Structured Reasoning in Vision-Language Models: A Comparative Study of Chain-of-Thought Prompting Formats

Ann Xiao
yx3493@nyu.edu

Xinyu Li
xl5280@nyu.edu

New York University

Abstract

Chain-of-Thought (CoT) prompting has been shown to improve reasoning in language models, but its role in multimodal scientific domains remains underexplored. In this work, we systematically evaluate how different CoT prompt structures affect performance in vision-language models (VLMs) across five chemistry tasks involving molecular image inputs. We compare four prompt formats—Explanation-first, Stepwise, Visual-first, and Baseline—using three open-source VLMs. Our findings show that structured prompts, particularly Explanation-first and Stepwise formats, tend to improve both accuracy and interpretability, while Visual-first and Baseline prompts often result in shallower or heuristic reasoning. We further analyze how prompt-task alignment influences factual correctness, reasoning depth, and user trust. These results highlight the importance of prompt design for enabling reliable scientific reasoning in multimodal AI systems.

All code, prompts, and evaluation data are available at: `https://github.com/AX1ao/ExplainIt`.

1 Introduction

Chain-of-Thought (CoT) prompting has been shown to enhance reasoning in language models, especially for arithmetic, logic, and commonsense problems [9]. By encouraging models to articulate step-by-step rationales, CoT prompts serve as cognitive scaffolds—structuring the model's reasoning in ways inspired by human problem-solving [4]. However, CoT's benefits in **multimodal scientific tasks**, where both visual perception and domain-specific logic are essential, remain poorly understood.

This gap is especially salient in chemistry, medicine, and related fields, where vision-language models (VLMs) are increasingly deployed. In such contexts, correct answers often require **interpreting molecular structures** while applying **specialized chemical reasoning** (e.g., resonance, nucleophilicity, reactivity mechanisms). Scientific tasks therefore impose dual demands: accurate visual grounding and structured domain reasoning. We hypothesize that the *form of the prompt*—how the model is cued to reason—can significantly affect whether outputs are accurate, interpretable, and trustworthy.

In this study, we systematically investigate how **different CoT prompt structures impact scientific reasoning in VLMs**. We design five chemistry-based tasks using molecular diagrams as input and compare four prompt formats: **Explanation-first**, **Stepwise**, **Visual-first**, and a **Baseline** condition with no scaffolding. The tasks span compound identification, mechanism prediction, and comparative reactivity, covering a range of reasoning demands.

Our results show that **prompt structure strongly influences both model accuracy and reasoning depth**, even in a zero-shot setting. Structured formats such as Explanation-first and Stepwise

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

consistently outperform Visual-first and Baseline prompts, especially in tasks requiring abstract domain logic. These findings suggest that **aligning prompt structure with task demands is critical** for building trustworthy scientific reasoning in multimodal models.

2 Related Work

Chain-of-Thought (CoT) prompting was introduced by Wei et al. [9] to improve symbolic reasoning in language models. By encouraging step-by-step verbalization, CoT prompting significantly improved performance on arithmetic and commonsense benchmarks. Zhu et al. [14] extended this approach to vision-language models (VLMs), proposing a two-stage pipeline that separates rationale generation from answer selection, establishing a basis for multimodal CoT prompting.

Subsequent work has introduced diverse CoT structures tailored to specific reasoning demands:

- **Compositional CoT** [5]: Constructs intermediate scene graphs before generating reasoning.
- **Duty-Distinct CoT** [12]: Divides reasoning into perception and inference phases.
- **Contrastive CoT** [11]: Promotes discrimination by prompting the model to compare similar images.
- **Answer-then-explain CoT (PEAR-CoT)** [3]: Inverts the CoT order, generating the answer first and then explaining it to reduce error propagation.

Other research has explored modularity and knowledge augmentation. Gao et al. [2] simulate expert agents that select reasoning paths. Mondal et al. [6] inject commonsense knowledge from ConceptNet into CoT chains (KAM-CoT), and Zhou et al. [13] propose *Image-of-Thought CoT*, which prioritizes perceptual inspection before reasoning.

In addition, Lu et al. [4] and Chen et al. [1] explore how CoT behaviors can be taught or distilled using lecture-style prompts and multi-grain scaffolds aligned to different reasoning levels. Wang et al. [8] demonstrate that even well-structured CoT prompts can be circumvented by adversarial inputs, raising questions about prompt robustness.

Most recently, Wang et al. [7] introduced **T-SciQ**, a hybrid prompting system that adaptively combines plan-based and standard CoT formats for science QA. Their results emphasize the importance of tailoring CoT structure to task complexity and logic.

Taxonomy and Scope. Yang et al. [10] provide a comprehensive taxonomy of CoT prompting strategies, classifying them into five families: *Stepwise*, *Modular*, *Visual-first*, *Contrastive*, and *Meta/Teaching-based*. These structures vary in how they scaffold reasoning and interface with model architecture. While most prior work focuses on general NLP or vision-language benchmarks, our study focuses specifically on chemistry-based multimodal reasoning and how CoT prompt structure impacts accuracy, interpretability, and trust.

Table 1: Taxonomy of CoT Structures in Multimodal Reasoning

Family	**Representative Works**	**Description**
Stepwise / Sequential	Wei et al., Lu et al., Li et al.	Classical CoT format with named substeps building toward a final answer.
Modular / Role-based	Zheng et al., Gao et al., Chen et al.	Reasoning is distributed across perception and inference modules or agents.
Visual-first / Tool-guided	Zhou et al., Mondal et al., Mitra et al.	Reasoning begins with perceptual features or visual tools.
Contrastive / Multi-Scene	Zhang et al.	CoT derived from comparing paired or contrasting images.
Meta / Teaching CoT	Lu et al., Chen et al., Yang et al.	CoT behavior is taught or distilled using explanation scaffolds.

3 Method

3.1 Task Design

We designed a five-part chemistry task suite to evaluate scientific reasoning in vision-language models (VLMs). Each task presents a pair of molecule images and requires the model to make a comparative or classificatory judgment based on domain-specific logic:

- **Molecule Identification**: Name the compound depicted in the image.
- **EAS Reactivity**: Predict which molecule is more reactive in electrophilic aromatic substitution.
- **Acid/Base Strength**: Determine which molecule is a stronger acid or base.
- **Functional Group Reactivity**: Identify which of two molecules is the better nucleophile.
- **SN1 Reactivity**: Judge which molecule is more likely to undergo an SN1 reaction.

These tasks span a range of reasoning types, from visual recognition to mechanism-driven inference, and serve as a foundation for analyzing how prompt structure influences model behavior.

3.2 Prompt Design: Chain-of-Thought Families

We evaluated four Chain-of-Thought (CoT) prompt formats, each reflecting a distinct structural strategy:

- **Stepwise Reasoning**: Prompts are divided into named substeps aligned with domain-relevant logic. Example: *Step 1: Identify nucleophilic atoms. Step 2: Assess resonance. Step 3: Decide and explain.*

- **Explanation-First Reasoning**: Prompts begin with relevant chemical principles, followed by application. Example: *SN1 requires a stable carbocation. Consider which molecule fits this principle.*

- **Visual-First Reasoning**: Prompts begin with visual inspection before invoking chemical concepts. Example: *Look for OH, NH_2, or NO_2 groups and use their positions to judge which ring is more reactive.*

- **Baseline (Non-CoT)**: Prompts use direct questioning with no explicit reasoning scaffolding. Example: *Which molecule is the stronger acid or base, and why?*

Excluded Formats. We excluded contrastive and training-time CoT formats (e.g., CoCoT, Meta-CoT). Contrastive prompting was unnecessary because Tasks 1–4 were already designed as pairwise comparisons. Our focus is strictly on inference-time, prompt-only interventions applied to comparative or single-image inputs.

External vs. Internal CoT. This study evaluates only externally visible reasoning produced in model outputs. We do not analyze latent reasoning processes, attention patterns, or internal activations.

3.3 Model and Inference Pipeline

We evaluated three open-source vision-language models in a zero-shot setting:

- **LLaVA-Med (v1.5, Mistral-7B)**: A biomedical VLM, executed on Google Colab using NVIDIA A100 GPU.
- **LLaVA-Onevision**: A general-purpose VLM with strong visual grounding, running on NYU CIMS CUDA2 cluster.
- **DeepSeek-VL**: A multilingual VLM, run on Google Colab using an NVIDIA A100 GPU.

Each model was tested across all five tasks using fixed decoding parameters (e.g., temperature, max tokens). No fine-tuning, in-context examples, or demonstrations were used. All prompt-image pairs were evaluated independently in a fully zero-shot setting.

While hardware environments varied slightly, prompt structure was the primary variable of interest. We note that minor variation in runtime behavior may arise from backend differences but does not affect our comparative analyses.

3.4 Scoring and Evaluation Protocol

Human evaluation was conducted using task-specific rubrics, carefully designed to assess both factual correctness and the quality of chemical reasoning.

- For **Task 0** and **Task 1**, outputs were scored on a 5-point scale ranging from 0 to 2.0, with intermediate intervals (0.5, 1.0, 1.5) to reflect more granular levels of partial correctness:
  - **0**: Completely incorrect; no meaningful recognition or reasoning.
  - **0.5**: Minimal recognition or relevance;
  - **1.0**: Some key structures or partial rationale captured, but final answer is incorrect.
  - **1.5**: Most relevant structures or reasoning steps are correct, with only minor flaws.
  - **2.0**: Fully correct identification or prediction, with consistent, detailed, and chemically sound reasoning.
- For **Task 2–4**, due to time constraints, we adopted a simpler 3-point rubric that maps to the same 0–2 score range:
  - **0**: Incorrect answer or off-topic/irrelevant explanation.
  - **1**: Partially correct answer with shallow, incomplete, or partially flawed reasoning.
  - **2**: Fully correct answer with reasonable and chemically valid justification.
  - **2*** (optional): Fully correct answer with particularly strong domain-consistent reasoning (used for internal annotation but not plotted separately).

To ensure consistency, all evaluations were performed by a single domain-aware rater using a shared rubric across tasks. Where applicable, cross-task calibration checks were conducted to reduce subjectivity and maintain scoring reliability. Additionally, for Task 0, we recorded the use of hedging language (e.g., *may*, *likely*, *suggested to be*) as a separate signal of epistemic uncertainty, especially when predictions were partially or fully incorrect.

3.5 Evaluation Goals

This experimental setup enables us to address the following questions:

- Which CoT prompt formats are most effective across different types of chemistry tasks?
- Do structured prompts generalize across tasks and model architectures?
- How does prompt structure affect factuality, reasoning depth, and interpretability?

4 Experiment

4.1 Dataset Description

A set of 55 chemical structure images was curated, all sourced from Wikimedia Commons to ensure standardization and public availability. To reduce potential confounding factors and ensure experimental consistency, all images were resized to approximately the same dimensions prior to model inference.

Task 0 uses the full dataset for evaluation on DeepSeek-VL and LLaVA-Med, while OneVision was evaluated on a subset due to computational constraints. Tasks 1 through 4 each involve 10 molecular comparison pairs and were evaluated across all three models. The dataset design supports controlled comparison of prompt strategies across diverse chemical reasoning contexts.

4.2 Prompt Pool and Coverage

Each chemistry task included 10 molecule pairs and designed 10–16 prompt variants per pair, covering all four CoT prompt families. These prompts were manually developed and refined to represent diverse reasoning strategies. Final prompt selections were based on pilot evaluations and were applied consistently across all models, resulting in 400–600 completions per task per model.

4.3 Prompt Selection and Filtering

Initial prompt selection was based on pilot runs using the OneVision model. For Tasks 0–3, we tested 20–50 variants per CoT format on representative image pairs and selected the top-performing prompt in each family based on average score and reasoning quality. These prompts were then used consistently across all three models for final evaluation.

Due to higher task complexity, Task 4 (SN1 Reactivity) used an extended selection procedure. We evaluated 20 variants per CoT format across the full test set using OneVision and selected the top three prompts per format. This resulted in 10 prompts per image pair (three CoT variants plus one baseline) for subsequent testing on DeepSeek-VL and LLaVA-Med.

This two-stage selection process ensured that prompts were both high-performing and adaptable across models and tasks, while minimizing overfitting to a specific configuration.

4.4 Grading Consistency and Limitations

While all tasks were scored using a 0–2 scale, there were structural differences between Task 0–1 and Task 2–4 that introduce potential inconsistencies. Task 0 and Task 1 employed a more fine-grained 5-point rubric (0, 0.5, 1.0, 1.5, 2.0) to capture partial recognition and nuanced chemical reasoning, whereas Tasks 2–4 used a coarser 3-point scale (0, 1, 2), with an optional 2* mark for outstanding responses. For cross-task comparison and averaging, we treated 1.5 as equivalent to 2.0, and 2.0 as equivalent to 2*, to reduce the scoring gap and better align the evaluative granularity.

All model outputs for each task were scored by a single domain-expert rater to ensure consistency within tasks. While rubric alignment and intra-task checks were applied, the use of a single annotator introduces subjectivity. Future work should adopt multi-rater protocols and inter-rater agreement measures to enhance scoring reliability.

5 Results

5.1 Overall Performance Summary

Structured Chain-of-Thought (CoT) prompts significantly outperformed baselines across all five tasks. **Explanation-first** and **Stepwise** prompts achieved the highest average scores, producing more accurate and chemically coherent outputs.

Table 2: Average Prompt Type Performance Across Tasks (OneVision)

Prompt Type	Average Score (/2)	Notes
Explanation-first	**2.0**	Consistently produced well-reasoned, accurate answers
Stepwise	**2.0**	Performed strongly in multi-step mechanistic reasoning
Baseline	1.6	Often guessed correctly but lacked deep reasoning
Visual-first	1.0	Relied on heuristics, lacked chemical grounding

Explanation-first prompts achieved near-perfect scores in four out of five tasks, particularly excelling in contexts requiring abstract comparisons or conceptual knowledge (e.g., resonance, conjugation, acidity). Stepwise prompts performed equally well in structured decision tasks (e.g., SN1 mechanisms, acid-base comparisons), though some responses omitted final decisions.

Fine-tuning prompts for Task 0 improved recognition accuracy for molecules like cholesterol and phenol. Structured prompts also reduced common errors such as misidentified functional groups, mistaken reactivity types, and heuristic overgeneralizations.

LLaVA-Med showed the greatest responsiveness to Explanation-first prompts. OneVision performed reliably with both Explanation and Stepwise formats. DeepSeek-VL struggled with vision-grounded prompts and performed best only when prompted with purely textual Explanation-first formats.

5.2 Task-by-Task Results

Task 0: Molecule Identification. Visual-first and Explanation-first prompts outperformed others after tuning: Visual-first helped anchor attention to rings and functional groups, while Explanation-first guided the model toward naming logic. Stepwise often stalled at substructure descriptions, and Baseline completions frequently relied on guessing or hallucinated content. Additionally, all CoT formats significantly improved hedging behavior, encouraging the use of uncertainty expressions and reducing confidently incorrect statements.

Task 1: EAS Reactivity. Explanation-first prompts effectively guided the model to apply principles like electron donation and resonance, performing well on subtle cases such as `Pyridine vs Benzene`. Stepwise prompts offered the highest overall accuracy through structured reasoning, while Baseline and Visual-first relied on shallow heuristics.

Task 2: Acid/Base Strength. Explanation-first and Stepwise both earned perfect scores, invoking conjugate base stability, pKa, and resonance. Baseline often relied on memorized rules; Visual-first repeated visual heuristics without mechanism.

Task 3: Nucleophilicity. Explanation-first led with precise orbital, resonance, and inductive logic. Stepwise followed closely with clear substep decomposition. Baseline sometimes guessed right but lacked reasoning. Visual-first prompts rarely connected visible structure to reactivity.

Task 4: SN1 Reactivity. A highlight for Stepwise prompts, which clearly walked through LG departure, intermediate formation, and stability. Explanation-first prompts gave high-level reasoning and were the only format to recover correct answers in harder cases (e.g., Pair 4). Visual-first completions were shallow and repetitive.

Summary. Across tasks, Explanation-first and Stepwise consistently delivered higher accuracy and deeper reasoning. Visual-first and Baseline formats were limited in logic depth and more error-prone on complex comparisons.

5.3 Prompt Format Comparison

Table 3: Average Prompt Scores by Task (OneVision)

Prompt Type	Task 0	Task 1	Task 2	Task 3	Task 4
Explanation-first	1.7	2.0	2.0	2.0	2.0
Stepwise	1.0	1.2	2.0	2.0	2.0
Baseline	1.2	1.3	1.6	1.5	1.6
Visual-first	1.1	1.1	1.0	0.9	1.0

Explanation-first. The most consistent and transferable format, particularly strong in abstract reasoning and when models had to cite principles before applying them. It reliably improved both correctness and clarity.

Stepwise. Excelled in tasks with clear substeps (acid/base, SN1). Some generations failed to decide, but when executed fully, they matched Explanation-first in accuracy.

Baseline. Competitive only in obvious or familiar comparisons. Reasoning was often shallow or incorrect even when the answer was right.

Visual-first. The weakest format overall. Responses typically repeated appearance-based heuristics without chemical logic, except in Task 0 where prompt tuning helped.

5.4 Qualitative Examples

EAS Reactivity. Visual: "More atoms near the ring = more reactive."
Explanation: "Aniline donates via lone pair; nitrobenzene withdraws. → Aniline more reactive."

Nucleophilicity. Visual: "More groups."
Stepwise: "Aniline's nitrogen is delocalized; $MeNH_2$ is localized → better nucleophile."

SN1 Reactivity. Visual: "Looks easier to leave."
Explanation: "LG stabilized via nitro resonance → better SN1 candidate."

Hallucination. Baseline: "Phenol has nitrogen."
Explanation/Stepwise: Rarely hallucinated due to structural guidance.

Overthinking. Some Explanation prompts led to speculative outputs, e.g., invoking sterics or alternate mechanisms unnecessarily.

5.5 Failure Modes

- **Hallucination**: Baseline and Visual prompts sometimes invented incorrect functional groups.
- **Heuristics**: Visual-first responses often defaulted to surface-level cues like group size or ring count.
- **Incomplete Chains**: Stepwise prompts occasionally omitted a final decision after listing valid logic.
- **Correct Answer, Wrong Logic**: Some outputs reached the right conclusion with flawed or circular reasoning.
- **Prompt Fragility**: Certain prompt variants (e.g., Explanation_3, Visual_2) produced erratic or nonsensical outputs.
- **Model Refusals**: DeepSeek-VL failed to respond to vision-grounded prompts and only completed text-based ones.

Summary. While structured prompts improved reasoning reliability, models remained vulnerable to hallucinations, overgeneralization, and shallow pattern-matching. Prompt structure is necessary but not sufficient for trustworthy scientific reasoning.

6 Discussion

6.1 Prompt Structure vs. Task Demands

Our results demonstrate that the effectiveness of a prompt format depends heavily on its alignment with the task's underlying reasoning structure.

Explanation-first prompts were most effective for tasks that required conceptual or principle-driven reasoning, such as SN1 reactivity and acid/base comparisons.

Stepwise prompts excelled in multi-step decision tasks but were less useful for open-ended recognition tasks like molecule identification.

Visual-first prompts showed some utility in perceptual tasks but often relied on superficial features in mechanistic reasoning tasks.

Baseline prompts occasionally produced correct answers but lacked logical depth and were more prone to hallucination.

Takeaway. Structured prompts enhance reasoning only when the structure complements the logic required by the task.

6.2 Emergent Reasoning Patterns and Failure Cases

Structured prompts revealed both positive reasoning behaviors and persistent limitations:

- **Mechanistic inference**: Some models correctly inferred resonance or rearrangement logic, even without explicit cues.

- **Speculation**: Certain Explanation-first prompts led to overconfident or unnecessary reasoning chains.
- **Template bias**: Stepwise prompts encouraged repeated phrasing regardless of context.
- **Incomplete chains**: Some responses failed to make a final judgment despite valid reasoning.
- **Heuristic fallback**: Without structure, models defaulted to unreliable visual cues.

Takeaway. Prompt structure unlocks deeper reasoning but can introduce rigidity, overthinking, or indecision if misaligned with task demands.

6.3 Interpretability and User Trust

Prompt format also shaped the interpretability and trustworthiness of outputs:

- **Reasoning alignment**: Explanation and Stepwise formats made it easier to trace how conclusions were derived.
- **Hallucination control**: Structured formats reduced unsupported claims.
- **Transparency vs. persuasion**: Stepwise reasoning was easier to verify; Explanation-first appeared more insightful but also more prone to speculative errors.
- **False confidence**: CoT outputs that sounded convincing could still be factually wrong.

Takeaway. CoT prompting improves interpretability but can also mislead users if the logic is fluent but flawed.

6.4 Limitations

- **Model scope**: We evaluated three open-source models; generalization to other architectures is not guaranteed.
- **Inference-only intervention**: We focused solely on prompt structure without fine-tuning.
- **Manual scoring**: While consistent, using a single rater introduces subjectivity.
- **Visual ambiguity**: Some properties (e.g., resonance) are not directly visible in 2D inputs.
- **Prompt selection**: Final evaluation used one best prompt per format per task (except Task 4). Broader sampling remains future work.
- **Surface-level analysis**: We evaluated only externalized reasoning, not internal model representations.

Model-specific instability. DeepSeek-VL frequently failed to process image inputs, particularly under vision-grounded prompts. The model often returned refusals (e.g., “I cannot analyze the image”) or generated generic text completions unrelated to the visual input. This limited its utility in vision-language reasoning tasks and reduced its comparability to models like OneVision and LLaVA-Med, which more consistently integrated image features during inference.

Takeaway. Our findings highlight important trends but are limited by model scope, evaluation design, and prompt coverage. Future work should explore more diverse tasks, scalable evaluation, and internal reasoning diagnostics.

7 Conclusion

We present a systematic evaluation of how Chain-of-Thought (CoT) prompt structure impacts reasoning in vision-language models applied to chemistry tasks. By comparing four prompt formats across five tasks and three models, we find that structured prompts—especially Explanation-first and Stepwise—consistently improve both accuracy and interpretability.

Explanation-first guided abstract reasoning grounded in domain principles, while Stepwise supported multi-step logic. In contrast, Visual-first and Baseline formats often defaulted to heuristics or produced shallow justifications.

Prompt structure not only affected factual performance, but also shaped how users interpret and trust model responses. However, structured prompting is not a universal solution—it can introduce rigidity or speculative logic when poorly matched to the task.

We conclude that **prompt-task alignment is essential** for enabling trustworthy scientific reasoning in multimodal models. CoT prompting holds significant promise, but its effectiveness depends on thoughtful design aligned with domain and task structure.

Author Contributions

Each team member's role is detailed as follows:

Ann Xiao was responsible for prompt type fine-tuning, generating model outputs for Tasks 0–4 using the OneVision model and for Tasks 1–4 using LLaVA-Med. She also led the analysis of Tasks 2–4 and contributed to writing the final report.

Xinyu Li conducted prompt selection and refinement for Task 0, generated LLaVA-Med outputs for Task 0 and all DeepSeek-VL outputs for Tasks 0–4. She performed analysis for Tasks 0–1, created all visualizations, and contributed to writing the final report.

Both authors reviewed and approved the final manuscript.

References

[1] Rui Chen and Yansong Feng. Chain-of-thought prompt distillation for multimodal named entity recognition and relation extraction. *arXiv preprint arXiv:2306.14122*, 2024.

[2] Xiaochuan Gao, Xitong Ye, Yujie Shen, and et al. Cantor: Inspiring multimodal chain-of-thought of mllm. *ACM MM*, 2024.

[3] Chunyang Li, Xiang Gu, Baobao Peng, and et al. Multimodal pear chain-of-thought reasoning. *ACM MM*, 2024.

[4] Pan Lu, Bosheng Zhang, Yuxiao Qian, and et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

[5] Tapas Nayak Mitra, Sayak Pramanick, and et al. Compositional chain-of-thought prompting for large multimodal models. *arXiv preprint arXiv:2311.17076*, 2023.

[6] Abhishek Mondal, Aayush Saraf, Ranit Paul, and et al. Kam-cot: Knowledge augmented multimodal chain-of-thought reasoning. *Journal of Web Intelligence*, 2024.

[7] Jiahao Wang, Yining Liu, Yiting Xie, Guangyi Tan, Chenguang Zhang, and Caiming Xiong. T-sciq: Teaching multimodal chain-of-thought reasoning for complex science question answering. *arXiv preprint arXiv:2403.12605*, 2024.

[8] Yuxuan Wang, Kun Hu, Chao Wu, and et al. Stop reasoning! when multimodal llm with chain-of-thought reasoning meets adversarial image. *arXiv preprint arXiv:2402.14899*, 2024.

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, and et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[10] Fan Yang, Xisen Li, Shizhu Liu, and et al. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2024.

[11] Yuhan Zhang, Xiangyu Jin, Huanyu Wang, and et al. Cocot: Contrastive chain-of-thought prompting for large multimodal models. *arXiv preprint arXiv:2401.02582*, 2024.

[12] Rui Zheng, Licheng Wang, Xiuyu Tang, and et al. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *arXiv preprint arXiv:2310.16436*, 2023.

[13] Boxin Zhou, Zhiqing Shen, Chong Yuan, and et al. Image-of-thought prompting for visual reasoning. *arXiv preprint arXiv:2405.13872*, 2024.

[14] Jiayao Zhu, Xin Chen, Yu Wang, and et al. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Appendix: Task-Level Statistics and Visualizations

Task 0: Molecule Identification

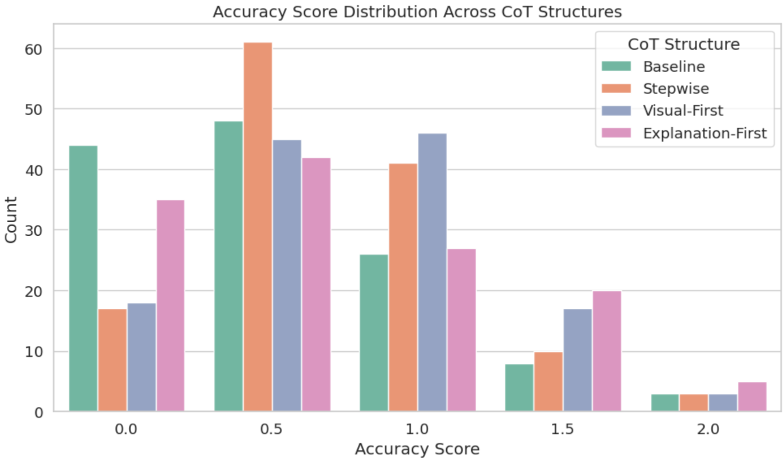

Figure A.1: Task 0 overall accuracy distribution.

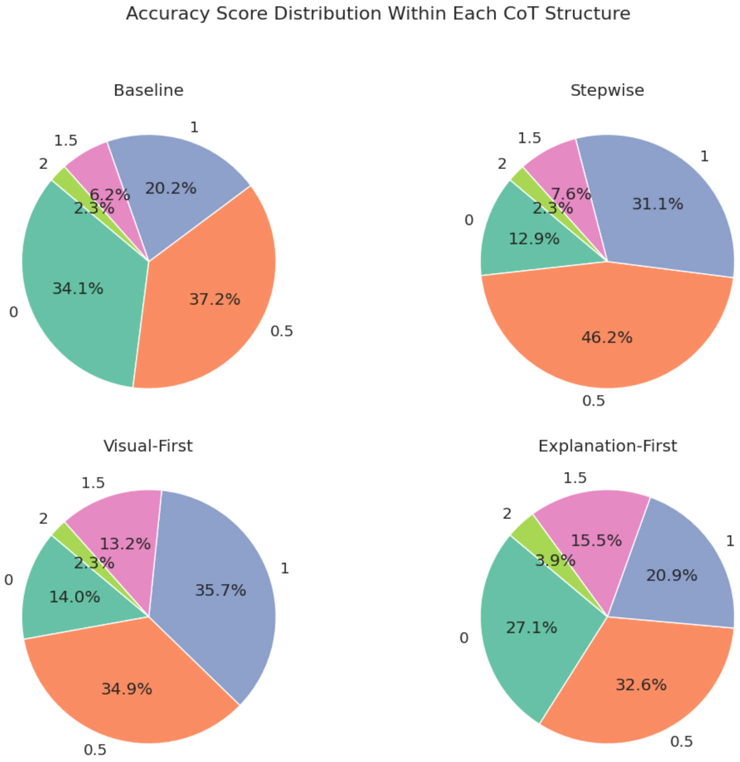

Figure A.2: Task 0 accuracy distribution by prompt type.

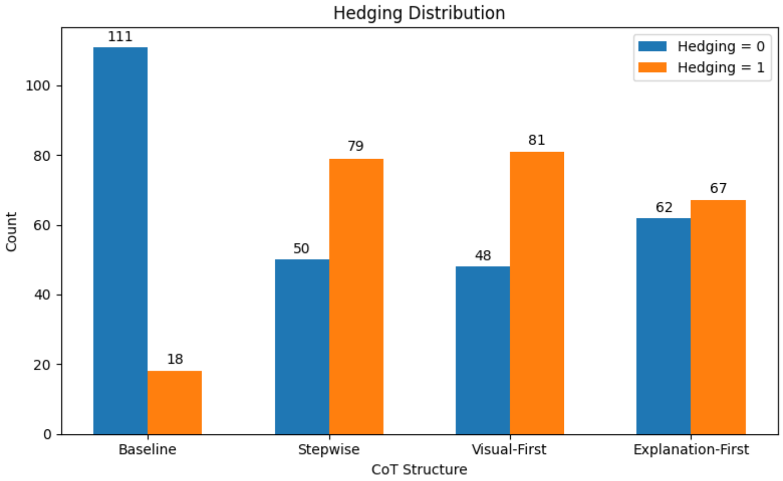

Figure A.3: Task 0 hedging behavior distribution.

Task 1: EAS Reactivity

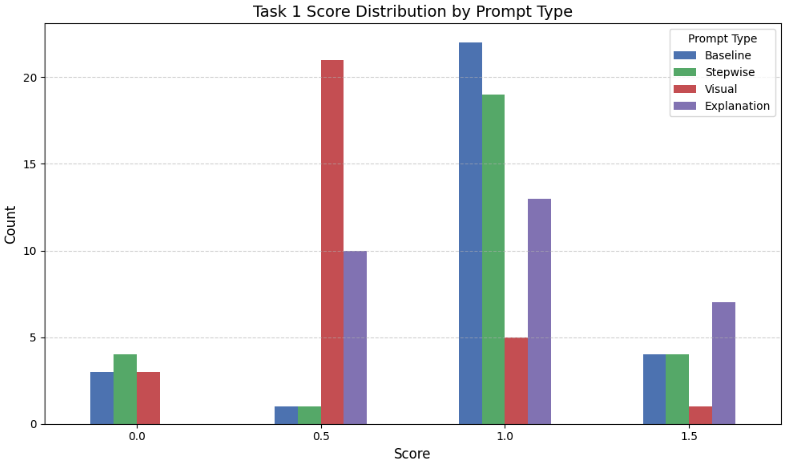

Figure A.4: Task 1 accuracy distribution by prompt type.

Task 2: Acid/Base Strength Comparison

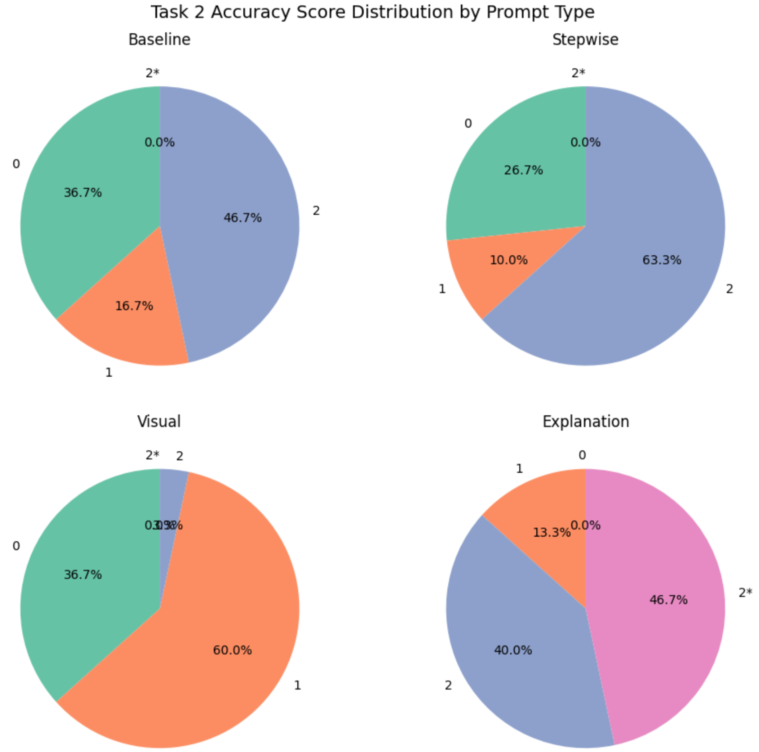

Figure A.5: Task 2 accuracy distribution by prompt type.

Task 3: Nucleophilicity (Functional Group Reactivity)

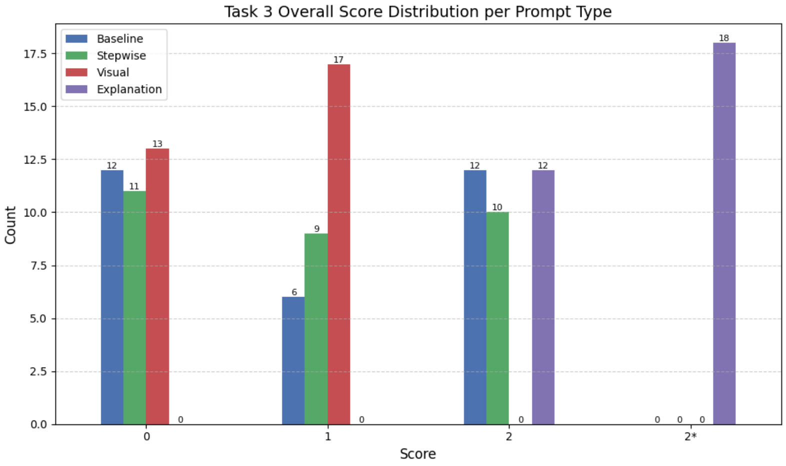

Figure A.6: Task 3 accuracy distribution by prompt type.

Task 4: SN1 Reaction Likelihood

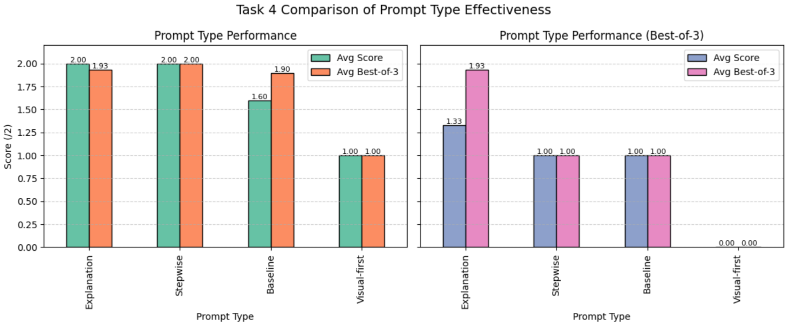

Figure A.7: Task 4 performance by prompt type.