# Can LLMs Reason Over Miscellaneous Structures Without Wasting Tokens?

## Scaling-Aware Adapter for Structure-Grounded LLM Reasoning

*Authors: Zihao Jing, Qiuhao Zeng, Ruiyi Fang, Yan Yi Li, Yan Sun, Boyu Wang, Pingzhao Hu*

ICML International Conference On Machine Learning

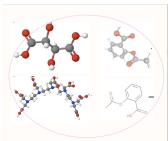
Western University · Canada

### Research Question

**Our Goal:** Design an adaptive connector between LLMs and structures, enabling LLMs to reason over diverse levels of structural complexity.

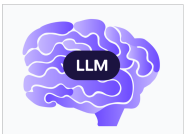
### Core Answer

Cuttlefish selects instruction-related regions, grows geometry-aware patches, and injects refined structural evidence into the LLM.



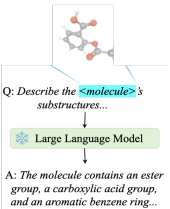
Entity 3D Structures

Understand

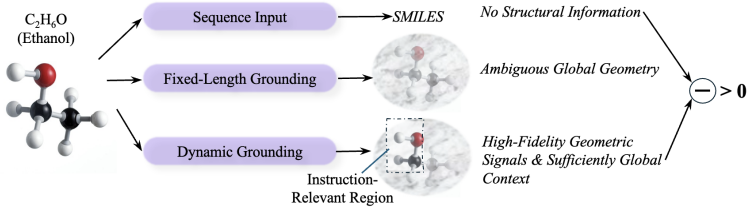



Foundation Models

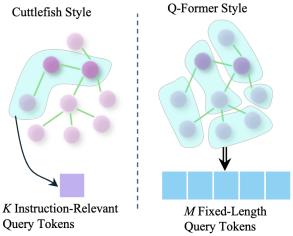
Inference/ Reasoning



# Challenges and Research Gaps



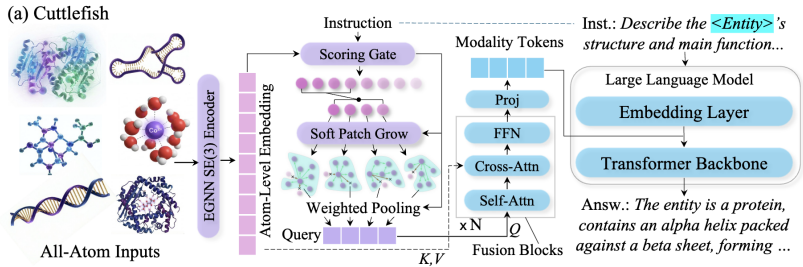

Sequence-only and fixed-length grounding miss instruction-related geometry.



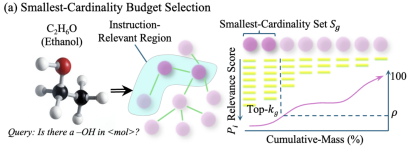

Dynamic selection is needed for entities of varying sizes.

| Instruction-Related Geometry | Sequence Inputs Are Not Enough | Fixed Budgets Hide Detail |
|---|---|---|
| Locating instruction-related geometric regions matters. The same entity may require different structural evidence for different requests and questions. | Without explicit geometry injection, LLMs produce structural descriptions that are not grounded in real coordinates. | Q-Former-style connectors compress every 3D entity equally, losing critical geometry for complex structures. |

# Methodology

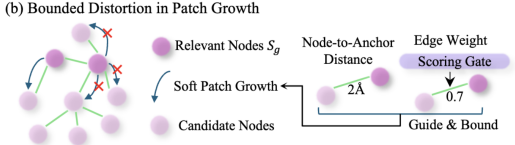



## 1. Select the Important Region



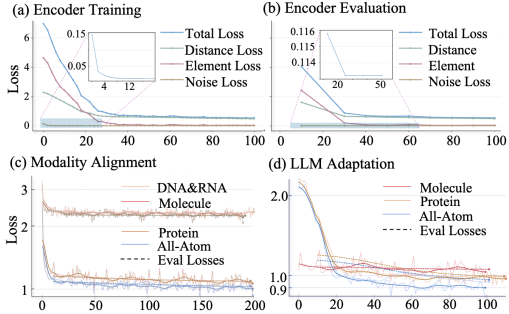

Instruction-conditioned gating selects anchor atoms by cumulative probability mass.

## 2. Refine the Geometry Detail

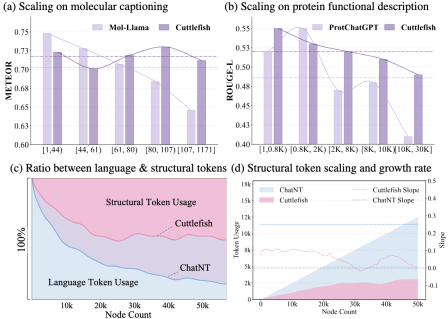



Soft patch growth expands anchors spatially; patch tokens retrieve geometry by cross-attention.

# Insight and Analysis




Mixed-modality tuning converges faster and more effectively than single-modality tuning.




Cuttlefish remains stable for large entities; adaptive tokens grow sublinearly compared with ChatNT.

| Convergence | Scaling Ability | Token Efficiency |
|---|---|---|
| Mixed-modality tuning converges faster; cross-entity chemical diversity improves all-atom generalization. | As entity size increases sharply, Cuttlefish maintains stable performance while the baseline declines. | Adaptive tokens scale sublinearly compared with ChatNT while preserving reasoning capacity. |

# Experiments and Results

| Model | Molecule | | Protein | | DNA | | RNA | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|
| | METEOR | BERT-S | METEOR | BERT-S | METEOR | BERT-S | METEOR | BERT-S | METEOR | BERT-S |
| *General LLM Baselines - Sequence Only - Non-Reasoning* | | | | | | | | | | |
| Qwen-2.5-7B-Instruct | 0.137 | 0.691 | 0.091 | 0.651 | 0.171 | 0.644 | 0.159 | 0.640 | 0.143 | 0.653 |
| Llama-3.1-8B-Instruct | 0.229 | 0.778 | 0.178 | 0.742 | 0.175 | 0.658 | 0.175 | 0.646 | 0.186 | 0.694 |
| Mistral-3-8B-Instruct-2512 | 0.185 | 0.732 | 0.134 | 0.714 | 0.156 | 0.665 | 0.192 | 0.652 | 0.172 | 0.683 |
| GLM-4-9B-0414 | 0.174 | 0.644 | 0.110 | 0.672 | 0.204 | 0.690 | 0.143 | 0.549 | 0.155 | 0.621 |
| *General LLM Baselines - Reasoning - Sequence Only* | | | | | | | | | | |
| Qwen3-8B | 0.103 | 0.715 | 0.044 | 0.679 | 0.131 | 0.665 | 0.128 | 0.655 | 0.107 | 0.674 |
| DeepSeek-R1-D-Qwen-7B | 0.160 | 0.721 | 0.141 | 0.707 | 0.206 | 0.659 | 0.170 | 0.687 | 0.169 | 0.692 |
| Mistral-3-8B-Reasoning-2512 | 0.147 | 0.753 | 0.108 | 0.811 | 0.264 | 0.806 | 0.180 | 0.674 | 0.176 | 0.744 |
| *General LLM Baselines - Reasoning - Sequence Only - Modality Enhanced Tokenizer* | | | | | | | | | | |
| Qwen3-8B | 0.158 | 0.709 | 0.027 | 0.573 | 0.177 | 0.792 | 0.095 | 0.773 | 0.110 | 0.724 |
| DeepSeek-R1-D-Qwen-7B | 0.204 | 0.673 | 0.219 | 0.617 | 0.147 | 0.644 | 0.283 | 0.688 | 0.227 | 0.662 |
| Mistral-3-8B-Reasoning-2512 | 0.185 | 0.823 | 0.192 | 0.755 | 0.149 | 0.756 | 0.288 | 0.579 | 0.220 | 0.698 |
| *Ours* | | | | | | | | | | |
| Cuttlefish + Qwen2.5-7B | 0.314 | 0.863 | 0.317 | 0.816 | 0.262 | 0.806 | 0.378 | 0.858 | 0.330 | 0.840 |
| Cuttlefish + Llama-3.1-8B-I | 0.391 | 0.875 | 0.417 | 0.896 | 0.529 | 0.816 | 0.403 | 0.868 | 0.428 | 0.864 |
| Cuttlefish + Mistral-3-8B-I | 0.415 | 0.653 | 0.333 | 0.858 | 0.358 | 0.852 | 0.220 | 0.694 | 0.310 | 0.750 |
| Cuttlefish + GLM-4-9B | 0.327 | 0.830 | 0.262 | 0.831 | 0.443 | 0.888 | 0.403 | 0.794 | 0.367 | 0.827 |
| Cuttlefish + Qwen3-8B | 0.389 | 0.853 | 0.377 | 0.888 | 0.391 | 0.860 | 0.491 | 0.890 | 0.428 | 0.876 |
| Cuttlefish + R1-7B | 0.342 | 0.812 | 0.327 | 0.801 | 0.330 | 0.802 | 0.422 | 0.799 | 0.369 | 0.803 |
| Cuttlefish + Mistral3-8B-R | 0.309 | 0.776 | 0.320 | 0.756 | 0.256 | 0.798 | 0.395 | 0.775 | 0.335 | 0.776 |

Entity captioning performance is evaluated on four modalities from the GEO-AT dataset.

**Geometry Grounding Is Essential**

Augmenting any LLM backbone with Cuttlefish yields substantial improvements, confirming structural grounding is necessary.

**Unified Superiority**

Cuttlefish achieves top-1 performance on most tasks across all modalities, outperforming modality-specific models.

# Experiments and Ablation Studies

| Connector | Token policy | Molecule↑ | Protein↑ | DNA&RNA↑ | Average↑ |
|---|---|---|---|---|---|
| Q-Former | Fixed 256 | 0.778 | 0.743 | 0.658 | 0.726 |
| Q-Former | Fixed 512 | 0.781 | 0.772 | 0.680 | 0.744 |
| Q-Former | Fixed 1024 | 0.809 | 0.784 | 0.712 | 0.768 |
| Q-Former | Fixed 2048 | 0.768 | 0.767 | 0.745 | 0.760 |
| Cuttlefish | Matched to 256 | 0.842 | 0.659 | 0.693 | 0.731 |
| Cuttlefish | Matched to 512 | 0.853 | 0.776 | 0.748 | 0.792 |
| Cuttlefish | Matched to 1024 | 0.875 | 0.798 | 0.750 | 0.808 |
| Cuttlefish | Matched to 2048 | 0.875 | 0.896 | 0.816 | 0.862 |

Matched-budget comparison: Cuttlefish outperforms fixed-length Q-Former at every budget.

| | Func.-group | | Mech.-expl. | | L.-range | | Avg. | |
|---|---|---|---|---|---|---|---|---|
| Model | HR | AR | HR | AR | HR | AR | HR | AR |
| *Molecule* | | | | | | | | |
| Mol-LLaMA | 0.12 | 0.94 | 0.23 | 0.98 | 0.19 | 1.00 | 0.18 | 0.97 |
| 3D-MoLM | 0.23 | 0.83 | 0.72 | 0.95 | 0.32 | 0.81 | 0.42 | 0.86 |
| Cuttlefish | 0.07 | 0.99 | 0.13 | 0.99 | 0.16 | 1.00 | 0.12 | 0.99 |
| *Protein* | | | | | | | | |
| ProtChatGPT | 0.10 | 0.94 | 0.20 | 0.96 | 0.24 | 1.00 | 0.18 | 0.97 |
| Prot2Chat | 0.06 | 0.95 | 0.71 | 0.89 | 0.33 | 0.76 | 0.37 | 0.86 |
| Cuttlefish | 0.04 | 0.97 | 0.16 | 0.98 | 0.20 | 0.98 | 0.13 | 0.98 |

Hallucination test: Cuttlefish achieves lower hallucination rates across scenarios.

## Adaptivity Drives Gains

At matched token budgets, Cuttlefish consistently outperforms fixed-length Q-Former, isolating adaptivity as the key factor.

## Hallucination Is Controlled

With adaptive multimodal alignment, Cuttlefish outperforms molecule and protein LLM baselines for hallucination control.

# Connect

Paper

Code

## Cuttlefish

Scaling-aware adapter for structure-grounded LLM reasoning


**Contact**

Zihao Jing *zjing24@uwo.ca* Web: *zihao-jing.org*

Corresponding: Pingzhao Hu *phu49@uwo.ca*


`github.com/zihao-jing/Cuttlefish`


**Support**

Canada Research Chairs Tier II Program; Canadian Institutes of Health Research; Natural Sciences and Engineering Research Council of Canada; Canada Foundation for Innovation JELF program.


Western University · Canada

NSERC CRSNG

CIHR IRSC
Canadian Institutes of Health Research | Instituts de recherche en santé du Canada