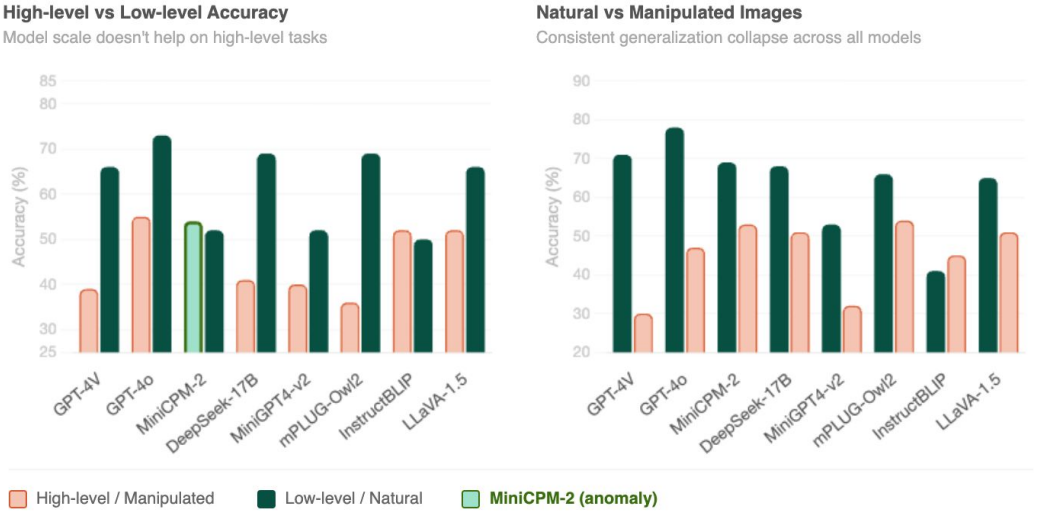
- 2 levels of visual perception
  - *low-level* addresses **factual and physical** understanding of individual image elements.
  - *high-level* requires **focused and combinatorial** reasoning across different regions of the image.



The high-level perception is where the LVLMs may fall short.

Diagnosing LVLMs via MVP-Bench

- GPT-4o (June, 2024): *high-level* 56% versus *low-level* 74%
- **Model scale doesn't help on high-level tasks:** MiniCPM-V-2 (3B) > larger models
- **Generalization collapse**: *natural images* 77% → *manipulated images* 49%



Diagnosing LVLMs via MVP-Bench

Implications

- The *high-level* visual perception deficit sits at the intersection of all three layers.
- Correctness–Appearance Gap ← confirmed at perception level

This is what the next generation of the scaling loop needs to address.

Content

- **PART 1**: Introduction
- **PART 2**: Layer One – Inference-time Scaling
- **PART 3**: Layer Two – Training-time Scaling
- **PART 4**: Layer Three – Architectural & Contextual Grounding
- **PART 5**: Reflections & Diagnosis
- **PART 6**: Future Directions & Closing

Future Directions

❑ **From *Static* Diagnosis to *Dynamic* Evaluation**

- **Recursive self-evaluation scaling**: To develop a meta-loop for the evaluator to ***co-evolve*** with the policy it supervises.
  - Incompleteness (*sparsity*) & Incorrectness (*bias*, *noise*) of "ground truth"
  - Co-evolving evaluators: Detecting the imperfection → Evolving

Heterogeneous Verification > Self-Referential Signals

MCTS-IPL & V-DPO

Coherence–Correctness Gap

Self-Evaluation Guided Search, MCTS-IPL, COrAL, Interleaved VL-CoT, & MVP-Bench

Trustworthy & Adaptability

Self-Evaluation Guided Search, ECHo, MCTS-IPL, & MVP-Bench

Future Directions

❑ **From *Static* Diagnosis to *Dynamic* Evaluation**

- **Recursive self-evaluation scaling**: To develop a meta-loop for the evaluator to ***co-evolve*** with the policy it supervises.
  - Incompleteness (*sparsity*) & Incorrectness (*bias*, *noise*) of "ground truth"
  - Co-evolving evaluators: Detecting the imperfection → Evolving
- **Environmental feedback via world models**: Rather than the evaluator discovering new dimensions *internally*, the world model supplies ***external*** feedback grounded in reality.

Future Directions

- ❑ **From *Static* Diagnosis to *Dynamic* Evaluation**
  - Recursive self-evaluation scaling.
  - Environmental feedback via world models.
- ❑ **From *Model Reasoning* to *Agentic Action***
  - Single model → Multi-agent system

  Self-Evaluation Guided Search, ECHo
  MCTS-IPL *vs.* V-DPO
- ❑ **Decoupling Reasoning from Language: Pre-Linguistic Foundations of Intelligence**

  MCTS-IPL & V-DPO
  - Exploit the separability of reasoning and (modality-based) knowledge at pretraining.
- ❑ **Native Multimodal Scaling**

  ECHo, V-DPO, Interleaved VL-CoT, & MVP-Bench
  - Joint vision–language (and other modalities) understanding & reasoning from pretraining.

Thanks to my Advisor, Committee Members, Collaborators, and Funding Support.

Yuxi Xie

National University of Singapore, Singapore

Advisor: Associate Professor Min-Yen Kan