ICLR

Mol-Instructions: A Large-Scale Biomolecular Instruction Dataset for Large Language Models

https://openreview.net/forum?id=Tlsdsb6l9n

Yin Fang*, Xiaozhuan Liang*, Ningyu Zhang† ✉, Kangwei Liu,

Rui Huang, Zhuo Chen, Xiaohui Fan, Huajun Chen† ✉

浙江大学 ZHEJIANG UNIVERSITY 1897

CONTENT

SEEKING TRUTH
PURSUING INNOVATION

01 Introduction & Background

02 Data Construction & Analysis

03 Experiments

04 Conclusion & Future Work

CONTENT

SEEKING TRUTH
PURSUING INNOVATION

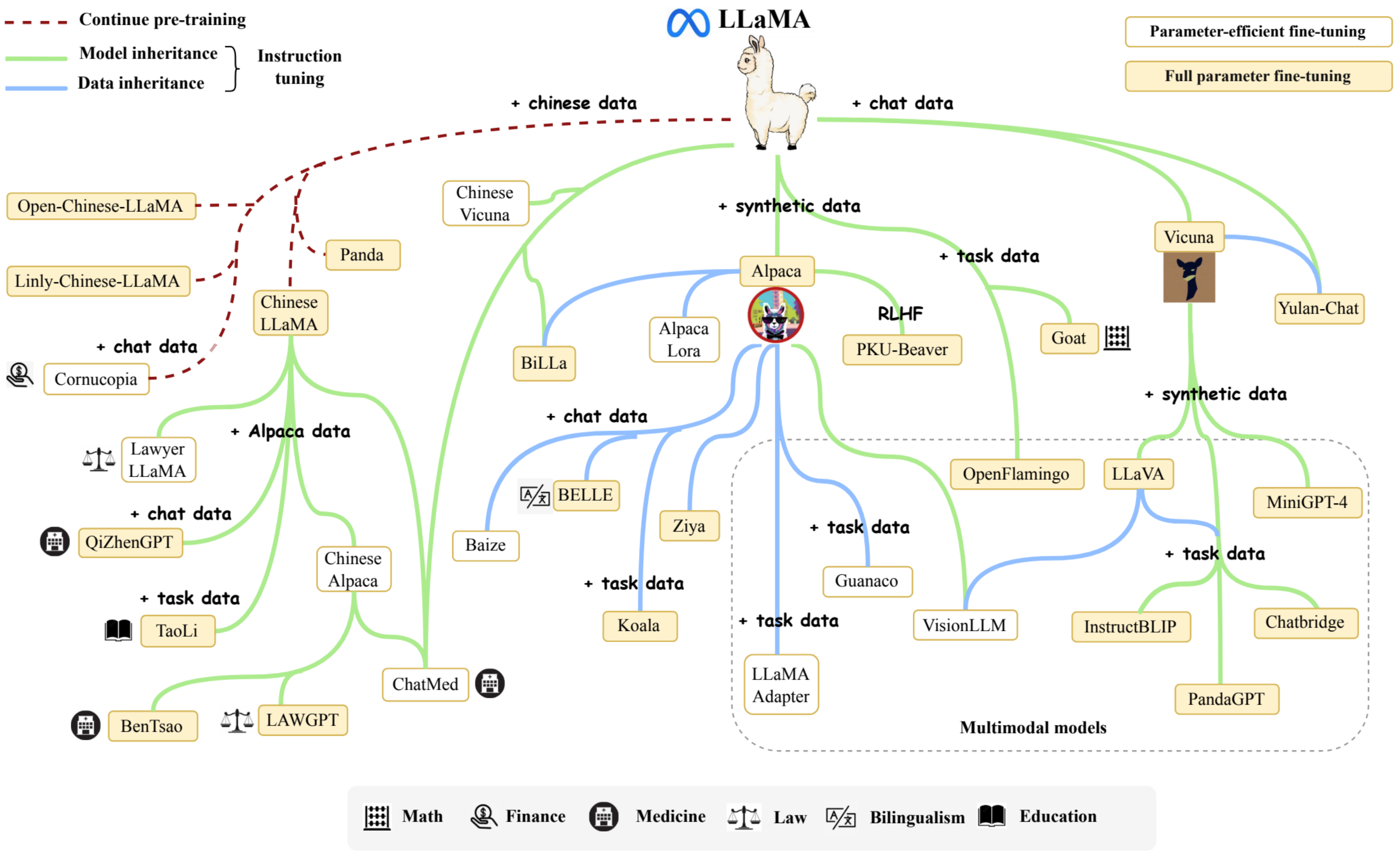
01 Introduction & Background

02 Data Construction & Analysis

03 Experiments

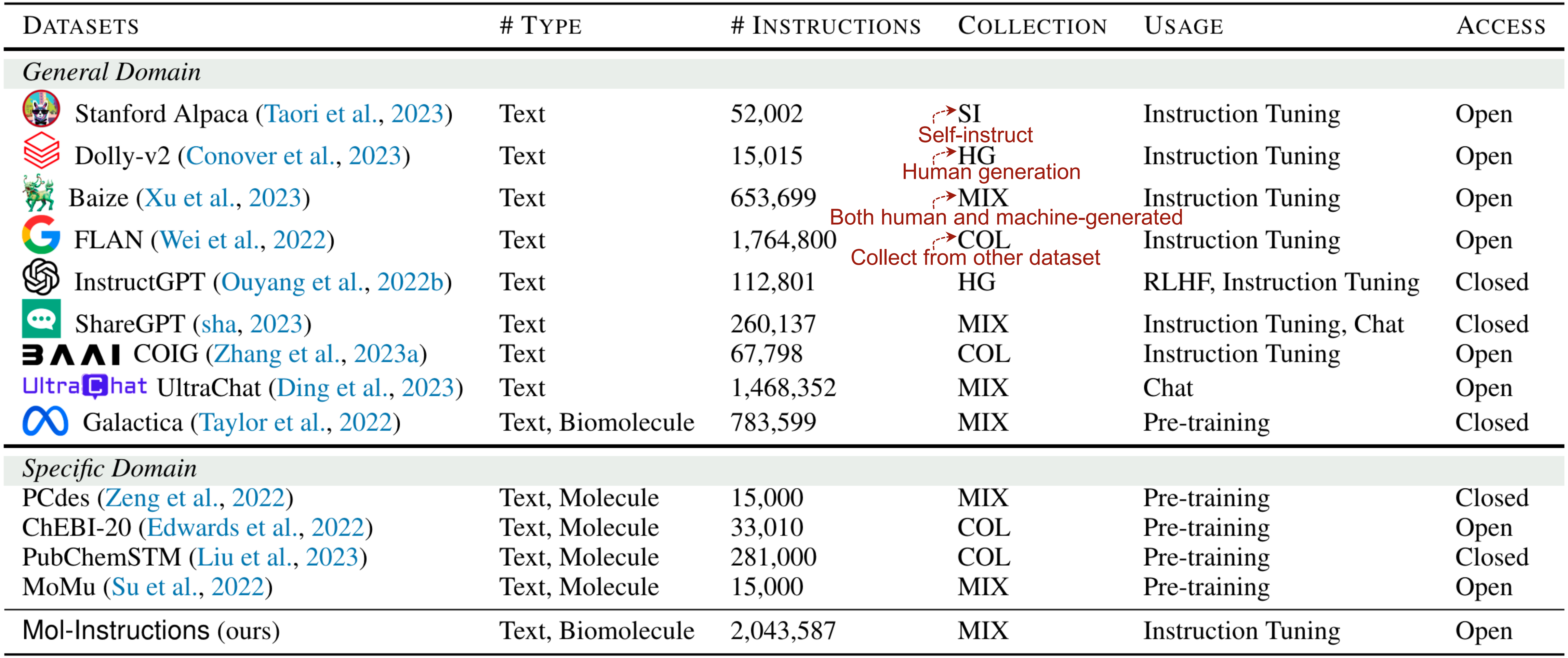
04 Conclusion & Future Work

Adapt LLMs for **Specific Domains**



A Survey of Large Language Models (2023)

Existing Instruction Datasets



Datasets	# Type	# Instructions	Collection	Usage	Access
General Domain					
Stanford Alpaca (Taori et al., 2023)	Text	52,002	SI	Instruction Tuning	Open
Dolly-v2 (Conover et al., 2023)	Text	15,015	HG	Instruction Tuning	Open
Baize (Xu et al., 2023)	Text	653,699	MIX	Instruction Tuning	Open
FLAN (Wei et al., 2022)	Text	1,764,800	COL	Instruction Tuning	Open
InstructGPT (Ouyang et al., 2022b)	Text	112,801	HG	RLHF, Instruction Tuning	Closed
ShareGPT (sha, 2023)	Text	260,137	MIX	Instruction Tuning, Chat	Closed
COIG (Zhang et al., 2023a)	Text	67,798	COL	Instruction Tuning	Open
UltraChat (Ding et al., 2023)	Text	1,468,352	MIX	Chat	Open
Galactica (Taylor et al., 2022)	Text, Biomolecule	783,599	MIX	Pre-training	Closed
Specific Domain					
PCdes (Zeng et al., 2022)	Text, Molecule	15,000	MIX	Pre-training	Closed
ChEBI-20 (Edwards et al., 2022)	Text, Molecule	33,010	COL	Pre-training	Open
PubChemSTM (Liu et al., 2023)	Text, Molecule	281,000	COL	Pre-training	Closed
MoMu (Su et al., 2022)	Text, Molecule	15,000	MIX	Pre-training	Open
Mol-Instructions (ours)	Text, Biomolecule	2,043,587	MIX	Instruction Tuning	Open

Self-instruct

Human generation

Both human and machine-generated

Collect from other dataset

CONTENT

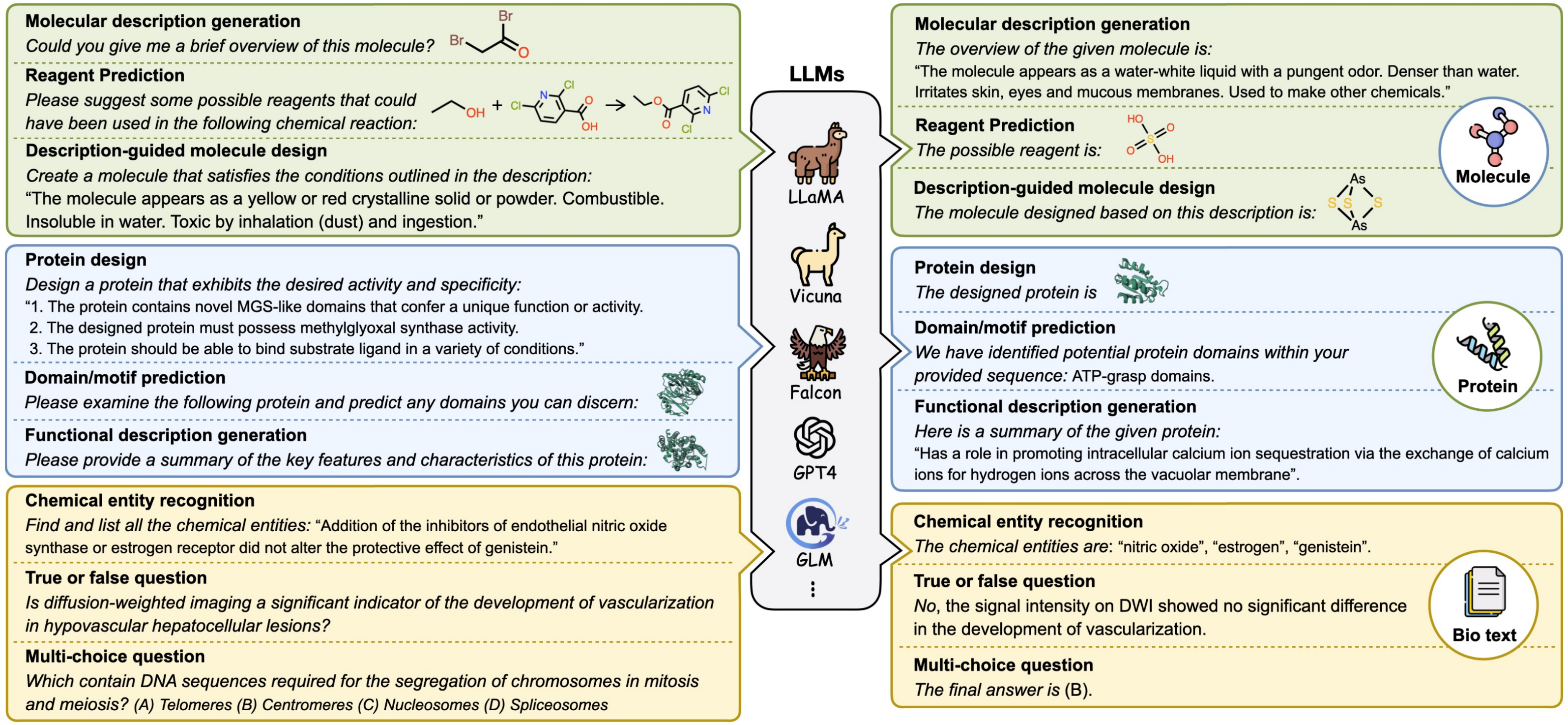
01 Introduction & Background

02 Data Construction & Analysis

03 Experiments

04 Conclusion & Future Work

Empowering LLMs with Mol-Instructions to unlock Biomolecular Domain



Molecular description generation
Could you give me a brief overview of this molecule?

Reagent Prediction
Please suggest some possible reagents that could have been used in the following chemical reaction:

Description-guided molecule design
Create a molecule that satisfies the conditions outlined in the description:
"The molecule appears as a yellow or red crystalline solid or powder. Combustible. Insoluble in water. Toxic by inhalation (dust) and ingestion."

Protein design
Design a protein that exhibits the desired activity and specificity:
"1. The protein contains novel MGS-like domains that confer a unique function or activity.
2. The designed protein must possess methylglyoxal synthase activity.
3. The protein should be able to bind substrate ligand in a variety of conditions."

Domain/motif prediction
Please examine the following protein and predict any domains you can discern:

Functional description generation
Please provide a summary of the key features and characteristics of this protein:

Chemical entity recognition
Find and list all the chemical entities: "Addition of the inhibitors of endothelial nitric oxide synthase or estrogen receptor did not alter the protective effect of genistein."

True or false question
Is diffusion-weighted imaging a significant indicator of the development of vascularization in hypovascular hepatocellular lesions?

Multi-choice question
Which contain DNA sequences required for the segregation of chromosomes in mitosis and meiosis? (A) Telomeres (B) Centromeres (C) Nucleosomes (D) Spliceosomes

LLMs

LLaMA
Vicuna
Falcon
GPT4
GLM
⋮

Molecular description generation
The overview of the given molecule is:
"The molecule appears as a water-white liquid with a pungent odor. Denser than water. Irritates skin, eyes and mucous membranes. Used to make other chemicals."

Reagent Prediction
The possible reagent is:

Description-guided molecule design
The molecule designed based on this description is:

Molecule

Protein design
The designed protein is

Domain/motif prediction
We have identified potential protein domains within your provided sequence: ATP-grasp domains.

Functional description generation
Here is a summary of the given protein:
"Has a role in promoting intracellular calcium ion sequestration via the exchange of calcium ions for hydrogen ions across the vacuolar membrane".

Protein

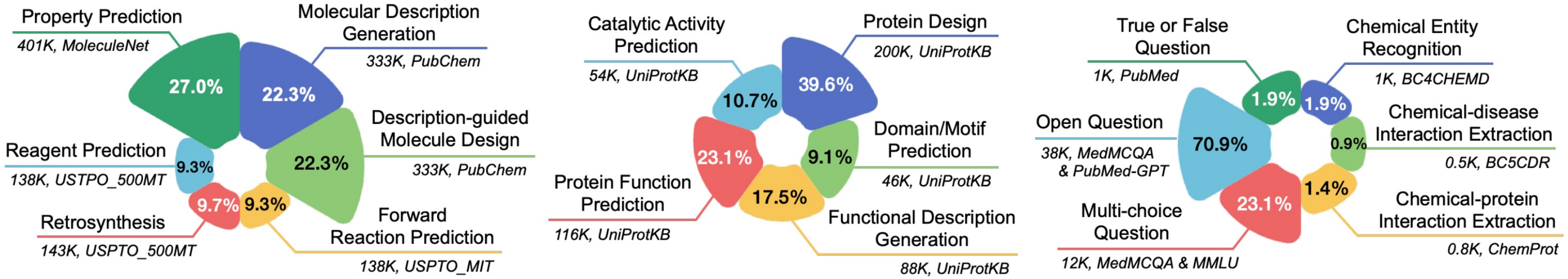
Chemical entity recognition
The chemical entities are: "nitric oxide", "estrogen", "genistein".

True or false question
No, the signal intensity on DWI showed no significant difference in the development of vascularization.

Multi-choice question
The final answer is (B).

Bio text

💡 Leveraging the powerful of LLMs to understand or even design biomolecules?



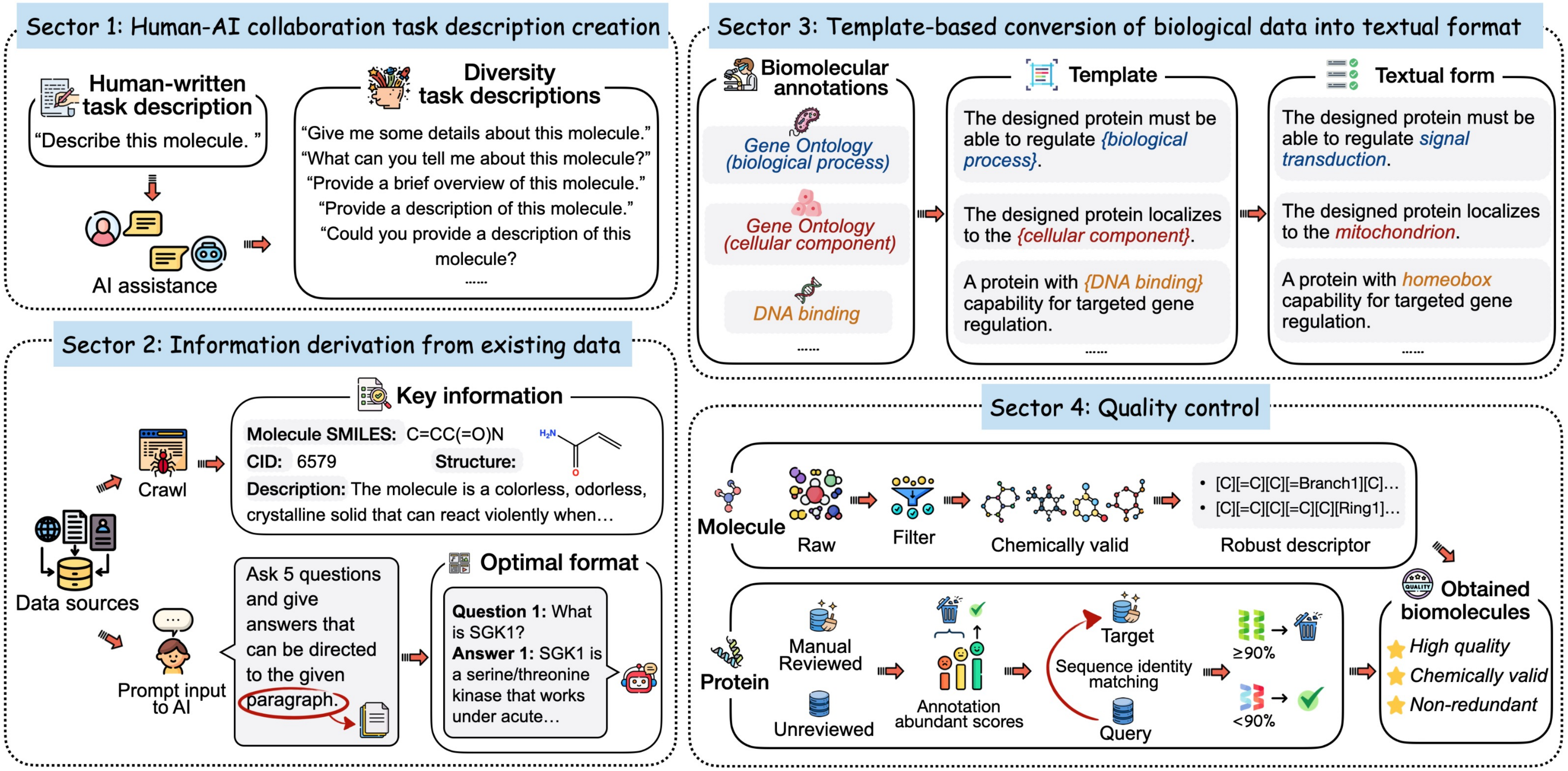
(a) Molecule-oriented Instructions (b) Protein-oriented Instructions (c) Biomolecular Text Instructions

Mol-Instructions includes tasks in three major categories, totaling 2,043,587 instruction data entries.

Data Constructions

Simulating the diversity of human needs and queries.

Converting structured annotations to text with templates.



Sector 1: Human-AI collaboration task description creation

Human-written task description

"Describe this molecule. "

AI assistance

Diversity task descriptions

"Give me some details about this molecule."
"What can you tell me about this molecule?"
"Provide a brief overview of this molecule."
"Provide a description of this molecule."
"Could you provide a description of this molecule?
......

Sector 2: Information derivation from existing data

Data sources

Crawl

Key information

Molecule SMILES: C=CC(=O)N

CID: 6579

Structure:

Description: The molecule is a colorless, odorless, crystalline solid that can react violently when…

Prompt input to AI

Ask 5 questions and give answers that can be directed to the given paragraph.

Optimal format

Question 1: What is SGK1? Answer 1: SGK1 is a serine/threonine kinase that works under acute…

Sector 3: Template-based conversion of biological data into textual format

Biomolecular annotations

Gene Ontology (biological process)

Gene Ontology (cellular component)

DNA binding

......

Template

The designed protein must be able to regulate {biological process}.

The designed protein localizes to the {cellular component}.

A protein with {DNA binding} capability for targeted gene regulation.

......

Textual form

The designed protein must be able to regulate signal transduction.

The designed protein localizes to the mitochondrion.

A protein with homeobox capability for targeted gene regulation.

......

Sector 4: Quality control

Molecule

Raw

Filter

Chemically valid

- [C][=C][C][=Branch1][C]…
- [C][=C][C][=C][C][Ring1]…

Robust descriptor

Protein

Manual Reviewed

Unreviewed

Annotation abundant scores

Target

Sequence identity matching

Query

≥90%

<90%

Obtained biomolecules

⭐ High quality

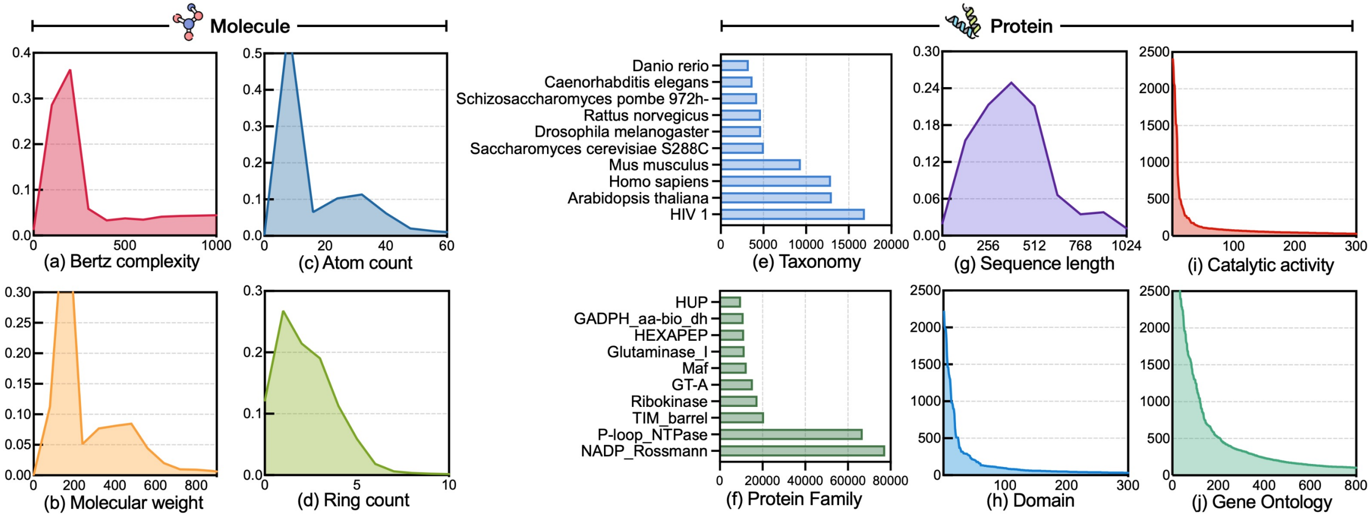
⭐ Chemically valid

⭐ Non-redundant

Transforming existing databases into instructions via preprocessing.

Ensuring sequence quality for molecules and proteins.

Data Analysis



	Features	Example
Molecule	Chemical properties	It combines with metals to make fluorides such as sodium fluoride and calcium fluoride.
	Physical properties	The molecule is a colorless, flammable gas that has a distinct, pungent smell.
	Applications	Used as a flavoring, solvent, and polymerization catalyst.
	Environment	The molecule is a metal that occurs naturally throughout the environment, in rocks, soil, water, and air.
	Safety	Lethal by inhalation and highly toxic or lethal by skin absorption.
	Formation	It is formed in foods that are rich in carbohydrates when they are fried, grilled, or baked.
Protein	Function	The designed protein must be able to regulate signal transduction.
	Subcellular location	The designed protein localizes to the mitochondrion.
	Structure	The target protein must exhibit Helix as its primary conformation.
	Family & Domain	The designed protein should contain PWWP domain that is essential for its function.
	PTM / Processing	Incorporate a signal peptide in the protein design.

Diversity of sequences

Coverage of descriptions

CONTENT

SEEKING TRUTH
PURSUING INNOVATION

01 Introduction & Background

02 Data Construction & Analysis

03 Experiments

04 Conclusion & Future Work

Results on Mol-Instructions

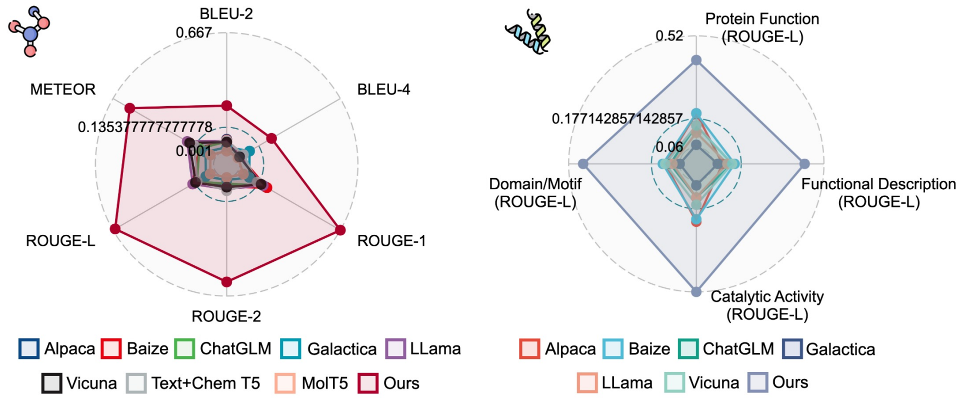
Molecular Generation

Model	Exact↑	BLEU↑	Levenshtein↓	RDK FTS↑	MACCS FTS↑	Morgan FTS↑	Validity↑
Description-guided Molecule Design							
Alpaca	0.000	0.004	51.088	0.006	0.029	0.000	0.002
Baize	0.000	0.006	53.796	0.000	0.000	0.000	0.002
ChatGLM	0.000	0.004	53.157	0.005	0.000	0.000	0.005
LLama	0.000	0.003	59.864	0.005	0.000	0.000	0.003
Vicuna	0.000	0.006	60.356	0.006	0.001	0.000	0.001
Galactica	0.000	0.192	44.152	0.135	0.248	0.088	0.992
Text+Chem T5	0.097	0.508	41.819	0.352	0.474	0.353	0.721
MolT5	0.112	0.546	38.276	0.400	0.538	0.295	0.773
Ours	0.002	0.345	41.367	0.231	0.412	0.147	1.000
Reagent Prediction							
Alpaca	0.000	0.026	29.037	0.029	0.016	0.001	0.186
Baize	0.000	0.051	30.628	0.022	0.018	0.004	0.099
ChatGLM	0.000	0.019	29.169	0.017	0.006	0.002	0.074
LLama	0.000	0.003	28.040	0.037	0.001	0.001	0.001
Vicuna	0.000	0.010	27.948	0.038	0.002	0.001	0.007
Galactica	0.000	0.141	30.760	0.036	0.127	0.051	0.995
Text+Chem T5	0.000	0.225	49.323	0.039	0.186	0.052	0.313
Ours	0.044	0.224	23.167	0.237	0.364	0.213	1.000
Forward Reaction Prediction							
Alpaca	0.000	0.065	41.989	0.004	0.024	0.008	0.138
Baize	0.000	0.044	41.500	0.004	0.025	0.009	0.097
ChatGLM	0.000	0.183	40.008	0.050	0.100	0.044	0.108
LLama	0.000	0.020	42.002	0.001	0.002	0.001	0.039
Vicuna	0.000	0.057	41.690	0.007	0.016	0.006	0.059
Galactica	0.000	0.468	35.021	0.156	0.257	0.097	0.946
Text+Chem T5	0.239	0.782	20.413	0.705	0.789	0.652	0.762
Ours	0.045	0.654	27.262	0.313	0.509	0.262	1.000
Retrosynthesis							
Alpaca	0.000	0.063	46.915	0.005	0.023	0.007	0.160
Baize	0.000	0.095	44.714	0.025	0.050	0.023	0.112
ChatGLM	0.000	0.117	48.365	0.056	0.075	0.043	0.046
LLama	0.000	0.036	46.844	0.018	0.029	0.017	0.010
Vicuna	0.000	0.057	46.877	0.025	0.030	0.021	0.017
Galactica	0.000	0.452	34.940	0.167	0.274	0.134	0.986
Text+Chem T5	0.141	0.765	24.043	0.685	0.765	0.585	0.698
Ours	0.009	0.705	31.227	0.283	0.487	0.230	1.000

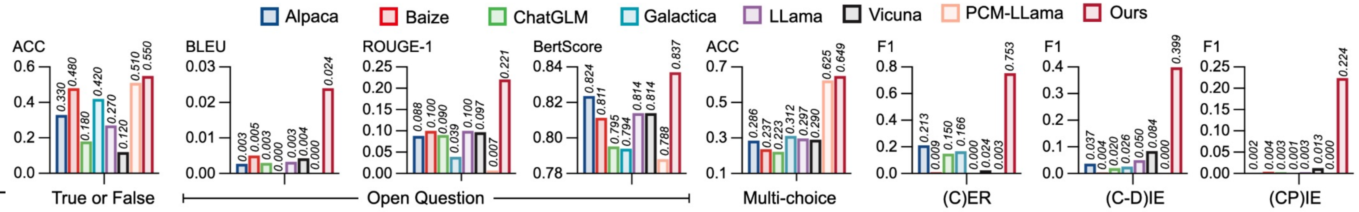
Property Prediction

Model	MAE ↓
Property Prediction	
Alpaca	322.109
Baize	261.343
ChatGLM	-
LLama	5.553
Vicuna	860.051
Galactica	0.568
Ours	↑0.555 **0.013**

Molecule & Protein Understanding



Biotext Natural Language Processing



CONTENT

SEEKING TRUTH
PURSUING INNOVATION

01 Introduction & Background

02 Data Construction & Analysis

03 Experiments

04 Conclusion & Future Work

Take Away

❑ This study bridges the gap in current resources and advances LLM training in the biomolecular domain:

  ❑ accessing cross-modal comprehension in general models

  ❑ advancing research and innovation in biomolecular design by collecting and organizing a wide range of design standards

  ❑ aiding models in understanding biomolecular properties and reactions without explicit programming

Limitations

❑ **Distinct representation spaces of text and biomolecules**

❑ **Limitations imposed by LoRA's training strategy**

Future Work

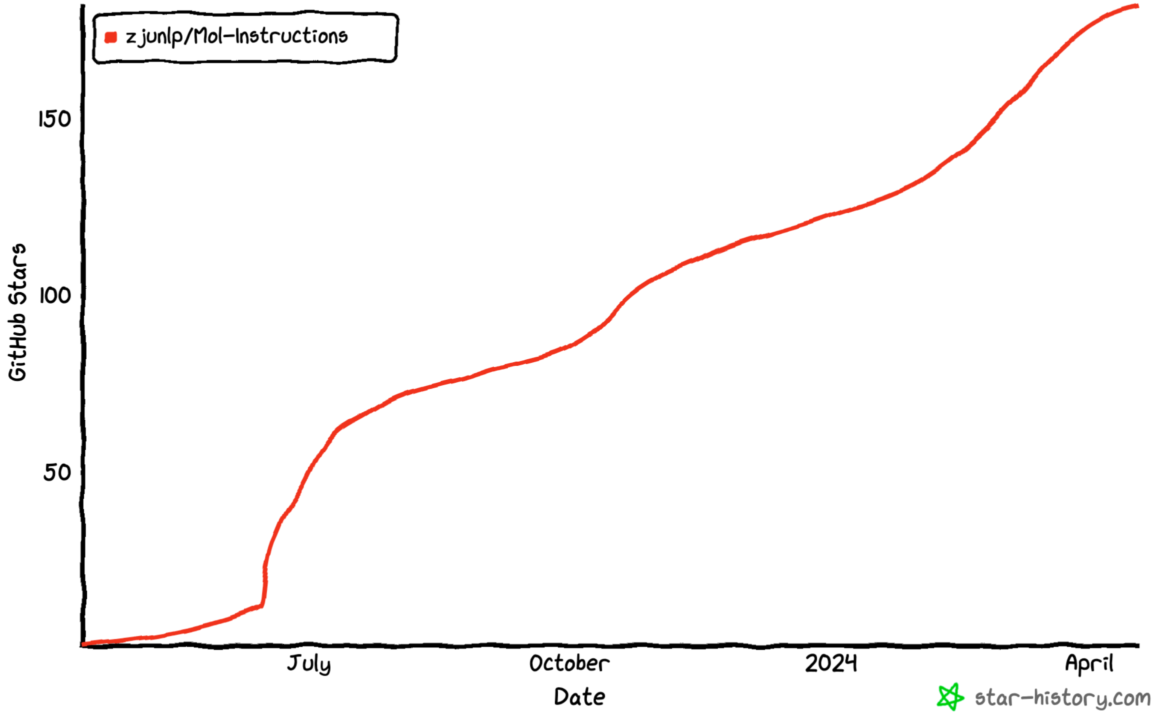
❑ Expand the vocabulary with technical terms

❑ Model with multimodal techniques

Open Source

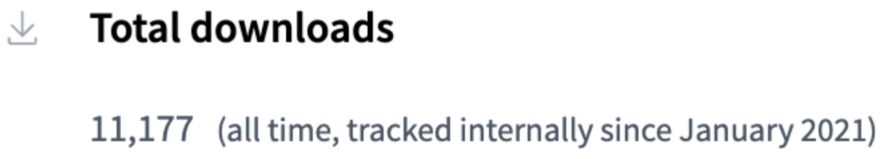
Github

github.com/zjunlp/Mol-Instructions



Hugging Face

zjunlp/Mol-Instructions



Total downloads

11,177 (all time, tracked internally since January 2021)

Thank you!

Code

Data

浙江大學
ZHEJIANG UNIVERSITY