

---

# Chain-of-Thoughts for Molecular Understanding

---

Yunhui Jang<sup>1</sup>, Jaehyung Kim<sup>2</sup>, Sungsoo Ahn<sup>1</sup>

<sup>1</sup>Pohang University of Science and Technology    <sup>2</sup>Yonsei University  
{uni5510, sungsoo.ahn}@postech.ac.kr,  
jaehyungk@yonsei.ac.kr

## Abstract

Large language models (LLMs) have achieved remarkable success in natural language processing tasks, and their specialized adaptations of chemistry have shown promising performance in molecular understanding tasks such as generating a text description of a molecule. However, reasoning based on structural information of molecules remains a significant challenge, e.g., even advanced LLMs such as GPT-4o struggle to identify functional groups which are crucial for inferring the molecular property of interest. To address this limitation, we propose STRUCTCOT, a structure-aware chain-of-thought (CoT) that enhances LLMs’ understanding of molecular structures by explicitly injecting the key structural features of molecules. Moreover, we introduce two fine-tuning frameworks for adapting the existing LLMs to use our STRUCTCOT. Our experiments demonstrate that STRUCTCOT leads to significant improvements and achieves state-of-the-art performance in both molecular understanding tasks.

## 1 Introduction

Large language models (LLMs; Touvron et al., 2023; OpenAI & et al., 2024; Raffel et al., 2020) have demonstrated remarkable performance across various tasks. To leverage their strong capabilities in chemistry, several prior works (Edwards et al., 2022; Christofidellis et al., 2023a; Fang et al., 2024; Pei et al., 2023) have proposed chemical LLMs that have shown superior performance in molecular understanding tasks such as molecule captioning (Mol2Text) and text-based molecule generation (Text2Mol) (Edwards et al., 2022), which are crucial for designing new molecules.

Reasoning based on molecular structures plays an important role in molecular understanding tasks in practice. For example, chemists consider a molecule potentially toxic if it contains a phenol group due to the formation of phenoxyl radicals and the compound’s ability to interact with biological membranes (Hansch et al., 2000). However, despite its significance, there exists a lack of studies that have studied reasoning in molecular understanding. In various other domains such as arithmetic and commonsense reasoning, chain-of-thought (CoT; Wei et al., 2022; Kojima et al., 2022) has shown that explicitly incorporating such reasoning steps significantly improves the performance of LLMs.

In detail, CoT aims to generate intermediate reasoning steps before arriving at a final answer. One might consider the naive adaptation of CoT prompting to include molecular structural information in reasoning. However, this approach is ineffective because even state-of-the-art LLMs (OpenAI & et al., 2024; Touvron et al., 2023) struggle to capture the structural details of molecules, indicating their lack of knowledge about molecules, as described in Figure 1 and Section 3.2. This hinders their ability to perform reasoning effectively in molecular understanding tasks. While some prior works (Ouyang et al., 2024; Jin et al., 2024; M. Bran et al., 2024) have proposed CoTs for chemistry, they are either not applicable or exhibit limited performance in molecular understanding tasks.

In this paper, we propose STRUCTCOT, a chain-of-thought that progressively sketches the structural features of molecules to solve molecular understanding tasks. STRUCTCOT consists of six key

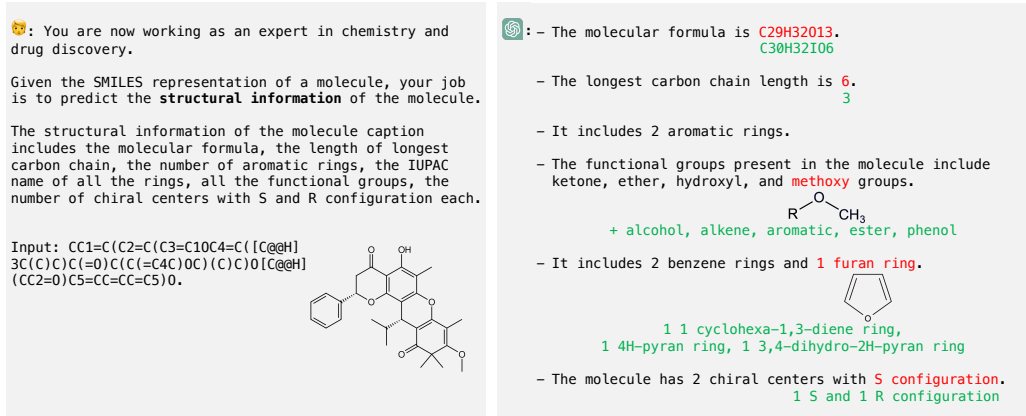


Figure 1: **The failure case of GPT-4o for the inference of structural information given the molecular SMILES.** The red color indicates the wrong generated structural information while the green color indicates the correct answer. Notably, the input for GPT-4o only includes the molecular SMILES, not the molecular graph.

structural elements, ranging from the primary structure to the smaller components. We propose to explicitly inject the appropriate structural information with STRUCTCoT to enhance the language models' understanding of molecules, which assists in inferring complex structural information.

Moreover, we propose two different fine-tuning frameworks to apply STRUCTCoT depending on the input and output of the given molecular understanding task, as illustrated in Figure 5. Both approaches share the same outline, including a reasoning module that generates STRUCTCoT and an answering module that generates the output from the input combined with STRUCTCoT. On the one hand, for the molecule captioning task, we use external tools like RDKit (Landrum et al., 2024) as the reasoning module, since they can exactly determine the structural information from the molecule. Therefore, one can simply attach a perfectly accurate STRUCTCoT to the input SMILES string and let the answering module generate the output.

On the other hand, for the text-based molecule generation task, one cannot acquire perfect STRUCTCoT as the molecule is not provided. Therefore, we propose to finetune the LLMs as the reasoning module that generates STRUCTCoT (Ho et al., 2023; Fu et al., 2023a; Magister et al., 2023). Then, we fine-tune the answering module to generate the answer given the text description and the acquired STRUCTCoT. Moreover, we incorporate a novel *matching-ratio-based rejection sampling* into the answering module, which forces the generated molecule to align with STRUCTCoT. Notably, the proposed rejection sampling leverages the deterministic nature of structural information for a given molecule.

In summary, our key contributions are as follows:

- We present the limitations of LLMs in understanding molecular structures by analyzing their capability to infer structural information.
- We introduce STRUCTCoT, a chain-of-thought that progressively sketches the structural information of molecules, making the generation process more intuitive.
- We propose to incorporate STRUCTCoT for molecule captioning by fine-tuning the answering module with the deterministic and perfectly accurate STRUCTCoT.
- We propose to incorporate STRUCTCoT for text-based molecule generation by applying CoT fine-tuning for the reasoning module, fine-tuning the answering module, and a novel matching ratio-based rejection sampling which further improves the performance.
- We validate the efficacy of STRUCTCoT and our fine-tuning framework by showing consistent improvements across chemistry-specific and general LLMs.

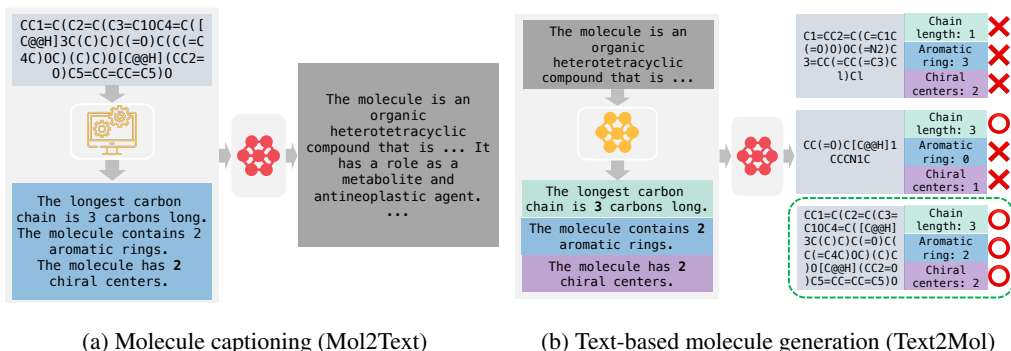


Figure 2: **Overview of the Fine-tuning Process of STRUCTCOT.** Light gray boxes represent SMILES strings; gray boxes represent text descriptions; colored boxes represent STRUCTCOT. The yellow ones are the reasoning module, and the red ones are the answering module. In (b), colors correspond to structural information elements. The third SMILES is selected after matching ratio-based rejection sampling for having the highest matching ratio (3/3).

## 2 Related work

**Large language models for chemistry.** General-purpose large language models (generalist LLMs) often struggle to solve basic chemistry problems and understand the molecules (White et al., 2023; Castro Nascimento & Pimentel, 2023; Guo et al., 2023). To address this issue, prior works have introduced specialist LLMs, i.e., chemical LLMs, by pre-training models on molecule-related texts (Edwards et al., 2022; Christofidellis et al., 2023b; Liu et al., 2023; Pei et al., 2023), through instruction tuning (Fang et al., 2024; Cao et al., 2023), and using retrieval-based in-context learning (Li et al., 2024). Our work focuses on reasoning processes that are broadly applicable to these specialist LLMs as well as generalist LLMs.

**Chain-of-thought reasoning.** Chain-of-thought (CoT) aims to generate intermediate reasoning steps before arriving at a final answer (Wei et al., 2022; Kojima et al., 2022). CoT not only enhances the reasoning capabilities of LLMs but also improves the overall quality of generated answers. Most prior works generated CoTs via few-shot learning based on the manually written CoTs (Wei et al., 2022) or by prompting LLMs with “Let’s think step by step.” (Kojima et al., 2022). In addition, several approaches have proposed to further enhance CoT, including techniques such as self-consistency (Wang et al., 2023), least-to-most prompting (Zhou et al., 2023), complexity-based prompting (Fu et al., 2023b), and self-polish (Xi et al., 2023).

**Chain-of-thought reasoning for chemistry.** Recently, a few works have extended CoT reasoning to address chemistry-related problems. For instance, Ouyang et al. (2024) proposed to employ the program-of-thoughts (PoT; Chen et al., 2023) to handle chemical question-answering tasks. Additionally, Jin et al. (2024) presented the protein chain of thought (ProCoT) to replicate the signaling pathways in the context of the protein-protein interaction (PPI) problem. Despite these advances, none of these works target molecular understanding tasks such as molecule captioning and text-based molecule generation. We note that some of the CoTs provided by M. Bran et al. (2024) can be applied for molecular understanding tasks, but their focus is on other tasks, e.g., organic synthesis using CoTs generated from tools like *LitSearch/WebSearch*, *PatentCheck*, *ReactionPlanner*, and *SMILES2Price*.<sup>1</sup>

## 3 Structure as milestones of LLM-based chemical reasoning

### 3.1 Examples of important structural information

Humans typically analyze a molecule by progressively mapping its structure, starting with primary elements like rings and long carbon chains, and then identifying smaller elements such as functional groups and chiral centers. Reflecting this approach, we identify six key elements of molecular structural information that are critical for chemical reasoning. To highlight the importance of these

<sup>1</sup>We experimented ChemCrow in molecular understanding tasks and ChemCrow shows limited performance.

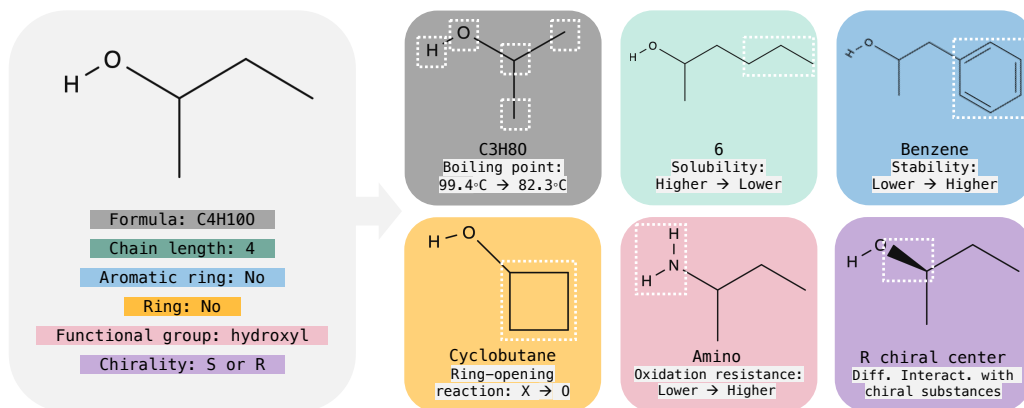


Figure 3: **Illustration of the Importance of Structural Information.** An example of replacing each structural information (described with a dashed box) of the molecule. From left to right and top to bottom, the described structural information are molecular formula, longest carbon chain length, aromatic ring, ring compounds, functional group, and chirality.

structural elements, we demonstrate how even slight modifications in molecular structure can lead to significant changes in chemical or physical properties, as shown in Figure 3.

**Molecular formula.** The molecular formula provides essential information about a molecule's composition, specifying the number and type of atoms present. This information is critical because it directly determines the molecular weight, which influences a molecule's physical properties. For instance, although 2-Butanol (C<sub>4</sub>H<sub>10</sub>O) and 2-Propanol (C<sub>3</sub>H<sub>8</sub>O) are composed of the same elements, i.e., carbon, hydrogen, and oxygen, their differing molecular formulas result in distinct molecular weights (74.1g/mol for 2-Butanol and 60.1g/mol for 2-Propanol). These differences lead to the change in boiling points, 99.4°C and 82.3°C, respectively, as shown in the gray part of Figure 3.

**Longest carbon chain.** The longest carbon chain (excluding atoms in ring systems) forms the molecular backbone where functional groups are attached. The length of this chain significantly influences properties like solubility. For example, extending the carbon chain of 2-Butanol from four to six carbons creates 2-Hexanol, which exhibits reduced solubility. This is illustrated in the green section of Figure 3.

**Aromatic rings.** Aromatic rings, such as benzene or pyridine, play a critical role in determining the stability and electronic properties of molecules. For instance, adding a benzene ring to 2-Butanol yields 1-Phenyl-2-Propanol, which has enhanced stability and greater oxidation resistance. This transformation is shown in the blue section of Figure 3.

**Ring compounds.** Similar to the longest carbon chain, ring structures often serve as the backbone where functional groups are attached. The ring system significantly affects molecular behavior and reactions. For example, although 2-Butanol and Cyclobutanol share the same number of carbons and oxygen, the ring in Cyclobutanol introduces a tendency toward ring-opening reactions, as depicted in the yellow section of Figure 3.

**Functional groups.** Functional groups, e.g., hydroxyl, amino, ester, etc., play a pivotal role in determining the chemical reactivity of molecules. For example, alcohols with a hydroxyl group (-OH) are prone to oxidize more while the molecules with an amino group (-NH<sub>2</sub>) are generally resistant to oxidation under mild conditions. A single replacement of a hydroxyl (-OH) group in 2-Butanol with an amino (-NH<sub>2</sub>) group leads to 2-Butanamine, which has increased oxidation resistance, as described in the red part of Figure 3.

**Chiral centers.** Chirality refers to the stereochemical property of a molecule that makes it non-superimposable on its mirror image, leading to different chemical behaviors. The chirality is determined by the chiral centers and their configurations, i.e., R- and S-configuration<sup>2</sup>, which describe the spatial arrangement of the groups around the chiral centers. This leads to different interactions between other molecules with chirality. For instance, (R)-2-Butanol and (S)-2-Butanol may interact differently with other chiral substances. This is described in the purple part of Figure 3.

<sup>2</sup>The names of R and S come from the Latin word *Rectus* and *Sinister*, which means right and left, respectively.

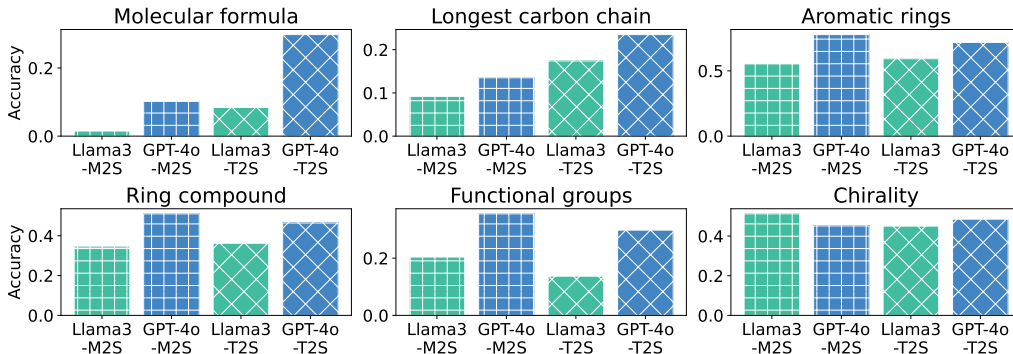


Figure 4: **Analysis of LLMs’ Understanding of Structural Information.** Colors indicate the architectures of the language models, with green and blue representing LLaMA3-8B and GPT-4, respectively. Patterns denote the input types: crossed patterns represent SMILES representations (*Molecule2Structure*), and diagonally crossed patterns represent molecule captions (*Text2Structure*).

### 3.2 Recent large language models do not understand structural information

Next, we demonstrate that even recent LLMs, i.e., GPT-4o (OpenAI & et al., 2024) and LLaMA3-8B-Instruct (Touvron et al., 2023), fail to infer important structural information from the given molecule and the text description of the molecule. We evaluated the LLMs by querying the structural information from the SMILES string (Weininger, 1988) and the text description, which was considered to be a simple task that could be solved by someone with a bachelor’s degree in chemistry.

As shown in Figure 4, both GPT-4o and LLaMA3-8B-Instruct failed to capture the structural information accurately. First, when the SMILES is given, both models performed best in counting the number of aromatic rings, with accuracies around 50% and 75%, respectively. However, their accuracies were significantly lower for other structural information. This implies that the molecular string does not provide sufficient information for LLMs to fully understand the molecular structures.

Similarly, when the text description is given, both models also fail to achieve a high accuracy in inferring the structural information. This indicates that LLMs cannot properly understand the structure of molecules even when provided with the text description of molecules. This points out the potential of the explicit use of a structural CoT to enhance molecular comprehension. Note that the detailed experimental settings and prompts for the analysis are provided in Appendix A.1 and the example failure case is demonstrated in Figure 1.

## 4 STRUCTCOT: structure-aware CoTs for molecules

### 4.1 STRUCTCOT

We introduce STRUCTCOT, a structure-aware CoT designed to enhance language models’ understanding of molecular structures. Each component of STRUCTCOT is illustrated in Figure 5 and further detailed in the following.

**Molecular formula** is expressed as “*The molecular formula is  $X_1N_1 \cdots X_MN_M$ .*”, where  $X_m$  and  $N_m$  represent the  $m$ -th atom type and the associated number of atoms, respectively.

**Length of the longest carbon chain** takes the following form: “*The longest carbon chain length is  $N$ .*”, where  $N$  denotes the length of the longest carbon chain of the molecule.

**Number of aromatic rings** takes the following form: “*It includes  $X$  aromatic ring(s).*”, where  $X$  denotes the number of aromatic rings in the molecule.

**Types of ring compounds** is expressed as “*It includes  $N_1 X_1$  rings,  $\cdots$ ,  $N_M X_M$  ring(s).*”, where  $X_m, N_m$  represents the International Union of Pure Applied Chemistry (IUPAC) name of the ring compound and the number of the rings, respectively.

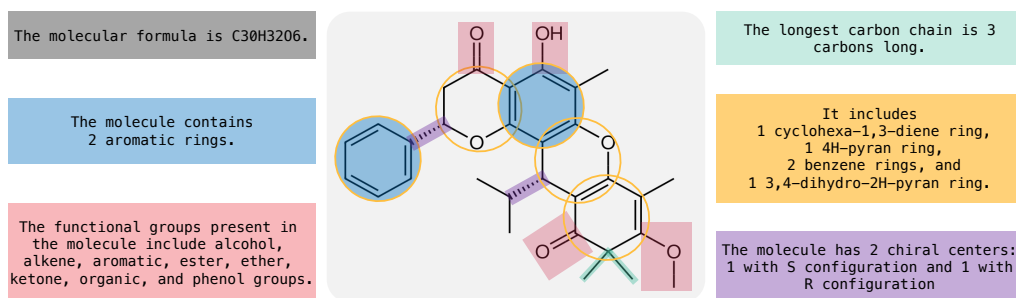


Figure 5: **The Six New Elements of STRUCTCOT: molecular formula, longest carbon chain length, aromatic rings, ring compounds, functional groups, and chirality.** The same color indicates the CoT and the corresponding structural information of the molecule. The order of the STRUCTCOT follows the order mentioned in the title of the figure, which progressively sketches the structure of molecules.

**Types of functional groups** is expressed as “*The functional groups of the molecule are  $X_1, X_2, \dots$ , and  $X_N$  group.*”, where  $X_n$  denotes the name of the functional group.<sup>3</sup>

**Number and types of chiral centers** is formulated as follows: “*The molecule has  $N$  chiral centers:  $N_S$  with  $S$  configuration and  $N_R$  with  $R$  configuration.*”, where  $N = N_S + N_R$ , and  $N_S$  and  $N_R$  denotes the number of chiral centers of  $S$  and  $R$  configurations, respectively.

## 4.2 Molecule captioning

Molecule captioning aims to generate an accurate and detailed text description of a given molecular SMILES string. We incorporate our STRUCTCOT scheme through (1) using the RDKit (Landrum et al., 2024) as a ground-truth reasoning module and (2) fine-tuning the answering module LLM with the generated CoT as an additional input. We provide the description in Figure 2a.

**Reasoning module.** One can obtain the true structural information of the given molecule from RDKit, which allows us to guide the answering module without uncertainty. This is natural as the structural information is deterministic given the molecule. Consequently, the obtained true structural information is used as STRUCTCOT. For this task, we consider the molecular weight CoT and IUPAC name CoTs (M. Bran et al., 2024) in addition to the CoTs described in Section 4.1.

**Answering module.** With the molecule and the acquired CoT as an input, we fine-tune the LLMs to generate the description of the molecule. In the experiments, we mainly consider chemical LLMs, i.e., MolT5 (Edwards et al., 2022) and ChemT5 (Christofidellis et al., 2023a), as the answering module.

## 4.3 Text-based molecule generation

Text-based molecule generation is the reverse process of molecule captioning, intending to generate the corresponding molecular string based on the given description. Following the two-stage framework that separates rationale generation and answer inference (Zhang et al., 2024), we first generate STRUCTCOT with the fine-tuned reasoning module and then attach this to the input and employ this as an input of the answering module. We provide the description in Figure 2b.

Notably, we do selectively use the CoT elements in STRUCTCOT. This is because the CoTs need to guarantee sufficient quality for the answering module to take advantage of the information from the generated CoTs. Therefore, we evaluate the abilities of the reasoning module to correctly generate the CoTs and exclude the ones with low accuracy (presented in Table 2): molecular formula CoT and the two CoTs proposed by M. Bran et al. (2024).

**Reasoning module.** For the reasoning module, following Ho et al. (2023); Fu et al. (2023a); Magister et al. (2023), we enable CoT reasoning of the models by fine-tuning the reasoning module on the STRUCTCOT as the molecule is not given. This is in contrast to the molecule captioning task where the exact structural information can be extracted from external tools with the given molecule. We mainly fine-tune the chemical LLMs, i.e., MolT5 and ChemT5 for this task.

<sup>3</sup>Note that we consider a wider range of functional groups compared to that of M. Bran et al. (2024).

**Answering module.** For the answering module, similar to that of molecule captioning, we fine-tune a chemical LLM to generate an appropriate molecule given the text description and generated STRUCTCOT. Moreover, we propose the *matching ratio-based rejection sampling*, which forces the generated molecule to align with STRUCTCOT, as described in the following.

The proposed matching ratio-based rejection sampling aims to match the structural information of the generated molecule with the given STRUCTCOT. In detail, we generate multiple  $k$  molecules using beam search and then score each molecule based on the matching ratio, which counts the number of matching structural information elements between STRUCTCOT and the generated molecule. Finally, we choose the best-scoring molecule as the final output. This approach also leverages the deterministic nature of structural information, i.e., we can easily compare the alignment between each structural information and the generated molecule. Notably, this differs from the prior works with iterative approaches (Wang et al., 2023; Xi et al., 2023; Sun et al., 2024), as we focus on the alignment between CoT and generated answer without needing to generate multiple rationales.

## 5 Experiments

In this section, we present our experiments on molecule captioning and text-based molecule generation tasks, including the experimental results, setting details, and ablation studies. We first explain the common settings shared by both tasks.

**Dataset.** Following prior works (Edwards et al., 2022; Christofidellis et al., 2023a), we employ the widely used CHEBI-20 dataset (Edwards et al., 2021), which consists of 33,010 pairs of molecular SMILES and their text descriptions. We also use the same train/validation/test split of 80%/10%/10%.

**Baselines.** We verify the performance enhancement of STRUCTCOT in two settings: specialist and generalist models. On one hand, we employed two popular specialist models, i.e., chemical LLMs: MolT5 (Edwards et al., 2022) and Text+CHem T5 (ChemT5; Christofidellis et al., 2023a). To validate the efficacy of our method across various model sizes, we used small (77M) and base (252M) for ChemT5 and base and large (800M) for MolT5. On the other hand, we employed two recent large language models: Llama3-8B-Instruct (Touvron et al., 2023) and GPT-4o (OpenAI & et al., 2024)<sup>4</sup> as our generalist models. Additionally, we include five baselines including RNN (Jain & Medsker, 1999), Transformer (Vaswani et al., 2017), T5 (Raffel et al., 2020), MolXPT (Liu et al., 2023), and BioT5 (Pei et al., 2023) to compare the absolute performance.

### 5.1 Molecule captioning

**Experimental setup and metrics.** For specialist models, we follow the method proposed in Section 4.2, i.e., to fine-tune the language model given the SMILES and ground-truth STRUCTCOT. For generalists, we cannot guarantee that the generated descriptions align with those in our training data. Therefore, we apply few-shot learning by attaching CoTs in the same way as for the specialist models. Performance is evaluated by comparing the generated captions with the ground-truth captions using six metrics: BLEU-2, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. We provide detailed experimental settings and prompts in Appendix A.2.

**Results.** We report the experimental results in Table 1. One can observe that adding our STRUCTCOT consistently improves performance for both specialist and generalist models. Surprisingly, despite BioT5 being pre-trained on a significantly larger dataset and sharing the same model size, our method, when incorporated with ChemT5-base, achieves competitive results without the need for any additional pre-training data. Moreover, MolT5-large incorporated with STRUCTCOT showed state-of-the-art performance. We provide examples of generated samples in Appendix B.1.

### 5.2 Text-based molecule generation

**Experimental setup and metrics.** For specialist models, we follow the method proposed in Section 4.3. For generalist models, we also apply few-shot learning by giving examples with structural information, to guide the language model to generate the structural information and then the answer molecule. The performance is evaluated using eight metrics: SMILES comparison metrics (BLEU,

<sup>4</sup>We used gpt-4o-2024-05-13.

Table 1: **Molecule Captioning Performance.**  $\Delta$  denotes the performance difference between the original model and the one incorporated with STRUCTCoT. **Teal** color indicates the improvement.

Models	BLEU-2		BLEU-4		ROUGE-1		ROUGE-2		ROUGE-L		METEOR	
	Metric	$\Delta$	Metric	$\Delta$	Metric	$\Delta$	Metric	$\Delta$	Metric	$\Delta$	Metric	$\Delta$
<i>Baselines (without CoTs)</i>												
RNN	0.251	-	0.176	-	0.450	-	0.278	-	0.394	-	0.363	-
T5-base	0.511	-	0.423	-	0.607	-	0.451	-	0.550	-	0.539	-
Transformer	0.061	-	0.027	-	0.204	-	0.087	-	0.186	-	0.114	-
MolXPT	0.594	-	0.505	-	0.660	-	0.511	-	0.597	-	0.626	-
BioT5	0.635	-	0.556	-	0.692	-	0.559	-	0.633	-	0.656	-
<i>Specialists (fine-tuning)</i>												
MolT5-base	0.540	-	0.457	-	0.634	-	0.485	-	0.578	-	0.569	-
+STRUCTCoT	0.592	0.052	0.507	0.050	0.667	0.043	0.523	0.038	0.606	0.028	0.619	0.050
MolT5-large	0.594	-	0.508	-	0.654	-	0.510	-	0.594	-	0.614	-
+STRUCTCoT	0.645	0.051	0.567	0.059	0.699	0.045	0.568	0.058	0.639	0.045	0.666	0.052
ChemT5-small	0.553	-	0.462	-	0.633	-	0.481	-	0.574	-	0.583	-
+STRUCTCoT	0.601	0.048	0.513	0.050	0.664	0.031	0.519	0.038	0.603	0.029	0.624	0.042
ChemT5-base	0.580	-	0.490	-	0.647	-	0.498	-	0.586	-	0.604	-
+STRUCTCoT	0.639	0.059	0.560	0.070	0.687	0.040	0.553	0.055	0.626	0.040	0.657	0.053
<i>Generalists (10-shot learning)</i>												
Llama3	0.211	-	0.117	-	0.367	-	0.183	-	0.308	-	0.257	-
+STRUCTCoT	0.259	0.048	0.158	0.041	0.401	0.034	0.208	0.025	0.324	0.016	0.341	0.084
GPT-4o	0.232	-	0.128	-	0.389	-	0.183	-	0.307	-	0.291	-
+STRUCTCoT	0.286	0.054	0.174	0.046	0.405	0.016	0.199	0.016	0.313	0.006	0.341	0.050

Table 2: **Reasoning Accuracy for Each Structural Information.**

Models	Form.	Chain	Arom.	Ring	Func.	Chiral.	Weight	Name
<i>Specialists (fine-tuning)</i>								
MolT5-base	0.458	0.922	0.926	0.930	0.957	0.798	0.606	0.512
MolT5-large	0.450	0.899	0.905	0.906	0.945	0.766	0.597	0.465
ChemT5-small	0.447	0.920	0.930	0.926	0.954	0.788	0.634	0.495
ChemT5-base	0.475	0.925	0.931	0.930	0.960	0.799	0.641	0.525
<i>Generalists</i>								
Llama3	0.084	0.174	0.593	0.362	0.137	0.450	0.435	0.015
GPT-4o	0.298	0.235	0.718	0.464	0.298	0.485	0.728	0.040

Exact, and Levenshtein distance), fingerprint similarity metrics (MACCS FTS, RDK FTS, and Morgan FTS), a molecular distribution metric (Frechet ChemNet Distance (FCD)), and the validity of the generated molecule. We provide detailed experimental settings and prompts in [Appendix A](#).

**Reasoning accuracy.** We measure the reasoning accuracy to filter out low-accuracy reasoning components that may misguide the answer module. The reasoning accuracies are provided in [Table 2](#). Our results show that fine-tuned specialist reasoners exhibit superior reasoning accuracy compared to the larger generalist models. However, even our reasoning modules failed to achieve high accuracy in molecular formula, molecular weight, and IUPAC name. Therefore, we filter out these three.

**Results.** The experimental results are reported in [Table 3](#). Incorporating our generated STRUCTCoT to the molecular description always improved performance. In particular, incorporating STRUCTCoT into the ChemT5-base achieves state-of-the-art performance compared to the recent baselines, validating the efficacy of our CoTs. Interestingly, our STRUCTCoT even improves the performance of smaller models beyond that of the vanilla larger models, e.g., MolT5-base+STRUCTCoT showed superior performance to MolT5-large. We provide examples of generated samples in [Appendix B.2](#).

Notably, the generalist models incorporated with STRUCTCoT do not show enhancement, due to their low reasoning accuracy described in [Section 3.2](#). This supports our assertion that correct reasoning improves performance, while incorrect reasoning may even degrade it.



Table 3: **Text-based Molecule Generation Performance.** The teal color indicates the improvement while the red color indicates the reduction. Note that it is natural to show no enhancement in generalist models as they lack reasoning ability as shown in Table 2.

Models	BLEU		Exact		Levenshtein ↓		MACCS FTS		RDK FTS		Morgan FTS		FCD ↓		Validity	
	Met.	Δ	Met.	Δ	Met.	Δ	Met.	Δ	Met.	Δ	Met.	Δ	Met.	Δ	Met.	Δ
<i>Baselines (without CoTs)</i>																
RNN	0.652	-	0.005	-	38.09	-	0.591	-	0.400	-	0.362	-	4.55	-	0.542	-
Transformer	0.499	-	0.000	-	57.66	-	0.480	-	0.320	-	0.217	-	11.32	-	0.906	-
T5-base	0.762	-	0.069	-	24.95	-	0.731	-	0.605	-	0.545	-	2.48	-	0.660	-
MolXPT	-	-	0.215	-	-	-	0.859	-	0.757	-	0.667	-	0.45	-	0.983	-
BioT5	0.867	-	0.413	-	15.10	-	0.886	-	0.801	-	0.734	-	0.43	-	<b>1.000</b>	-
<i>Specialists (fine-tuning)</i>																
MolT5-base	0.769	-	0.081	-	24.46	-	0.721	-	0.588	-	0.529	-	2.18	-	0.772	-
+STRUCTCoT	0.863	0.094	0.385	0.304	13.91	10.55	0.918	0.197	0.843	0.255	0.783	0.254	0.29	1.89	0.983	0.211
MolT5-large	0.854	-	0.311	-	16.07	-	0.834	-	0.746	-	0.684	-	1.20	-	0.905	-
+STRUCTCoT	0.840	0.014	0.364	0.053	15.91	0.16	0.902	0.068	0.820	0.074	0.759	0.075	0.40	0.80	0.985	0.080
ChemT5-small	0.739	-	0.157	-	28.54	-	0.859	-	0.736	-	0.660	-	0.07	-	0.776	-
+STRUCTCoT	0.874	0.135	0.381	0.224	13.22	15.32	0.918	0.059	0.845	0.109	0.787	0.127	0.29	0.22	0.976	0.200
ChemT5-base	0.750	-	0.212	-	27.39	-	0.874	-	0.767	-	0.697	-	0.06	-	0.792	-
+STRUCTCoT	<b>0.878</b>	0.128	<b>0.421</b>	0.209	<b>12.76</b>	14.63	<b>0.924</b>	0.050	<b>0.856</b>	0.089	<b>0.804</b>	0.107	0.26	0.20	0.982	0.190
<i>Generalists (10-shot learning)</i>																
Llama3	0.251	-	0.007	-	117.30	-	0.586	-	0.352	-	0.276	-	13.11	-	0.629	-
+STRUCTCoT	0.259	0.008	0.008	0.001	109.77	7.53	0.579	0.007	0.279	0.073	0.344	0.068	4.47	8.64	0.669	0.040
GPT-4o	0.521	-	0.079	-	40.87	-	0.797	-	0.496	-	0.583	-	3.67	-	0.881	-
+STRUCTCoT	0.509	0.012	0.088	0.009	41.68	0.081	0.783	0.014	0.498	0.002	0.571	0.012	1.57	2.10	0.846	0.035

Table 4: **Comparison to ChemCrow.** The left part shows the molecule captioning results and the right part shows the text-based molecule generation results.

	Mol2Text			Text2Mol					
	BLEU2	ROUGE2	METEOR	BLEU	Leven. ↓	MACCS	Morgan	FCD ↓	Valid.
GPT-4o	0.232	0.183	0.291	0.521	40.87	0.797	0.583	3.67	0.881
ChemCrow (GPT-4o)	0.162	0.097	0.225	0.306	56.46	0.772	0.555	2.31	0.851
Ours (GPT-4o)	0.286	0.199	0.341	0.509	41.68	0.783	0.571	1.57	0.846
Ours (MolT5-large)	<b>0.645</b>	<b>0.568</b>	<b>0.666</b>	<b>0.840</b>	<b>15.91</b>	<b>0.902</b>	<b>0.759</b>	<b>0.40</b>	<b>0.985</b>

### 5.3 Ablation study

**Comparison to ChemCrow.** To validate the efficacy of our STRUCTCoT, we compare our method with ChemCrow (M. Bran et al., 2024), which has employed CoTs for various chemical tasks. The comparative results are provided in Table 4. We select some representative metrics (e.g., ROUGE2 among ROUGE1, 2, and 4) due to the space limit and the remaining metrics are provided in Appendix B.3. One can observe that ChemCrow shows limited performance in both molecule understanding tasks. We provide the experimental details in Appendix A.4.

#### Matching ratio-based rejection sampling.

Here, we discuss the efficacy of matching ratio-based rejection sampling in the answer module of text-based molecule generation. We compare the results without and with matching ratio-based rejection sampling ( $k = 5$ ) for ChemT5-small under the same setting including the hyperparameters. As demonstrated in Figure 6, the matching ratio-based rejection sampling improves performance by encouraging the generated molecule to follow the given STRUCTCoT.

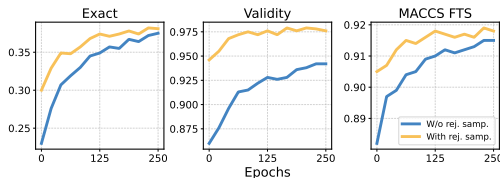


Figure 6: **Efficacy of Matching Ratio-based Rejection Sampling.**

## 6 Conclusion

In this paper, we introduced STRUCTCoT, a structure-aware chain-of-thought framework that enhances language models’ understanding of molecular structures by explicitly incorporating key structural features. By fine-tuning domain-specific specialist models with STRUCTCoT, we achieved consistent improvements in molecule captioning and text-based molecule generation tasks. This work underscores the effectiveness of small, domain-specific models in capturing molecular structures and offers a solution for molecular reasoning.

## Acknowledgments

This work was partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. IITP-2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2022R1C1C1013366), Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MSIT) (No. 2022R1A6A1A03052954), and the National Research Foundation of Korea(NRF) grant funded by the Ministry of Education(MSIT) (No. RS-2024-00436165).

## References

- He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery, 2023. 3
- Cayque Monteiro Castro Nascimento and AndréSilva Pimentel. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6): 1649–1655, 03 2023. doi: 10.1021/acs.jcim.3c00285. URL <https://doi.org/10.1021/acs.jcim.3c00285>. 3
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=YfZ4ZPt8zd>. 3
- Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6140–6157. PMLR, 23–29 Jul 2023a. URL <https://proceedings.mlr.press/v202/christofidellis23a.html>. 1, 6, 7
- Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6140–6157. PMLR, 23–29 Jul 2023b. URL <https://proceedings.mlr.press/v202/christofidellis23a.html>. 3
- Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2Mol: Cross-modal molecule retrieval with natural language queries. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.47. URL <https://aclanthology.org/2021.emnlp-main.47>. 7
- Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.26. URL <https://aclanthology.org/2022.emnlp-main.26>. 1, 3, 6, 7
- Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=Tl5dsb619n>. 1, 3
- Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International*

- Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10421–10430. PMLR, 23–29 Jul 2023a. URL <https://proceedings.mlr.press/v202/fu23d.html>. 2, 6
- Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023b. URL <https://openreview.net/forum?id=yflicZHC-19>. 3
- Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023. 3
- Corwin Hansch, Susan Mckarns, Carr Smith, and David Doolittle. Comparative qsar evidence for a free-radical mechanism of phenol-induced toxicity. *Chemico-Biological Interactions*, 127:61–72, 07 2000. doi: 10.1016/S0009-2797(00)00171-X. 1
- Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.830. URL <https://aclanthology.org/2023.acl-long.830>. 2, 6
- L. C. Jain and L. R. Medsker. *Recurrent Neural Networks: Design and Applications*. CRC Press, Inc., USA, 1st edition, 1999. ISBN 0849371813. 7
- Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. ProLLM: Protein chain-of-thoughts enhanced LLM for protein-protein interaction prediction. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=2nTzomzjjb>. 1, 3
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=e2TBb5y0yFf>. 1, 3
- Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, NadineSchneider, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Vincent F. Scalfani, Rachel Walker, Daniel Probst, Kazuya Ujihara, Axel Pahl, guillaume godin, Juuso Lehtivarjo, tadhurst cdd, François Bérenger, and Jonathan Bisson. rdkit/rdkit: 2024\_09\_1 (q3 2024) release beta, September 2024. URL <https://doi.org/10.5281/zenodo.13820100>. 2, 6
- Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–13, 2024. ISSN 2326-3865. doi: 10.1109/tkde.2024.3393356. URL <http://dx.doi.org/10.1109/TKDE.2024.3393356>. 3, 15, 16
- Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. MolXPT: Wrapping molecules with text for generative pre-training. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1606–1616, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.138. URL <https://aclanthology.org/2023.acl-short.138>. 3, 7
- Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8. URL <https://doi.org/10.1038/s42256-024-00832-8>. 1, 3, 6, 9, 17

- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.151. URL <https://aclanthology.org/2023.acl-short.151>. 2, 6
- OpenAI and Josh Achiam et al. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>. 1, 5, 7, 14
- Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Yejin Choi, Jiawei Han, and Lianhui Qin. Structured chemistry reasoning with large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=7R3pzxTS1g>. 1, 3
- Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1102–1123, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.70. URL <https://aclanthology.org/2023.emnlp-main.70>. 1, 3, 7
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>. 1, 7
- Jiashuo Sun, Yi Luo, Yeyun Gong, Chen Lin, Yelong Shen, Jian Guo, and Nan Duan. Enhancing chain-of-thoughts prompting with iterative bootstrapping in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4074–4101, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.257. URL <https://aclanthology.org/2024.findings-naacl.257>. 7
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>. 1, 5, 7, 14
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf). 7
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>. 3, 7
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf). 1, 3
- David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 02 1988. 5

- Andrew D. White, Glen M. Hocky, Heta A. Gandhi, Mehrad Ansari, Sam Cox, Geemi P. Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, and Willmor J. Peña Ccoa. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2:368–376, 2023. doi: 10.1039/D2DD00087C. URL <http://dx.doi.org/10.1039/D2DD00087C>. 3
- Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Jia Liu, Tao Gui, Qi Zhang, and Xuanjing Huang. Self-Polish: Enhance reasoning in large language models via problem refinement. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11383–11406, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.762. URL <https://aclanthology.org/2023.findings-emnlp.762>. 3, 7
- Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=y1pPWFVfvR>. 6
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=WZH7099tgfM>. 3

## A Experimental details

In this section, we provide the details of the experiments. All experimental code related to this paper is available at <https://anonymous.4open.science/r/MolStructCoT>.

### A.1 Structure information analysis

Here, we describe the detailed settings for the analysis in Section 3.1. To evaluate the understanding of two recent LLMs: Llama3-8B-Instruct (Touvron et al., 2023) and GPT-4o (OpenAI & et al., 2024), we prompt the LLMs to infer the structural information from the given molecular SMILES string and text description of the molecule.

**Prompts given SMILES string.** First, we asked LLMs to infer the structural information from the SMILES string, with the prompt described in Table 6.

Table 5: **Prompts for structure information analysis given SMILES string.**

---

**Head prompt:** You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule, your job is to predict the structural information of the molecule.

The structural information of the molecule caption includes the molecular formula, the length of the longest carbon chain, the number of aromatic rings, the IUPAC name of all the rings, all the functional groups, the number of chiral centers with S and R configurations each, the molecular weight, the IUPAC name of the molecule.

The functional group and ring IUPAC names should be on the list. The number of chiral centers should also be format {"S": , "R": }.

Your response should only be in the JSON format following {"molecular formula": , "functional group": , "longest carbon chain length": , "aromatic ring": , "ring IUPAC name":, "chiral": {"S": , "R": }, "weight": , "IUPAC name": }.

THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE. DO NOT CHANGE THE JSON KEY NAMES.

**Input prompt:** Input: <SMILES>

---

**Prompts given text description of molecules.** Next, we asked LLMs to infer the structural information from the text description of the molecule, with the prompt described in Table 5.

Table 6: **Prompts for structure information analysis given text description.**

---

**Head prompt:** You are now working as an excellent expert in chemistry and drug discovery. Given the caption of a molecule, your job is to predict the structural information of the molecule. The molecule caption is a sentence that describes the molecule, which mainly describes the molecule’s structures, properties, and production.

The structural information of the molecule caption includes the molecular formula, the length of the longest carbon chain, the number of aromatic rings, the IUPAC name of all the rings, all the functional groups, the number of chiral centers with S and R configurations each, the molecular weight, the IUPAC name of the molecule.

The functional group and ring IUPAC names should be on the list. The number of chiral centers should also be format {"S": , "R": }.

Your response should only be in the JSON format following {"molecular formula": , "functional group": , "longest carbon chain length": , "aromatic ring": , "ring IUPAC name":, "chiral": {"S": , "R": }, "weight": , "IUPAC name": }.

THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE. DO NOT CHANGE THE JSON KEY NAMES.

**Input prompt:** Input: <Description>

---

## A.2 Molecule captioning

Here, we describe the detailed settings for the experiments of molecule captioning in Section 5.1. Note that we used four A100-80GB GPUs for fine-tuning.

**Hyperparameters.** The hyperparameters for the specialist models are provided in Table 7. Note that MolT5-large was not trained for the same number of epochs as the other models due to limited time constraints.

Table 7: **Hyperparameters for molecule captioning.**

Hyperparameter	MolT5-base	MolT5-large	ChemT5-small	ChemT5-base
Batch size	8	4	8	8
Learning rate	$2e^{-4}$	$2e^{-4}$	$6e^{-4}$	$6e^{-4}$
Epochs	250	160	250	250
Warmup ratio	0	0	0.1	0.1
Weight decay	0.01	0.01	0	0
Lr scheduler	linear	linear	linear	linear

**Prompts.** The prompts used for the generalist models are described in Table 11. We primarily followed the prompt presented by Li et al. (2024).

Table 8: **Prompts for generalist models in text-based molecule generation task.**

**Head prompt:** You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule and structural description of the molecule, your job is to predict the caption of the molecule. The molecule caption is a sentence that describes the molecule, which mainly describes the molecule’s structures, properties, and production.

Example 1:

Instruction: Given the SMILES representation of a molecule, predict the caption of the molecule.

Input: <SMILES><STRUCTCOT >

Your output should be: {"caption": <Description>}

...

Example  $k$ :

Instruction: Given the SMILES representation of a molecule, predict the caption of the molecule.

Input: <SMILES><STRUCTCOT >

Your output should be: {"caption": <Description>}

Your response should only be in the JSON format above; THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR RESPONSE.

**Input prompt:** Input: <SMILES><STRUCTCOT >

### A.3 Text-based molecule generation

Here, we described the detailed settings for the experiments of text-based molecule generation in Section 4.3. Note that we also used four A100-80GB GPUs for fine-tuning.

**Hyperparameters.** The hyperparameters for the reasoning and answering module for the specialist models are provided in Table 9 and Table 10, respectively. Note that MolT5-large was not trained for the same number of epochs as the other models due to limited time constraints.

Table 9: ~~Hyperparameters for the reasoning module of text-based molecule generation.~~

Hyperparameter	MolT5-base	MolT5-large	ChemT5-small	ChemT5-base
Batch size	8	4	8	8
Learning rate	$1e^{-3}$	$1e^{-3}$	$6e^{-4}$	$6e^{-4}$
Epochs	250	120	250	250
Warmup ratio	0.1	0.1	0	0
Weight decay	0	0	0	0
Lr scheduler	cosine	cosine	linear	linear

Table 10: ~~Hyperparameters for the reasoning module of text-based molecule generation.~~

Hyperparameter	MolT5-base	MolT5-large	ChemT5-small	ChemT5-base
Batch size	8	4	8	8
Learning rate	$1e^{-3}$	$1e^{-3}$	$6e^{-4}$	$6e^{-4}$
Epochs	250	200	250	250
Warmup ratio	0.1	0.1	0	0
Weight decay	0	0	0	0
Lr scheduler	cosine	cosine	linear	linear

**Prompts.** The prompts used for the generalist models are described in Table 8. We also primarily followed the prompt presented by Li et al. (2024).

Table 11: ~~Prompts for the generalist models in molecule captioning task.~~

**Head prompt:** You are now working as an excellent expert in chemistry and drug discovery. Given the caption of a molecule, your job is to predict the SMILES representation of the molecule. The molecule caption is a sentence that describes the molecule, which mainly describes the molecule’s structures, properties, and production.

You can infer the molecule SMILES representation from the caption.

Before you infer the molecule SMILES representation, YOU SHOULD FIRST GENERATE the molecular formula, the length of the longest carbon chain, the number of aromatic rings, the IUPAC name of all the rings, all the functional groups, the number of chiral centers with S and R configurations each, the molecular weight, the IUPAC name of the molecule.

Example 1: Instruction: Given the caption of a molecule, predict the SMILES representation of the molecule.

Input: <Description><STRUCTCOT >

Your output should be: {"molecule": <SMILES>}

...

Example  $k$ : Instruction: Given the caption of a molecule, predict the SMILES representation of the molecule.

Input: <Description><STRUCTCOT >

Your output should be: {"molecule": <SMILES>}

You should FIRST generate the structural information following the examples above, and then provide the JSON format of the molecule SMILES based on that.

NOTE THAT THE SMILES REPRESENTATION MUST BE IN THE JSON format above {"molecule": }. THERE SHOULD BE NO OTHER CONTENT INCLUDED IN YOUR JSON. DO NOT CHANGE THE JSON KEY NAME.

**Input prompt:** Input: <Description>



#### A.4 Ablation study

Here, we describe the detailed settings for the ablation study.

**Prompts for ChemCrow.** The prompts used for ChemCrow (M. Bran et al., 2024) are described in Table 12 and Table 13. Notably, it was not able to apply few-shot learning for ChemCrow as it was not applicable as the original prompt proposed in ChemCrow does not include any few-shot setting.

---

Table 12: **Prompts for molecule captioning with ChemCrow.**

---

**Head prompt:** Given the SMILES representation of a molecule and structural description of the molecule, your job is to predict the caption of the molecule.

"Final Answer" follows the format: Final Answer: {"caption": }

**Input prompt:** The SMILES representation of the molecule is as follows: : <SMILES>

---

---

Table 13: **Prompts for text-based molecule generation with ChemCrow.**

---

**Head prompt:** Given the caption of a molecule, your job is to predict the SMILES representation of the molecule.

The molecule caption is a sentence that describes the molecule, which mainly describes the molecule's structures, properties, and production.

You can infer the molecule SMILES representation from the caption.

"Final Answer" follows the format: Final Answer: {"molecule": }

**Input prompt:** The caption is as follows: <Description>

---

## B Additional experimental results

In this section, we provide additional experimental results including several concrete examples of generated samples.

### B.1 Molecule captioning

Here, we show the samples of molecule captioning, i.e., generated text descriptions of given molecules in Figure 7. Notably, we show the generated samples from base-sized models for fair comparison.

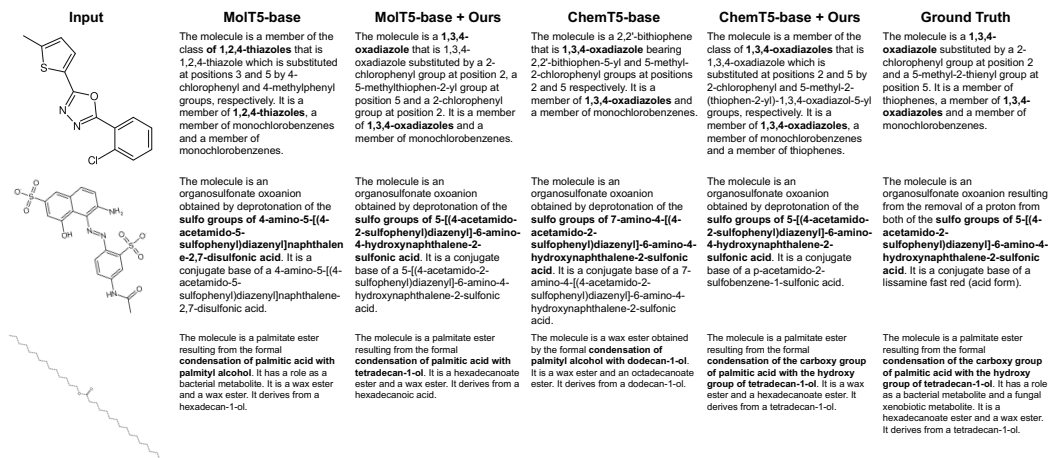


Figure 7: The generated samples of molecule captioning.

## B.2 Text-based molecule generation

Here, we show the samples of text-based molecule generation, i.e., generated molecules for the given text description in Figure 8. Notably, we show the generated samples from base-sized models for fair comparison.

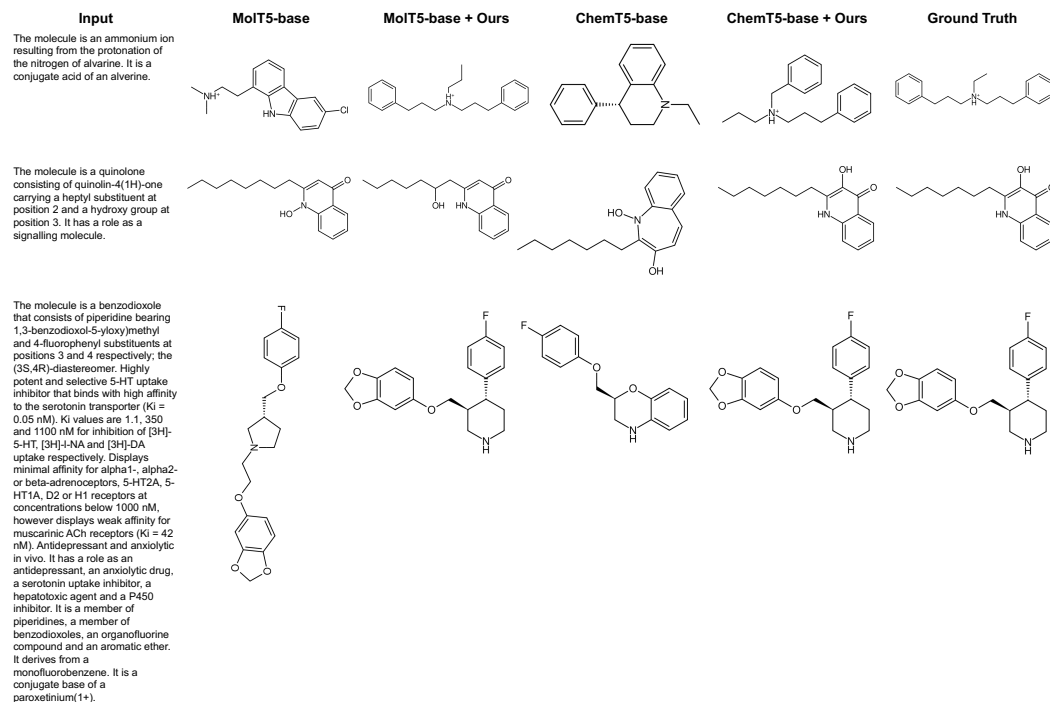


Figure 8: The generated samples of text-based molecule generation.

## B.3 Ablation study

Here, we provide all the metrics in ChemCrow comparison ablation study. The molecule captioning results are in Table 14 and the text-based molecule generation results are in Table 15.

Table 14: **Comparison to ChemCrow in molecule captioning.** The specialist model indicates our results from MolT5-large while the generalist model indicates the one from GPT-4o.

Models	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
ChemCrow (GPT-4o)	0.162	0.078	0.299	0.097	0.211	0.225
Ours (GPT-4o)	0.249	0.139	0.386	0.179	0.300	0.303
Ours (MolT5-large)	0.645	0.567	0.699	0.568	0.639	0.666

Table 15: **Comparison to ChemCrow in text-based molecule generation.** The specialist model indicates our results from GPT-4o while the generalist model indicates the one from .

Models	BLEU	Exact	Levenshtein ↓	MACCS FTS	RDK FTS	Morgan FTS	FCD ↓	Validity
ChemCrow (GPT-4o)	0.306	0.194	56.46	0.772	0.632	0.555	2.31	0.851
Ours (GPT-4o)	0.509	0.088	41.68	0.783	0.498	0.571	1.57	0.846
Ours (MolT5-large)	0.840	0.364	15.91	0.902	0.820	0.759	0.40	0.985