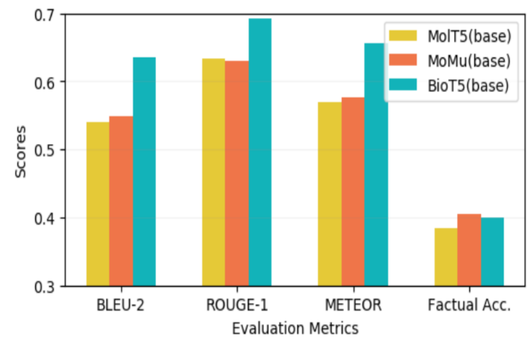
MoleculeQA: A Dataset to Evaluate Factual Accuracy in Molecular Comprehension

Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, Yu Li
Shenzhen International Graduate School, Tsinghua University
International Digital Economy Academy (IDEA) Hong Kong University of Science and Technology

Introduction and Motivation

- Molecular LLMs exhibit a high degree of hallucination in generated content.
- Existing benchmarks are insufficient for measuring the level of hallucination.
- We propose MoleculeQA to verify the accuracy of existing LLMs on molecular knowledge:
  - A taxonomy focused on molecular knowledge.
  - A high-quality QA benchmark based on the taxonomy.
  - Comprehensive evaluation of existing molecular LLMs.

CID	Ground-Truth	Generated
9810996	The molecule is a dipeptide composed of N-(3,3-dimethylbutyl)-L-aspartic acid and methyl L-phenylalanate units joined by a peptide linkage.	The molecule is a dipeptide obtained by formal condensation of the alpha-carboxy group of N-(3,3-dimethylbutyl)-L-phenylalanine with ethanol.
10129879	The molecule is the stable isotope of potassium with relative atomic mass 38.963707.	The molecule is the stable isotope of tellurium with relative atomic mass 124.904425.



MoleculeQA Dataset

Taxonomy Construction

Process Molecular Corpus

- Original Corpus:
  - *ChEBI-20* dataset +*T3DB, FDA, DrugBank*
- Topic Extraction:
  - Rule-based Program + Few-shot prompt

Normalize & Structuralize

- *Topic Filtering: 1000 -> 587 topics*
- Cluster + Merge: Hierarchical 3-layer Taxonomy
- Multi-round verification by human experts.

Dataset Construction

Extract Content & Reassign Topic

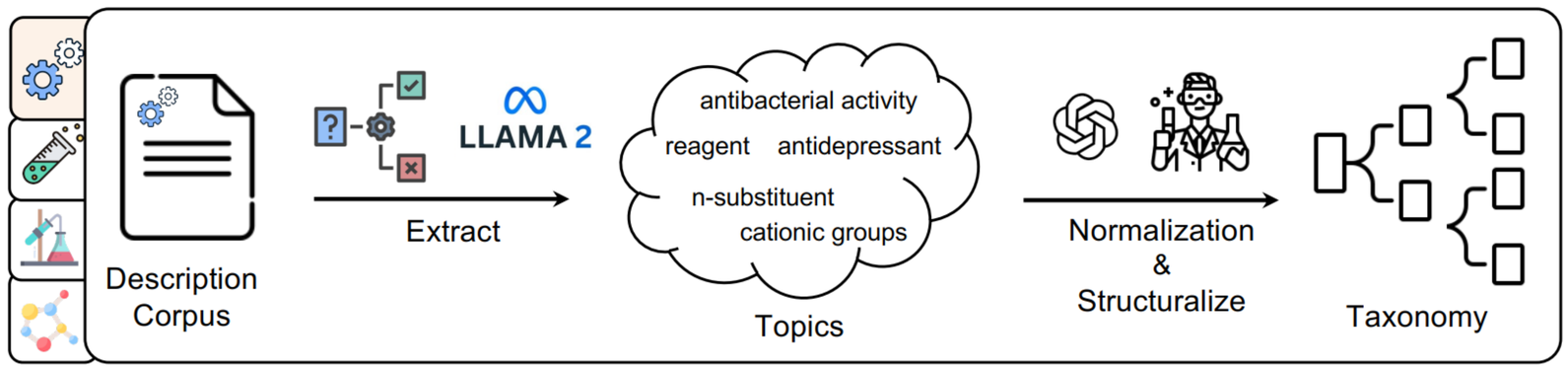
- Former extracted topic can be out of taxonomy
- Former extracted topic-content pair can not be directly queried

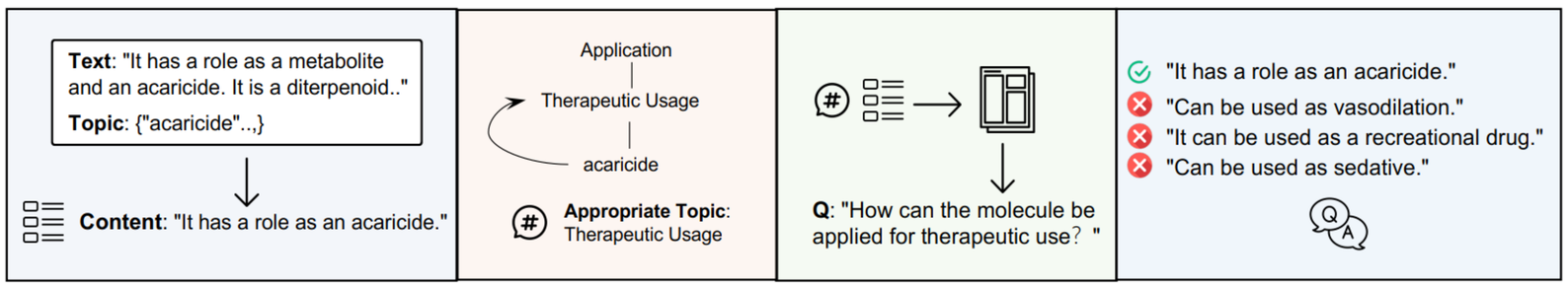
Question-Answer pair Construction

- Design question acceding to topic
- Extract right answer from original description
- Select negative candidates from other mols by topic
- Multi-round verification by LLMs and annotators

MoleculeQA

- Largest existing QA benchmark for molecule LLMs
- Data distribution among topics and categories
- Strict manual verification to ensure quality





(a) Extract Content (b) Reassign topic (c) Design question (d) Collect Pos. & Neg. answers

Taxonomy	Reference Description	Extracted Question	Positive Answer	Negative Answer
Property→ Antiviral activity	It has been shown to *exhibit inhibitory effects on the viral neuraminidases from two influenza viral strains, H1N1 and H9N2*.	Which kind of **antiviral activity** does this molecule have/exhibit?	It exhibits inhibitory effects on the viral neuraminidases from two influenza viral strains, H1N1 and H9N2.	It is used for the treatment of cytomegalovirus (CMV) retinitis in AIDS patients.
Structure→ Backbone	The molecule is a heparan sulfate composed of a backbone of *repeating beta-D-glucuronosyl-(1->4)- N-sulfonyl-alpha-D-glucosamine units joined by (1->4)-linkages*.	Which kind of **backbone** does this molecule have?	It has a backbone of repeating beta-D-glucuronosyl-(1->4)-N-sulfonyl-alpha-D-glucosamine units joined by (1->4)-linkages	It has a backbone of repeating alpha-L-iduronosyl-(1->4)-N-sulfonyl-alpha-D-glucosamine units joined by (1->4)-linkages.

Benchmarks	# QA	Sophistication
MMLU(Chem)	534	College, High school, Medicine
MMMU(Chem)	638	Inorganic, Organic, Physical
ScienceQA	867	Solution, Reaction, Molecule
ChemistryQA	4,500	Reaction, Molecule, Physics
MoleculeQA	61,574	Structure, Source, Property, Application

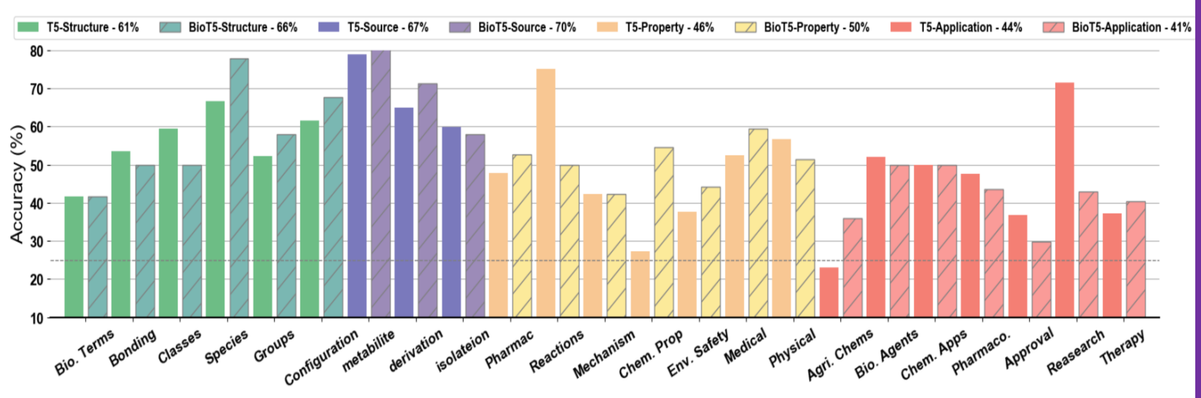
Aspects	Structure	Property	Application	Source	Total
# Train	32,176	4,838	1,917	11,062	49,993
# Dev	3,314	698	558	1,225	5,795
# Test	3,113	731	599	1,343	5,786
Avg. Q Tokens	7.96	9.02	7.90	7.00	7.74
Avg. A Tokens	9.50	10.98	11.93	7.96	9.42

Metric	Annotator 1	Annotator 2	Agreement (κ)
Consistency	99.0	99.0	1.0
Discrimination	97.0	96.0	0.85

Experimental Result

Model	# Trainable Params	Implementation	Structure	Source	Property	Application	Total
Random	–	–	24.41	22.30	23.04	24.57	24.03
Molecular LLM							
MolT5-small	80M	full ft	49.59	64.18	46.51	40.90	51.69
MolT5-base	250M	full ft	58.01	65.85	45.14	42.24	55.39
MoMu-small	82M	full ft	52.71	63.44	44.87	40.57	52.96
MoMu-base	252M	full ft	61.58	65.30	43.78	43.07	57.43
BioT5-base	252M	full ft	65.98	69.24	**49.11**	40.73	62.03
MolCA-125M	100M	LoRA ft	65.54	67.34	45.77	40.33	60.30
MolCA-1.3B	110M	LoRA ft	**71.12**	**70.98**	47.81	**43.17**	**64.79**
BioMedGPT-LM-7B	40M	LoRA ft	54.19	60.01	38.85	40.90	52.23
General LLM							
T5-small	60M	full ft	55.51	64.41	45.42	38.56	54.55
T5-base	220M	full ft	**60.42**	**66.42**	45.83	**43.74**	**58.24**
OPT-125M	125M	full ft	38.58	55.92	41.04	28.73	42.93
OPT-350M	331M	full ft	44.39	60.83	**46.24**	40.57	48.05
GALACTICA-6.7B	12.5M	LoRA ft	32.35	41.92	31.05	28.21	33.96
BLOOM-7.1B	27.5M	LoRA ft	35.01	47.51	31.46	33.56	37.31
Pythia-6.9B	29.4M	LoRA ft	42.79	58.90	38.58	39.07	45.61
Mol-Instruction-7B	40M	LoRA ft	37.46	47.36	32.69	29.88	38.37
Llama-2-7B-chat	40M	LoRA ft	28.75	39.84	31.33	27.71	31.54
Llama-2-13B-chat	63M	LoRA ft	34.37	43.86	31.05	29.72	35.67
Vicuna-v1.5-7B	40M	LoRA ft	34.89	44.15	34.20	31.55	36.61
Vicuna-v1.5-13B	63M	LoRA ft	37.01	43.19	30.64	31.55	37.07
Large-scale Universal Models							
Mixtral-8×7B-Instruct-v0.1	–	10-shot	23.32	31.87	32.89	29.96	27.79
GPT-3.5-1106-turbo	–	10-shot	25.60	37.60	28.04	32.22	29.29
GPT-4-1106-preview	–	10-shot	**60.94**	**50.19**	**35.57**	**43.91**	**53.47**

Factual accuracy of existing molecular LLMs is far from satisfactory



Accuracies among different topics reflects where model excels and fails

Model	Structure	Property	Application	Source
MolT5-base	63/0/34	1/4/3	7/15/8	20/10/30
MolT5-base-DPO	59/0/38	0/2/6	10/13/7	17/ 8/35
BioT5-base	62/0/35	2/3/3	9/12/9	16/13/31
BioT5-base-DPO	57/1/39	1/2/5	11/10/9	14/14/33

Apart from evaluation, MoleculeQA can be applied to reduce hallucination