CSCE 689 - Special Topics in NLP for Science

Lecture 6: Scientific Knowledge Extraction

Yu Zhang

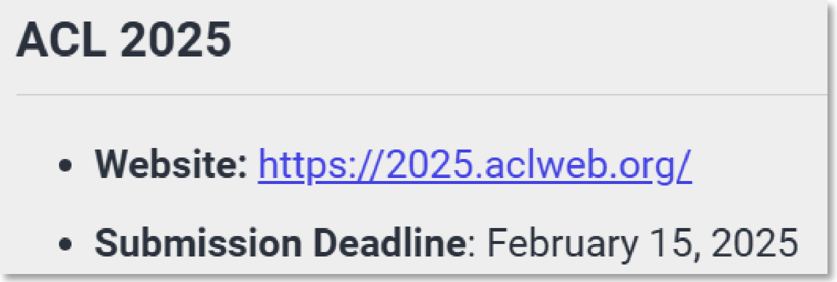
yuzhang@tamu.edu

February 4, 2025

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html

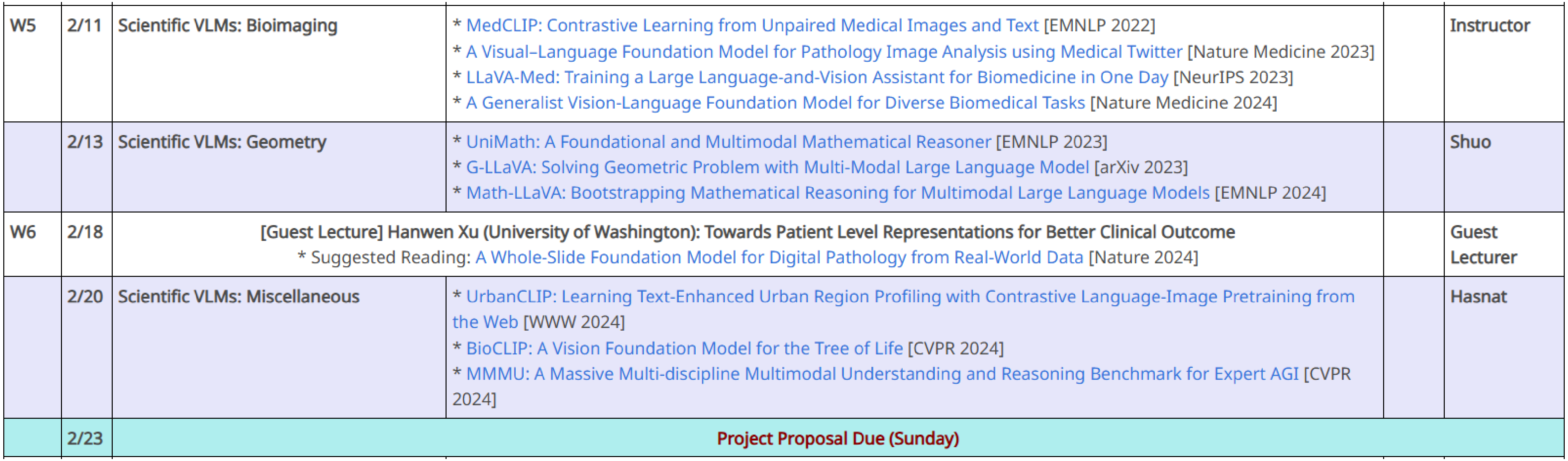
Project Proposal Deadline Change (2/16 → 2/23)

- Reason: 2/15 is the deadline of the ACL conference, and many students are working on their paper submissions.



ACL 2025

- **Website:** https://2025.aclweb.org/
- **Submission Deadline**: February 15, 2025



W5	2/11	Scientific VLMs: Bioimaging	* MedCLIP: Contrastive Learning from Unpaired Medical Images and Text [EMNLP 2022] * A Visual–Language Foundation Model for Pathology Image Analysis using Medical Twitter [Nature Medicine 2023] * LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day [NeurIPS 2023] * A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks [Nature Medicine 2024]		Instructor
	2/13	Scientific VLMs: Geometry	* UniMath: A Foundational and Multimodal Mathematical Reasoner [EMNLP 2023] * G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model [arXiv 2023] * Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models [EMNLP 2024]		Shuo
W6	2/18	[Guest Lecture] Hanwen Xu (University of Washington): Towards Patient Level Representations for Better Clinical Outcome * Suggested Reading: A Whole-Slide Foundation Model for Digital Pathology from Real-World Data [Nature 2024]			Guest Lecturer
	2/20	Scientific VLMs: Miscellaneous	* UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web [WWW 2024] * BioCLIP: A Vision Foundation Model for the Tree of Life [CVPR 2024] * MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI [CVPR 2024]		Hasnat
	2/23	Project Proposal Due (Sunday)			

Agenda

- Fundamental Scientific Information Extraction Tasks
  - Named Entity Recognition: AIONER
  - Relation Extraction: SciER
- Advanced Scientific Information Extraction Tasks
  - Chemical Reaction Extraction: ReactIE
  - Action Extraction: ActionIE

Agenda

- Fundamental Scientific Information Extraction Tasks
  - Named Entity Recognition: AIONER
  - Relation Extraction: SciER
- Advanced Scientific Information Extraction Tasks
  - Chemical Reaction Extraction: ReactIE
  - Action Extraction: ActionIE

Recap: Named Entity Recognition

- Named Entity Recognition (NER): Given a sentence, find entities (i.e., token spans) of certain types (e.g., chemical, disease, gene, species, variant, cell line).

… human complement factor H deficiency associated with hemolytic uremic syndrome …

DISEASE DISEASE

Input	human	complement	factor	H	deficiency	associated	with
Output	B-DISEASE	I-DISEASE	I-DISEASE	I-DISEASE	I-DISEASE	O	O

- The BIO schema: B (beginning of an entity), I (in an entity), O (out of an entity)

- NER → predicting a label for each token in the sentence

Real-World NER Datasets

NCBI-disease [1]	B-Disease, I-Disease, O	→	Extract Disease
NLM-Chem [2]	B-Chemical, I-Chemical, O	→	Extract Chemical
Linnaeus [3]	B-Species, I-Species, O	→	Extract Species
BC5CDR [4]	B-Disease, I-Disease, B-Chemical, I-Chemical, O	→	Extract Disease and Chemical

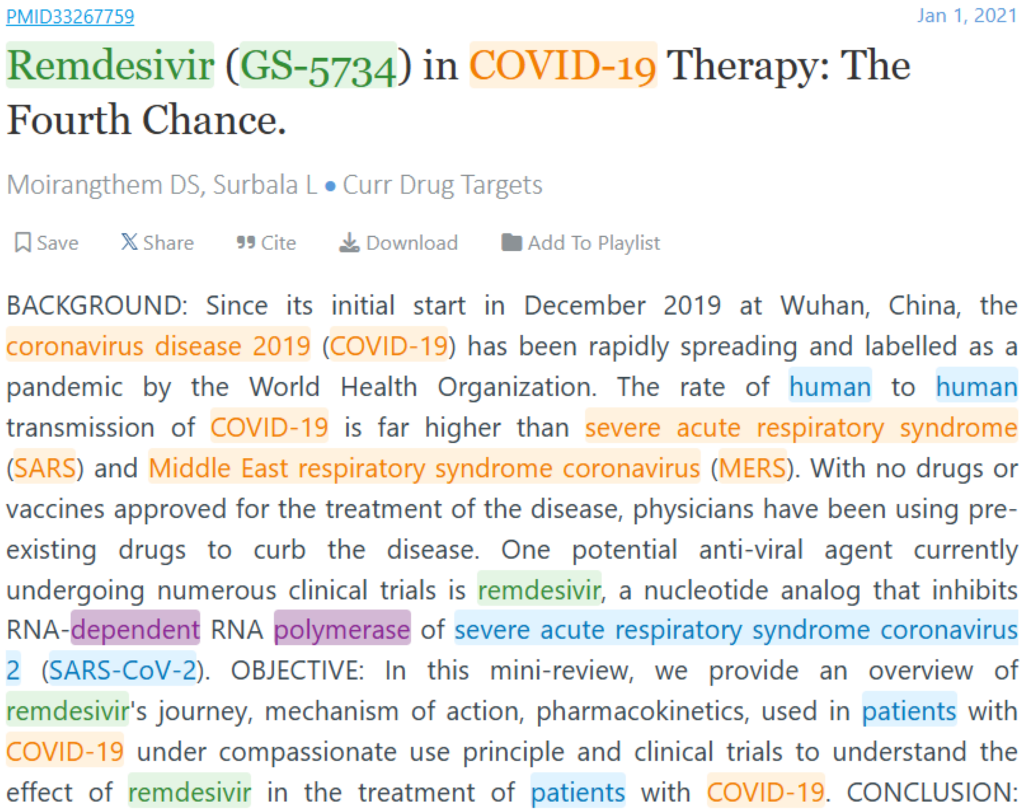
…

[1] *NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization.* Journal of Biomedical Informatics 2014.
[2] *NLM-Chem, A New Resource for Chemical Entity Recognition in PubMed Full Text Literature.* Scientific Data 2021.
[3] *Linnaeus: A Species Name Identification System for Biomedical Literature.* BMC Bioinformatics 2010.
[4] *BioCreative V CDR Task Corpus: A Resource for Chemical Disease Relation Extraction.* Database 2016.

Real-World NER Application

- How to train an NER model using these datasets that can recognize all annotated entity types?

https://www.ncbi.nlm.nih.gov/research/pubtator3



PMID33267759

Jan 1, 2021

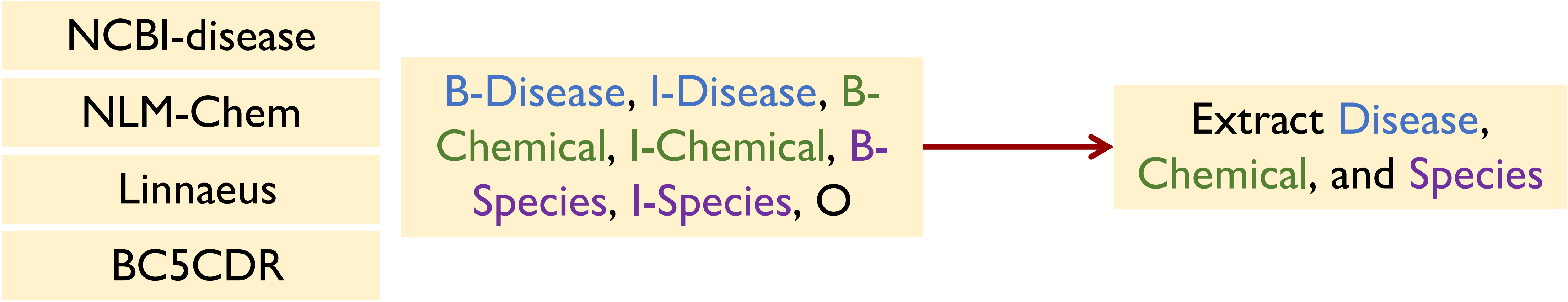
Remdesivir (GS-5734) in COVID-19 Therapy: The Fourth Chance.

Moirangthem DS, Surbala L • Curr Drug Targets

BACKGROUND: Since its initial start in December 2019 at Wuhan, China, the coronavirus disease 2019 (COVID-19) has been rapidly spreading and labelled as a pandemic by the World Health Organization. The rate of human to human transmission of COVID-19 is far higher than severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome coronavirus (MERS). With no drugs or vaccines approved for the treatment of the disease, physicians have been using pre-existing drugs to curb the disease. One potential anti-viral agent currently undergoing numerous clinical trials is remdesivir, a nucleotide analog that inhibits RNA-dependent RNA polymerase of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). OBJECTIVE: In this mini-review, we provide an overview of remdesivir's journey, mechanism of action, pharmacokinetics, used in patients with COVID-19 under compassionate use principle and clinical trials to understand the effect of remdesivir in the treatment of patients with COVID-19. CONCLUSION:

PubTator 3.0: An AI-Powered Literature Resource for Unlocking Biomedical Knowledge. Nucleic Acids Research 2024.

Bad Solution 1: Directly Combining All Training Data Together



NCBI-disease

NLM-Chem

Linnaeus

BC5CDR

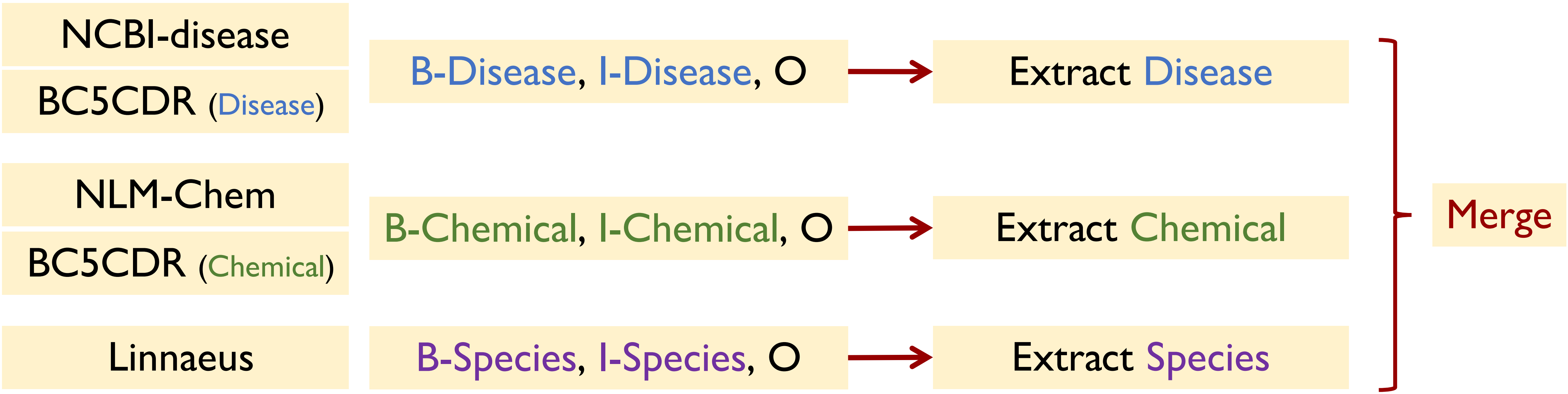
B-Disease, I-Disease, B-Chemical, I-Chemical, B-Species, I-Species, O

Extract Disease, Chemical, and Species

Text	in	rats	,	nitrofurantoin	causes	pulmonary	toxicity
Annotations in Linnaeus	O	B-Species	O	O	O	O	O
Correct Annotations	O	B-Species	O	B-Chemical	O	B-Disease	I-Disease

- Incorrect annotations (false negatives) used as training data!!

Bad Solution 2: Merging the Results of Each Entity Type



NCBI-disease

BC5CDR (Disease)

B-Disease, I-Disease, O → Extract Disease

NLM-Chem

BC5CDR (Chemical)

B-Chemical, I-Chemical, O → Extract Chemical

Linnaeus

B-Species, I-Species, O → Extract Species

Merge

Text	in	rats	,	nitrofurantoin	causes	pulmonary	toxicity
Prediction of Disease	O	O	O	O	O	B-Disease	I-Disease
Prediction of Chemical	O	O	O	B-Chemical	I-Chemical	I-Chemical	O

- Cannot handle conflicts among predictions!!

Why is the task non-trivial?

- The meanings of “O” are different in different datasets.

NCBI-disease	B-Disease, I-Disease, O	O: NOT Disease
NLM-Chem	B-Chemical, I-Chemical, O	O: NOT Chemical
Linnaeus	B-Species, I-Species, O	O: NOT Species
BC5CDR	B-Disease, I-Disease, B-Chemical, I-Chemical, O	O: NOT Disease and NOT Chemical

Rewriting the Label "O"

- Assume we have an annotated sentence from BC5CDR

Text	in	rats	,	nitrofurantoin	causes	pulmonary	toxicity
Annotations	O	O	O	B-Chemical	O	B-Disease	I-Disease

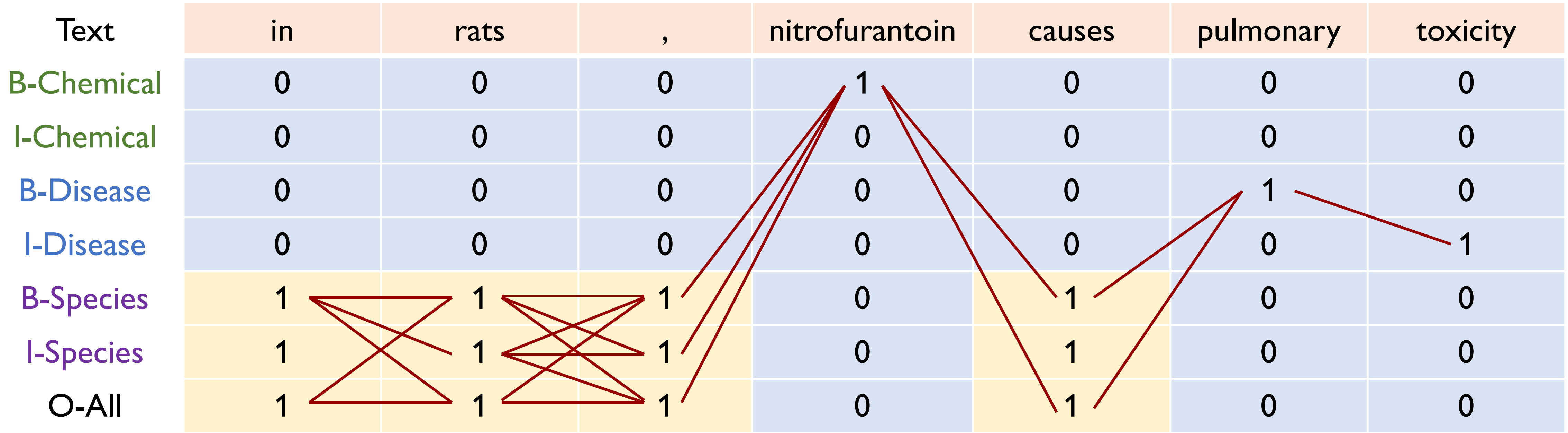
- If we are just performing Chemical and Disease NER, we can adopt a "one-hot" representation of the ground truth.

Text	in	rats	,	nitrofurantoin	causes	pulmonary	toxicity
B-Chemical	0	0	0	1	0	0	0
I-Chemical	0	0	0	0	0	0	0
B-Disease	0	0	0	0	0	1	0
I-Disease	0	0	0	0	0	0	1
O	1	1	1	0	1	0	0

Rewriting the Label “O”

- If we are performing all-type NER, “O” should be interpreted as “anything else”.

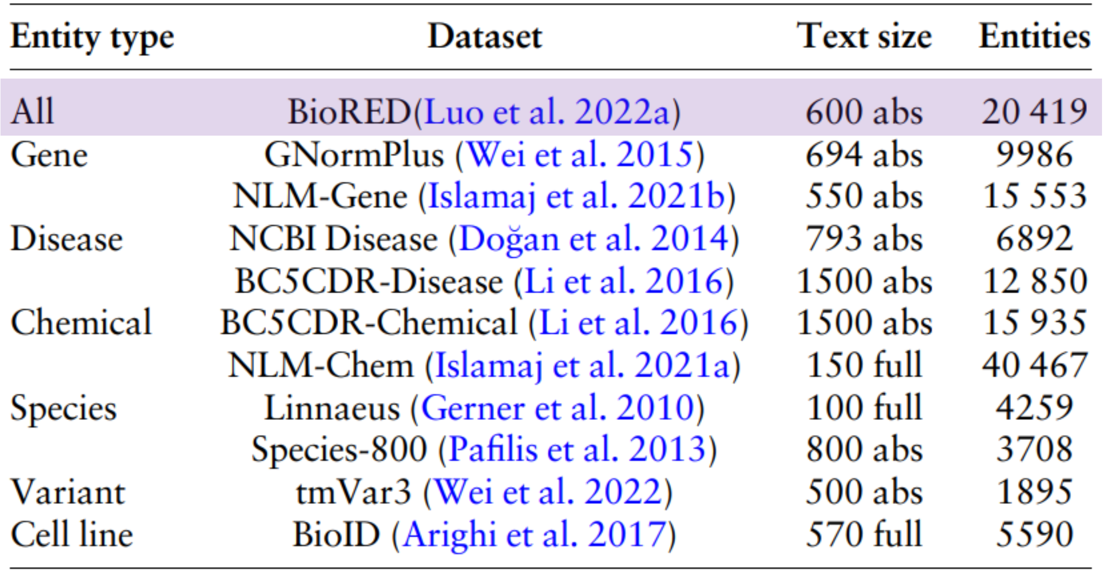
Text	in	rats	,	nitrofurantoin	causes	pulmonary	toxicity
Annotations	O	O	O	B-Chemical	O	B-Disease	I-Disease



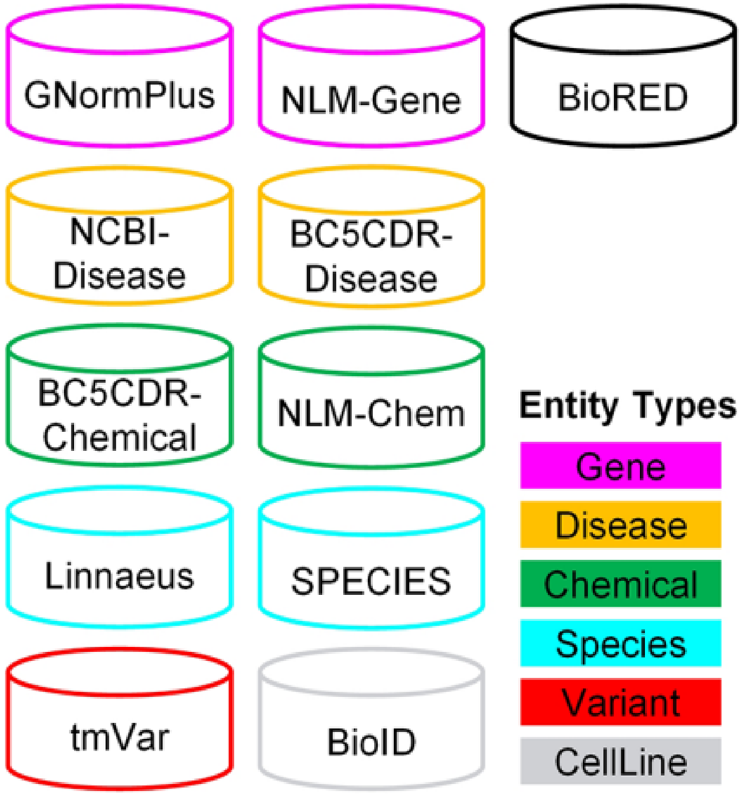
Text	in	rats	,	nitrofurantoin	causes	pulmonary	toxicity
B-Chemical	0	0	0	1	0	0	0
I-Chemical	0	0	0	0	0	0	0
B-Disease	0	0	0	0	0	1	0
I-Disease	0	0	0	0	0	0	1
B-Species	1	1	1	0	1	0	0
I-Species	1	1	1	0	1	0	0
O-All	1	1	1	0	1	0	0

Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets. EMNLP 2018.

What if you have an annotated dataset for all-type NER?



Entity type	Dataset	Text size	Entities
All	BioRED(Luo et al. 2022a)	600 abs	20 419
Gene	GNormPlus (Wei et al. 2015)	694 abs	9986
	NLM-Gene (Islamaj et al. 2021b)	550 abs	15 553
Disease	NCBI Disease (Doğan et al. 2014)	793 abs	6892
	BC5CDR-Disease (Li et al. 2016)	1500 abs	12 850
Chemical	BC5CDR-Chemical (Li et al. 2016)	1500 abs	15 935
	NLM-Chem (Islamaj et al. 2021a)	150 full	40 467
Species	Linnaeus (Gerner et al. 2010)	100 full	4259
	Species-800 (Pafilis et al. 2013)	800 abs	3708
Variant	tmVar3 (Wei et al. 2022)	500 abs	1895
Cell line	BioID (Arighi et al. 2017)	570 full	5590



BioRED: A Rich Biomedical Relation Extraction Dataset. Briefings in Bioinformatics 2022.

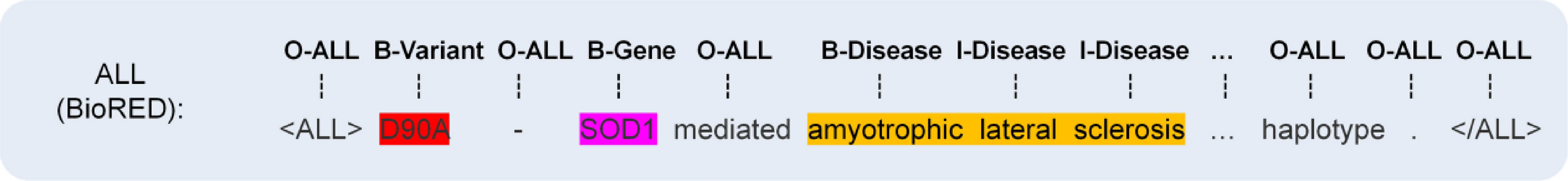
What if you have an annotated dataset for all-type NER?

NCBI-disease BC5CDR (Disease)	B-Disease, I-Disease, O-Disease
NLM-Chem BC5CDR (Chemical)	B-Chemical, I-Chemical, O-Chemical
Linnaeus Species-800	B-Species, I-Species, O-Species
	…
BioRED	B-Disease, I-Disease, B-Chemical, I-Chemical, B-Species, I-Species, B-Gene, I-Gene, B-Variant, I-Variant, …, O-All

AIONER: All-in-One Scheme-Based Biomedical Named Entity Recognition using Deep Learning. Bioinformatics 2023.

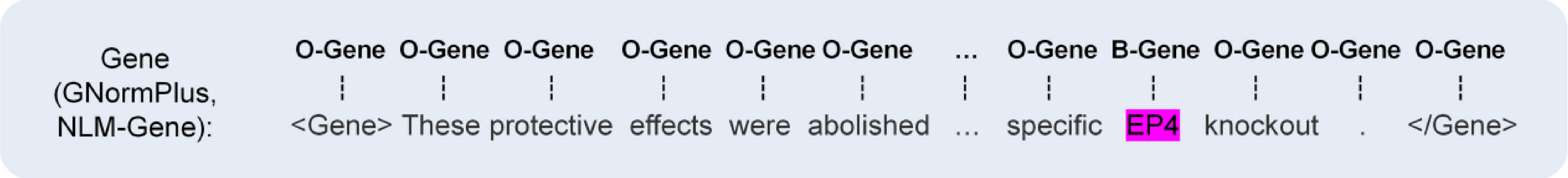
Supporting Both One-Type and All-Type NER

- Prepend/append special tokens to the sentence to indicate your task
- For all-type NER (e.g., BioRED)



	O-ALL	B-Variant	O-ALL	B-Gene	O-ALL	B-Disease	I-Disease	I-Disease	…	O-ALL	O-ALL	O-ALL
ALL (BioRED):	<ALL>	D90A	-	SOD1	mediated	amyotrophic	lateral	sclerosis	…	haplotype	.	</ALL>

- For one-type NER (e.g., other training sets)



	O-Gene	O-Gene	O-Gene	O-Gene	O-Gene	O-Gene	…	O-Gene	B-Gene	O-Gene	O-Gene	O-Gene
Gene (GNormPlus, NLM-Gene):	<Gene>	These	protective	effects	were	abolished	…	specific	EP4	knockout	.	</Gene>

AIONER: All-in-One Scheme-Based Biomedical Named Entity Recognition using Deep Learning. Bioinformatics 2023.

Performance of AIONER

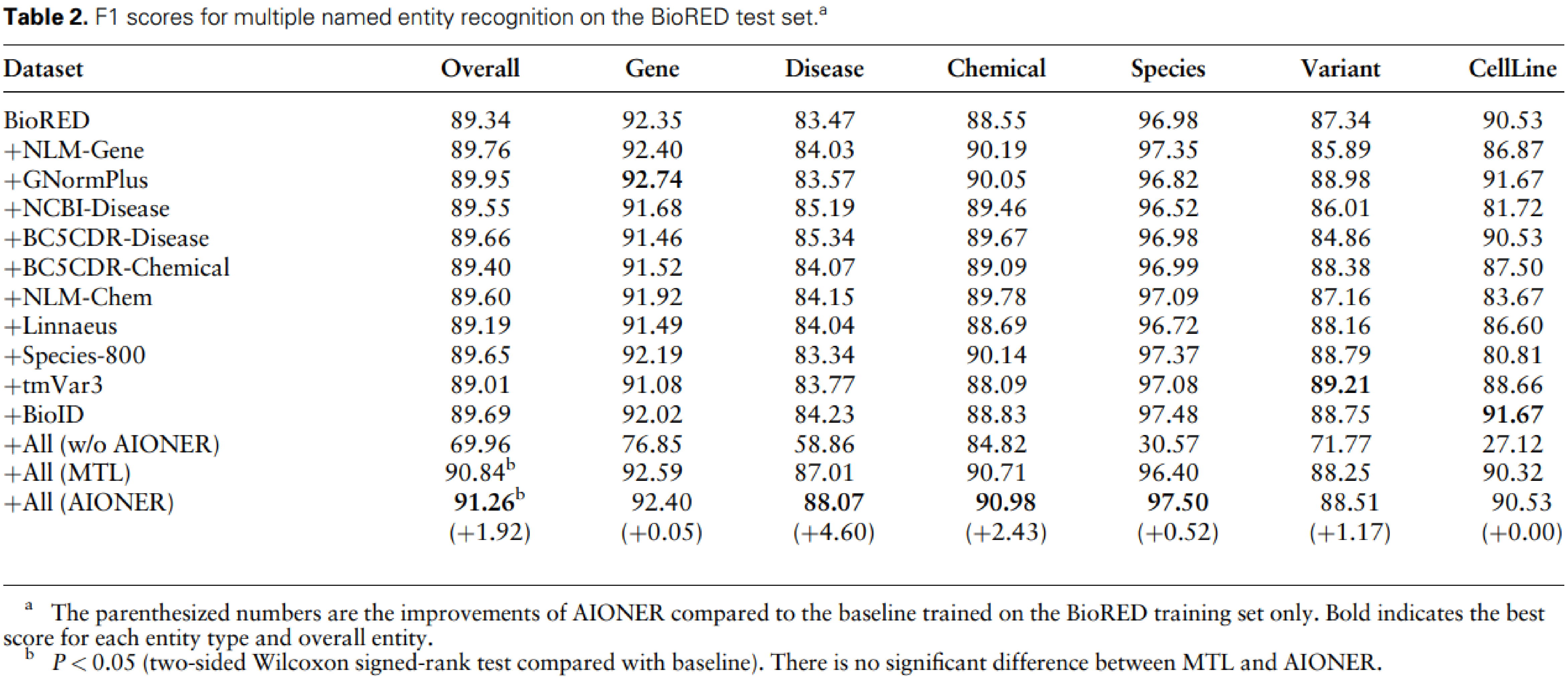


Table 2. F1 scores for multiple named entity recognition on the BioRED test set.[a]

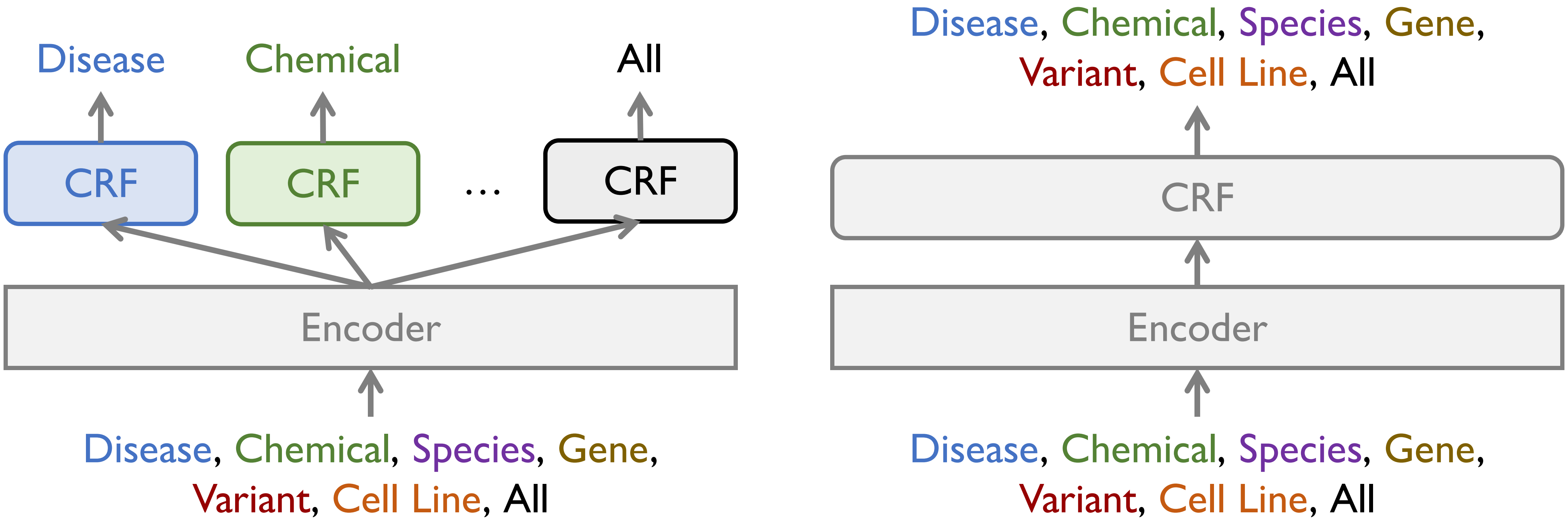
Dataset	Overall	Gene	Disease	Chemical	Species	Variant	CellLine
BioRED	89.34	92.35	83.47	88.55	96.98	87.34	90.53
+NLM-Gene	89.76	92.40	84.03	90.19	97.35	85.89	86.87
+GNormPlus	89.95	**92.74**	83.57	90.05	96.82	88.98	91.67
+NCBI-Disease	89.55	91.68	85.19	89.46	96.52	86.01	81.72
+BC5CDR-Disease	89.66	91.46	85.34	89.67	96.98	84.86	90.53
+BC5CDR-Chemical	89.40	91.52	84.07	89.09	96.99	88.38	87.50
+NLM-Chem	89.60	91.92	84.15	89.78	97.09	87.16	83.67
+Linnaeus	89.19	91.49	84.04	88.69	96.72	88.16	86.60
+Species-800	89.65	92.19	83.34	90.14	97.37	88.79	80.81
+tmVar3	89.01	91.08	83.77	88.09	97.08	**89.21**	88.66
+BioID	89.69	92.02	84.23	88.83	97.48	88.75	**91.67**
+All (w/o AIONER)	69.96	76.85	58.86	84.82	30.57	71.77	27.12
+All (MTL)	90.84[b]	92.59	87.01	90.71	96.40	88.25	90.32
+All (AIONER)	**91.26**[b] (+1.92)	92.40 (+0.05)	**88.07** (+4.60)	**90.98** (+2.43)	**97.50** (+0.52)	88.51 (+1.17)	90.53 (+0.00)

[a] The parenthesized numbers are the improvements of AIONER compared to the baseline trained on the BioRED training set only. Bold indicates the best score for each entity type and overall entity.

[b] $P < 0.05$ (two-sided Wilcoxon signed-rank test compared with baseline). There is no significant difference between MTL and AIONER.

AIONER: All-in-One Scheme-Based Biomedical Named Entity Recognition using Deep Learning. Bioinformatics 2023.

Multi-Task Learning vs. AIONER



Cross-Type Biomedical Named Entity Recognition with Deep Multi-Task Learning. Bioinformatics 2019.

Performance of AIONER

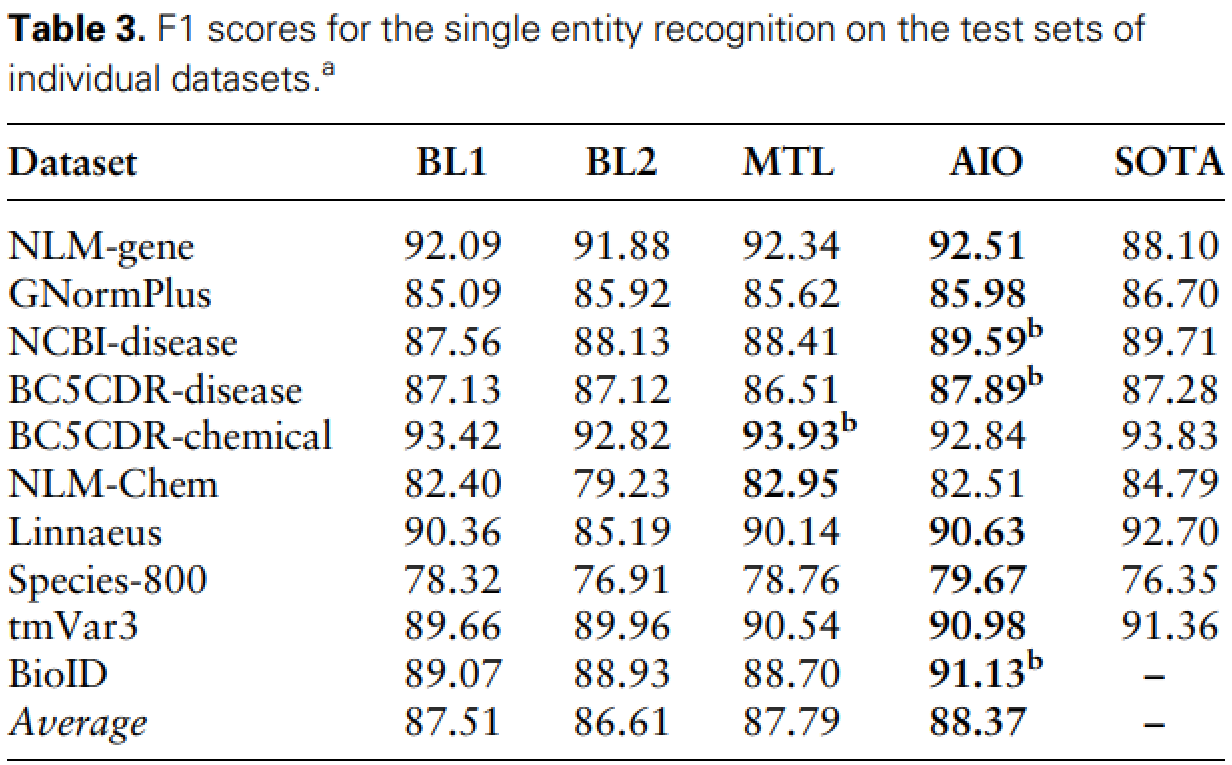


Table 3. F1 scores for the single entity recognition on the test sets of individual datasets.[a]

Dataset	BL1	BL2	MTL	AIO	SOTA
NLM-gene	92.09	91.88	92.34	**92.51**	88.10
GNormPlus	85.09	85.92	85.62	**85.98**	86.70
NCBI-disease	87.56	88.13	88.41	**89.59**[b]	89.71
BC5CDR-disease	87.13	87.12	86.51	**87.89**[b]	87.28
BC5CDR-chemical	93.42	92.82	**93.93**[b]	92.84	93.83
NLM-Chem	82.40	79.23	**82.95**	82.51	84.79
Linnaeus	90.36	85.19	90.14	**90.63**	92.70
Species-800	78.32	76.91	78.76	**79.67**	76.35
tmVar3	89.66	89.96	90.54	**90.98**	91.36
BioID	89.07	88.93	88.70	**91.13**[b]	–
Average	87.51	86.61	87.79	**88.37**	–

AIONER: All-in-One Scheme-Based Biomedical Named Entity Recognition using Deep Learning. Bioinformatics 2023.

Take-Away Messages

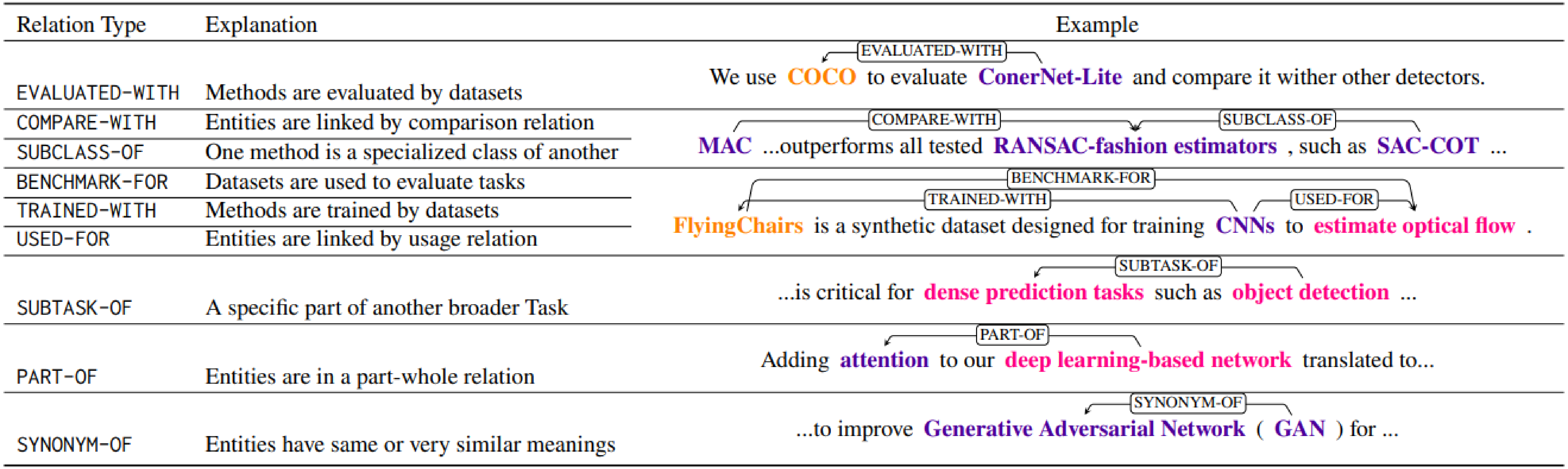
- Unlike general-domain NER, there are lots of partially-annotated datasets for scientific NER. Simple heuristics to combine these datasets together usually cannot work because of the ambiguity of the label "O".
- Rewriting the label "O" in training sets makes partial annotations useful in both all-type and one-type NER.
- Limitation:
  - AIONER still relies on at least one all-type annotated training set. If all training sets are partially typed, one should adopt marginal likelihood training of CRF.
  - *Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets.* EMNLP 2018.

Agenda

- Fundamental Scientific Information Extraction Tasks
  - Named Entity Recognition: AIONER
  - Relation Extraction: SciER
- Advanced Scientific Information Extraction Tasks
  - Chemical Reaction Extraction: ReactIE
  - Action Extraction: ActionIE

Constructing an NER Benchmark for Computer Science

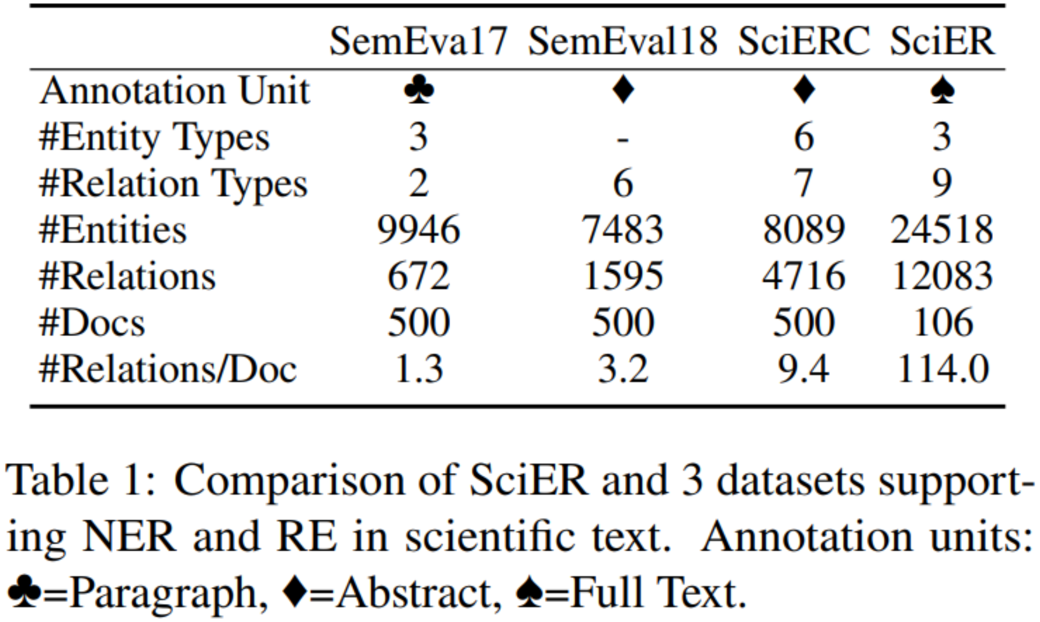
- 3 types of entities: DATASET, METHOD, and TASK
- 9 types of relations


EVALUATED-WITH
COMPARE-WITH
SUBCLASS-OF
BENCHMARK-FOR
TRAINED-WITH
USED-FOR
SUBTASK-OF
PART-OF
SYNONYM-OF

Relation Type	Explanation	Example
EVALUATED-WITH	Methods are evaluated by datasets	We use COCO to evaluate ConerNet-Lite and compare it wither other detectors.
COMPARE-WITH	Entities are linked by comparison relation	MAC ...outperforms all tested RANSAC-fashion estimators , such as SAC-COT ...
SUBCLASS-OF	One method is a specialized class of another	
BENCHMARK-FOR	Datasets are used to evaluate tasks	FlyingChairs is a synthetic dataset designed for training CNNs to estimate optical flow .
TRAINED-WITH	Methods are trained by datasets	
USED-FOR	Entities are linked by usage relation	
SUBTASK-OF	A specific part of another broader Task	...is critical for dense prediction tasks such as object detection ...
PART-OF	Entities are in a part-whole relation	Adding attention to our deep learning-based network translated to...
SYNONYM-OF	Entities have same or very similar meanings	...to improve Generative Adversarial Network (GAN) for ...

SciER: An Entity and Relation Extraction Dataset for Datasets, Methods, and Tasks in Scientific Documents. EMNLP 2024.

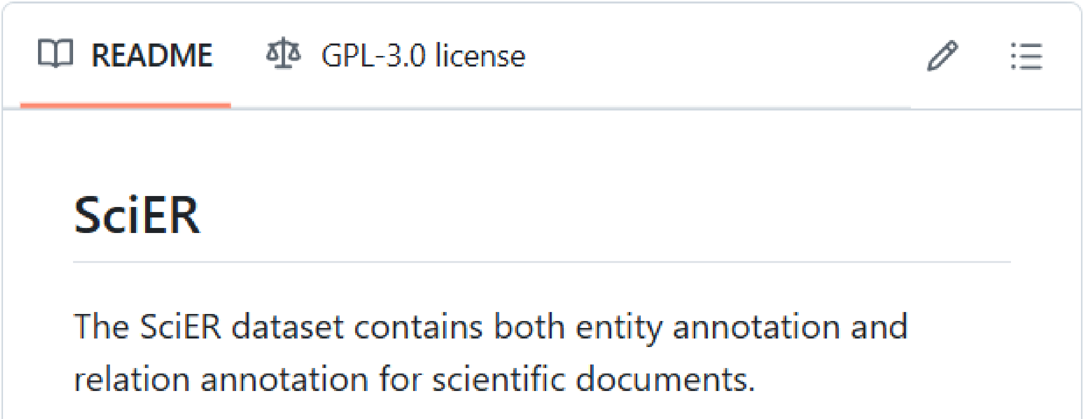
More Details of SciER



	SemEval17	SemEval18	SciERC	SciER
Annotation Unit	♣	♦	♦	♠
#Entity Types	3	-	6	3
#Relation Types	2	6	7	9
#Entities	9946	7483	8089	24518
#Relations	672	1595	4716	12083
#Docs	500	500	500	106
#Relations/Doc	1.3	3.2	9.4	114.0

Table 1: Comparison of SciER and 3 datasets supporting NER and RE in scientific text. Annotation units: ♣=Paragraph, ♦=Abstract, ♠=Full Text.

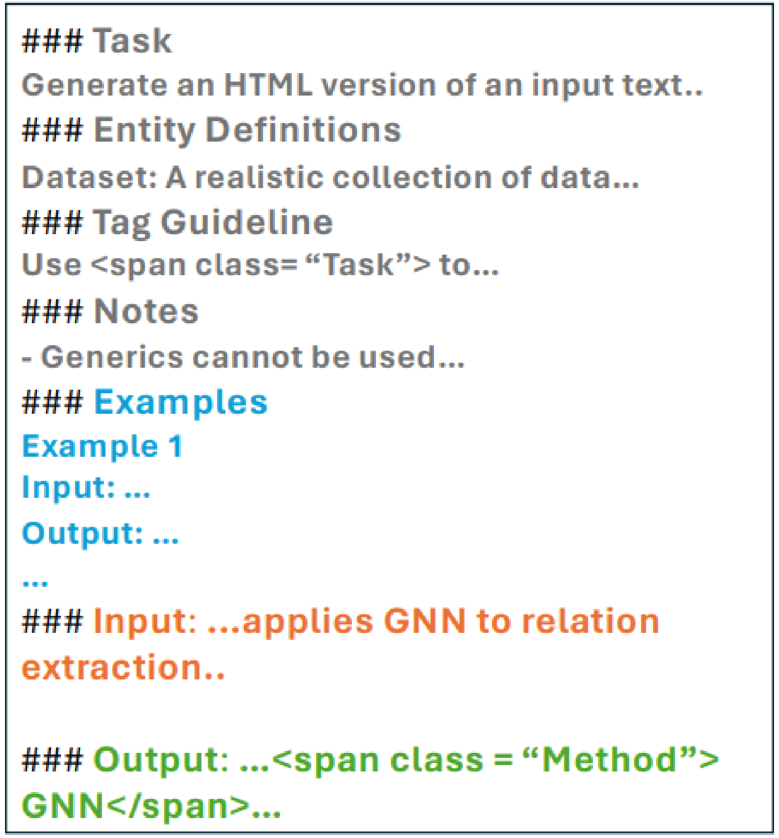
https://github.com/edzq/SciER



SciER

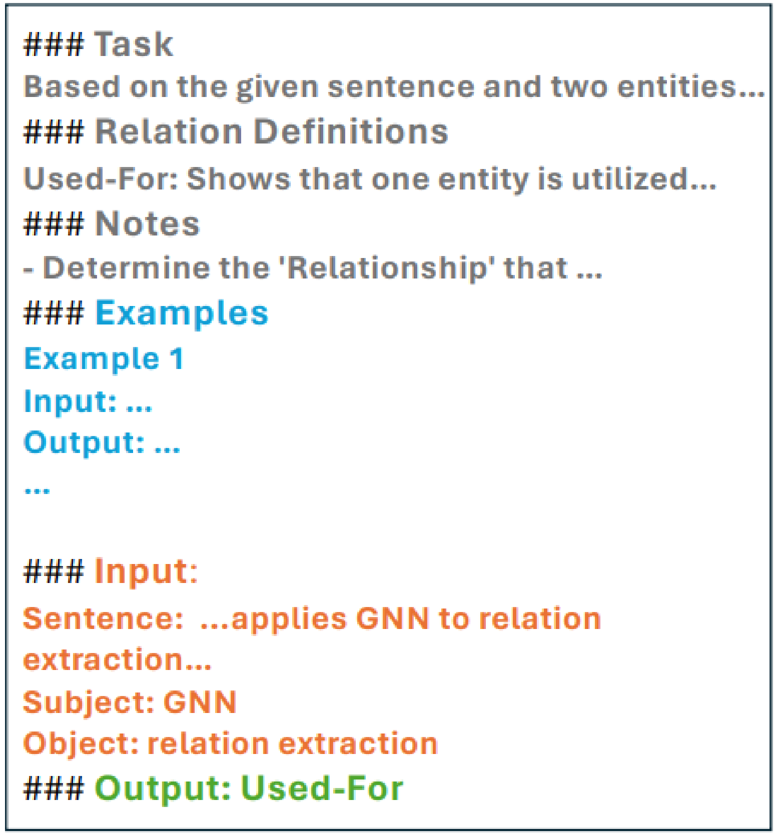
The SciER dataset contains both entity annotation and relation annotation for scientific documents.

How to use LLM in-context learning to perform NER and RE?



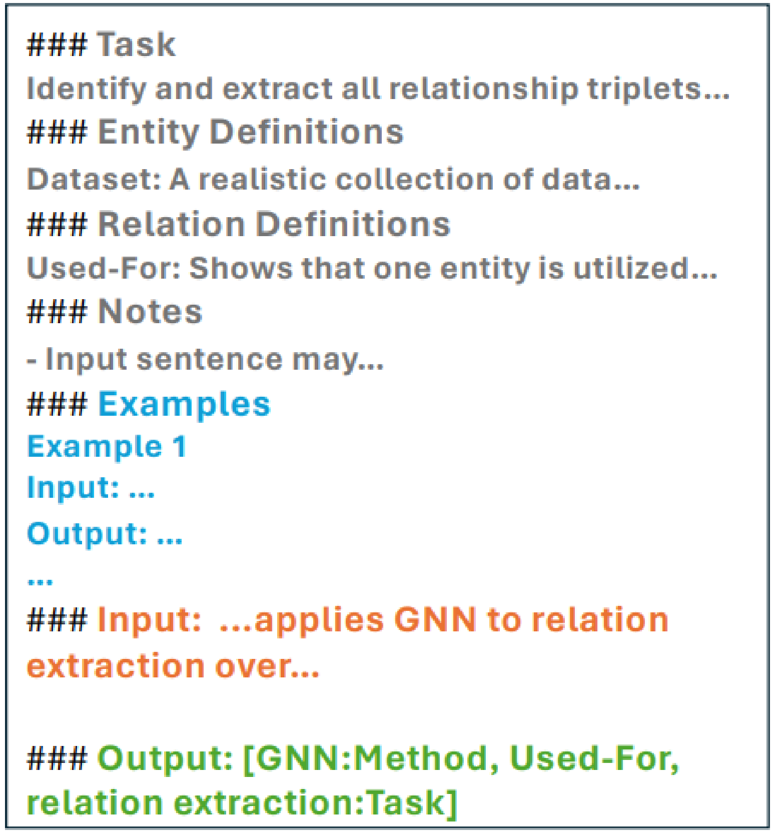
Task
Generate an HTML version of an input text..
Entity Definitions
Dataset: A realistic collection of data...
Tag Guideline
Use <span class= “Task”> to...
Notes
- Generics cannot be used...
Examples
Example 1
Input: ...
Output: ...
...
Input: ...applies GNN to relation extraction..

Output: ...<span class = “Method”> GNN</span>...



Task
Based on the given sentence and two entities...
Relation Definitions
Used-For: Shows that one entity is utilized...
Notes
- Determine the 'Relationship' that ...
Examples
Example 1
Input: ...
Output: ...
...

Input:
Sentence: ...applies GNN to relation extraction...
Subject: GNN
Object: relation extraction
Output: Used-For



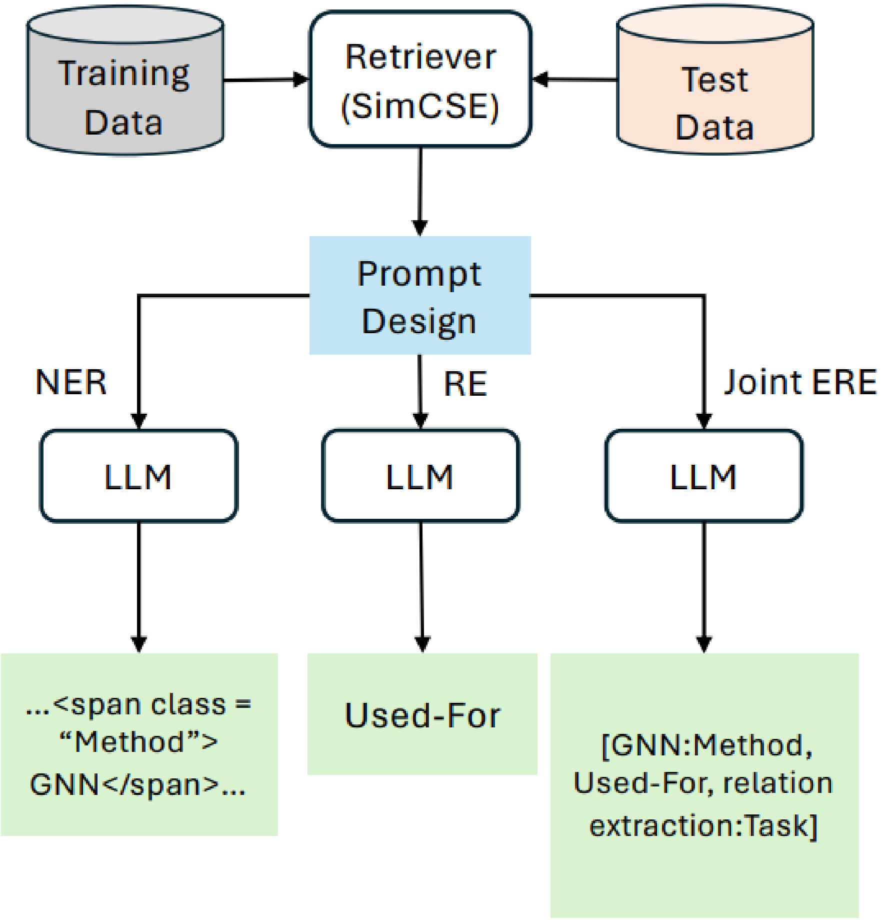
Task
Identify and extract all relationship triplets...
Entity Definitions
Dataset: A realistic collection of data...
Relation Definitions
Used-For: Shows that one entity is utilized...
Notes
- Input sentence may...
Examples
Example 1
Input: ...
Output: ...
...
Input: ...applies GNN to relation extraction over...

Output: [GNN:Method, Used-For, relation extraction:Task]

SciER:An Entity and Relation Extraction Dataset for Datasets, Methods, and Tasks in Scientific Documents. EMNLP 2024.

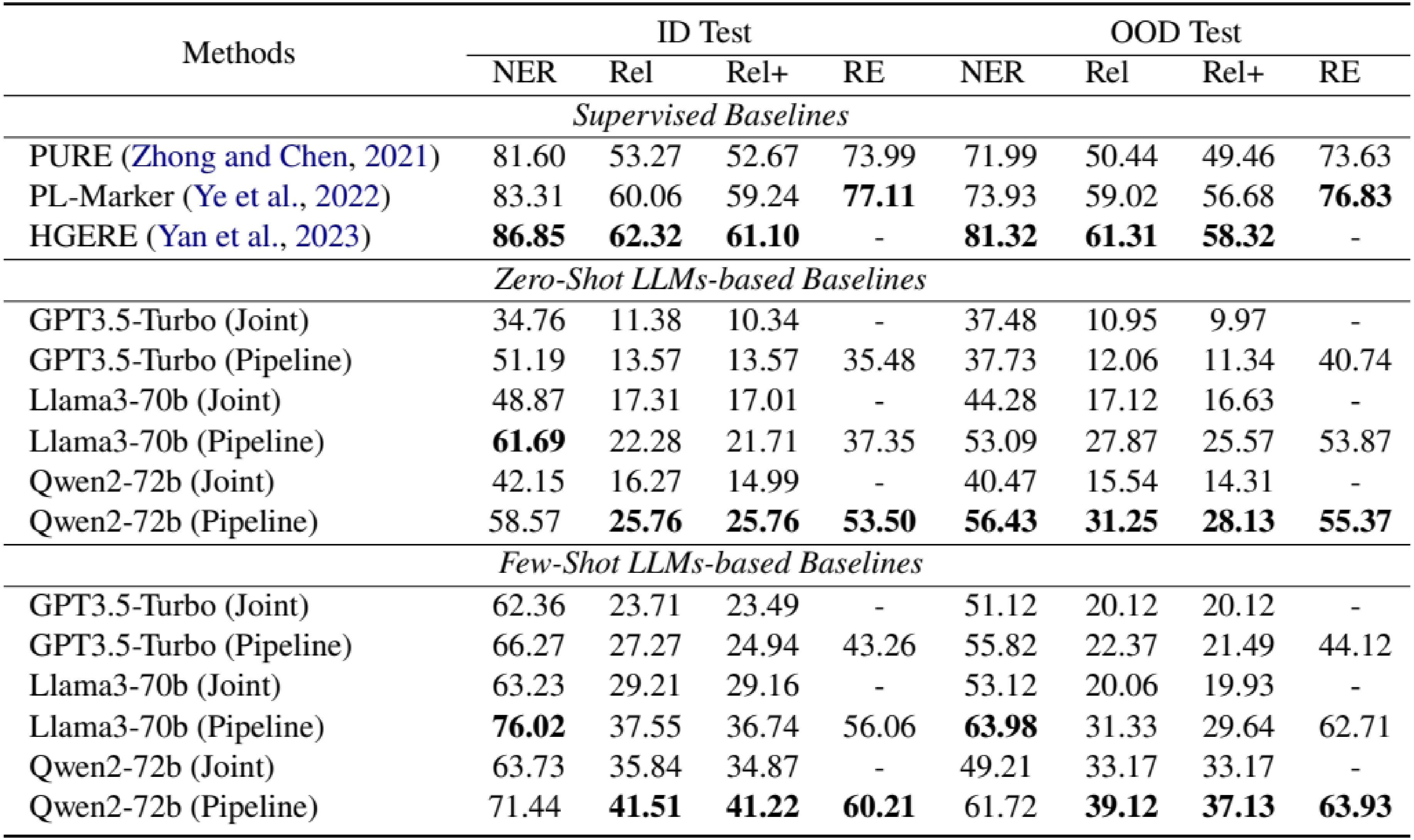
Retrieving In-Context Learning Examples

- LLMs cannot take all annotated samples for in-context learning.
- Use sentence embeddings to retrieve the most similar annotated samples.
  - Purely based on text
  - No entity/relation information used to learn the embeddings


Training Data
Retriever (SimCSE)
Test Data
Prompt Design
NER
RE
Joint ERE
LLM
LLM
LLM
...<span class = "Method"> GNN</span>...
Used-For
[GNN:Method, Used-For, relation extraction:Task]

SciER: An Entity and Relation Extraction Dataset for Datasets, Methods, and Tasks in Scientific Documents. EMNLP 2024.

Performance of LLMs on SciER



Methods	ID Test				OOD Test			
	NER	Rel	Rel+	RE	NER	Rel	Rel+	RE
Supervised Baselines								
PURE (Zhong and Chen, 2021)	81.60	53.27	52.67	73.99	71.99	50.44	49.46	73.63
PL-Marker (Ye et al., 2022)	83.31	60.06	59.24	**77.11**	73.93	59.02	56.68	**76.83**
HGERE (Yan et al., 2023)	**86.85**	**62.32**	**61.10**	-	**81.32**	**61.31**	**58.32**	-
Zero-Shot LLMs-based Baselines								
GPT3.5-Turbo (Joint)	34.76	11.38	10.34	-	37.48	10.95	9.97	-
GPT3.5-Turbo (Pipeline)	51.19	13.57	13.57	35.48	37.73	12.06	11.34	40.74
Llama3-70b (Joint)	48.87	17.31	17.01	-	44.28	17.12	16.63	-
Llama3-70b (Pipeline)	**61.69**	22.28	21.71	37.35	53.09	27.87	25.57	53.87
Qwen2-72b (Joint)	42.15	16.27	14.99	-	40.47	15.54	14.31	-
Qwen2-72b (Pipeline)	58.57	**25.76**	**25.76**	**53.50**	**56.43**	**31.25**	**28.13**	**55.37**
Few-Shot LLMs-based Baselines								
GPT3.5-Turbo (Joint)	62.36	23.71	23.49	-	51.12	20.12	20.12	-
GPT3.5-Turbo (Pipeline)	66.27	27.27	24.94	43.26	55.82	22.37	21.49	44.12
Llama3-70b (Joint)	63.23	29.21	29.16	-	53.12	20.06	19.93	-
Llama3-70b (Pipeline)	**76.02**	37.55	36.74	56.06	**63.98**	31.33	29.64	62.71
Qwen2-72b (Joint)	63.73	35.84	34.87	-	49.21	33.17	33.17	-
Qwen2-72b (Pipeline)	71.44	**41.51**	**41.22**	**60.21**	61.72	**39.12**	**37.13**	**63.93**

Joint: Joint ERE

Pipeline: NER→RE

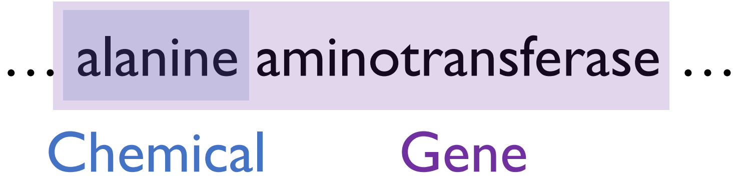
Rel/Rel+: end-to-end RE performance

RE: RE performance given ground-truth NER results

Take-Away Messages

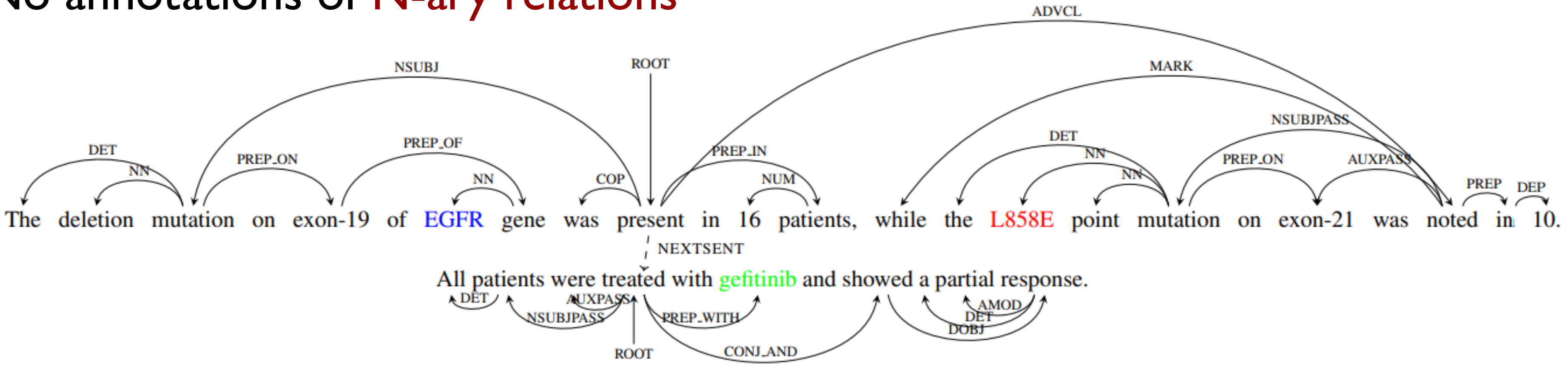
- A new benchmark for Dataset, Method, and Task entity recognition from CS papers
- Annotations on full-text papers provide us with richer types of relations
- Provide a straightforward approach that uses LLM in-context learning for NER and RE
- LLMs with zero or a few examples still significantly underperform fully supervised SOTA.

- Limitations:
    - No annotations of nested entities
    - No annotations of N-ary relations



… alanine aminotransferase …

Chemical Gene

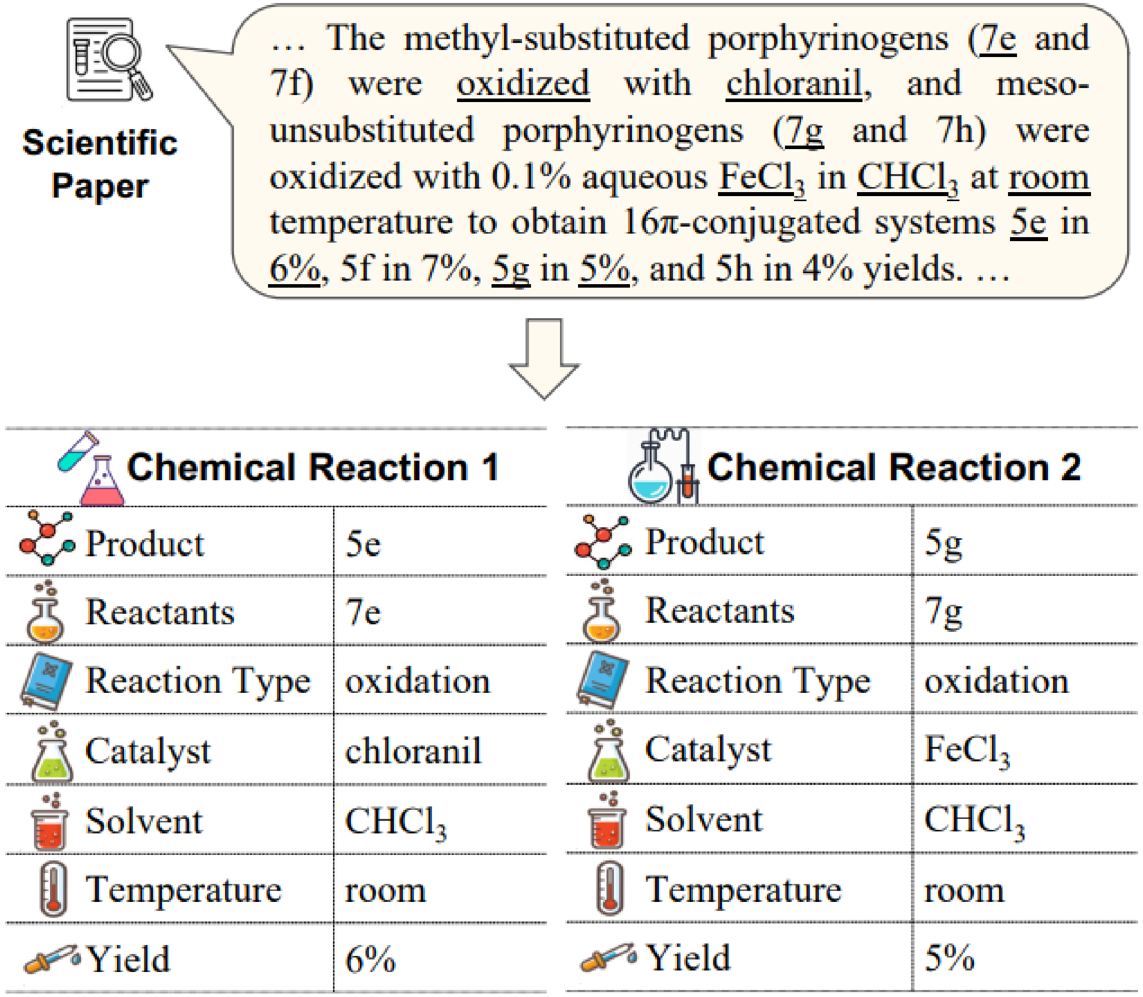


The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10.
All patients were treated with gefitinib and showed a partial response.
ADVCL
NSUBJ
ROOT
MARK
NSUBJPASS
DET
NN
PREP_ON
PREP_OF
NN
COP
PREP_IN
NUM
DET
NN
NN
PREP_ON
AUXPASS
PREP
DEP
NEXTSENT
DET
AUXPASS
NSUBJPASS
PREP_WITH
AMOD
DET
DOBJ
ROOT
CONJ_AND

Agenda

- Fundamental Scientific Information Extraction Tasks
  - Named Entity Recognition: AIONER
  - Relation Extraction: SciER
- Advanced Scientific Information Extraction Tasks
  - Chemical Reaction Extraction: ReactIE
  - Action Extraction: ActionIE

Chemical Reaction Extraction



Scientific Paper

… The methyl-substituted porphyrinogens (7e and 7f) were oxidized with chloranil, and meso-unsubstituted porphyrinogens (7g and 7h) were oxidized with 0.1% aqueous $FeCl_3$ in $CHCl_3$ at room temperature to obtain 16π-conjugated systems 5e in 6%, 5f in 7%, 5g in 5%, and 5h in 4% yields. …

Chemical Reaction 1	
Product	5e
Reactants	7e
Reaction Type	oxidation
Catalyst	chloranil
Solvent	$CHCl_3$
Temperature	room
Yield	6%

Chemical Reaction 2	
Product	5g
Reactants	7g
Reaction Type	oxidation
Catalyst	$FeCl_3$
Solvent	$CHCl_3$
Temperature	room
Yield	5%

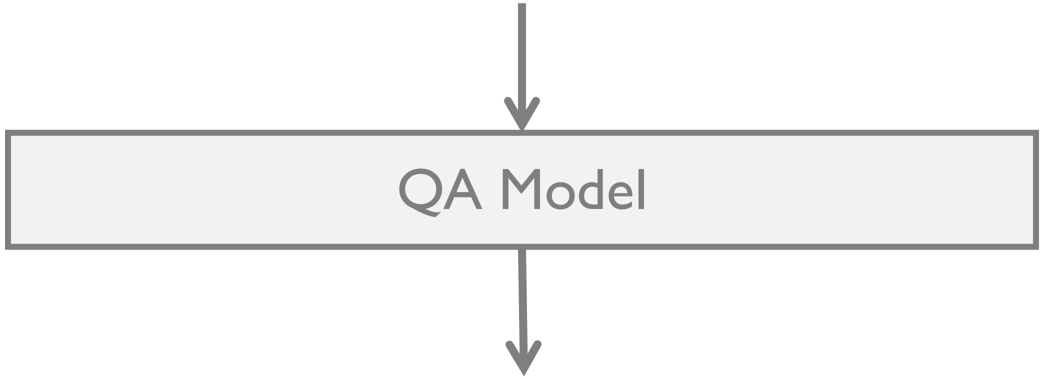
- Pre-define some attributes to be extracted (1st column)
  - Product, Reactants, Reaction Type, Catalyst, …
- Get the values of these attributes from text (2nd column)
  - Some words need necessary conversion (e.g., "*oxidized*"→ "*oxidation*")
  - No longer a sequence labeling task
- (attribute, value) pairs also widely exist in other domains
  - (Task, NER), (Metric, F1), …

ReactIE: Enhancing Chemical Reaction Extraction with Weak Supervision. ACL 2023 Findings.

Chemical Reaction Extraction as a QA Task

- Context (scientific papers): Bromolysis of the C(sp2)-Si bond of 3 with NBS produced bromide 4 as a colorless solid …
- Question: What is the product of the chemical reaction in the text?

you can replace “product” with any other attributes (e.g., “catalyst”)



QA Model

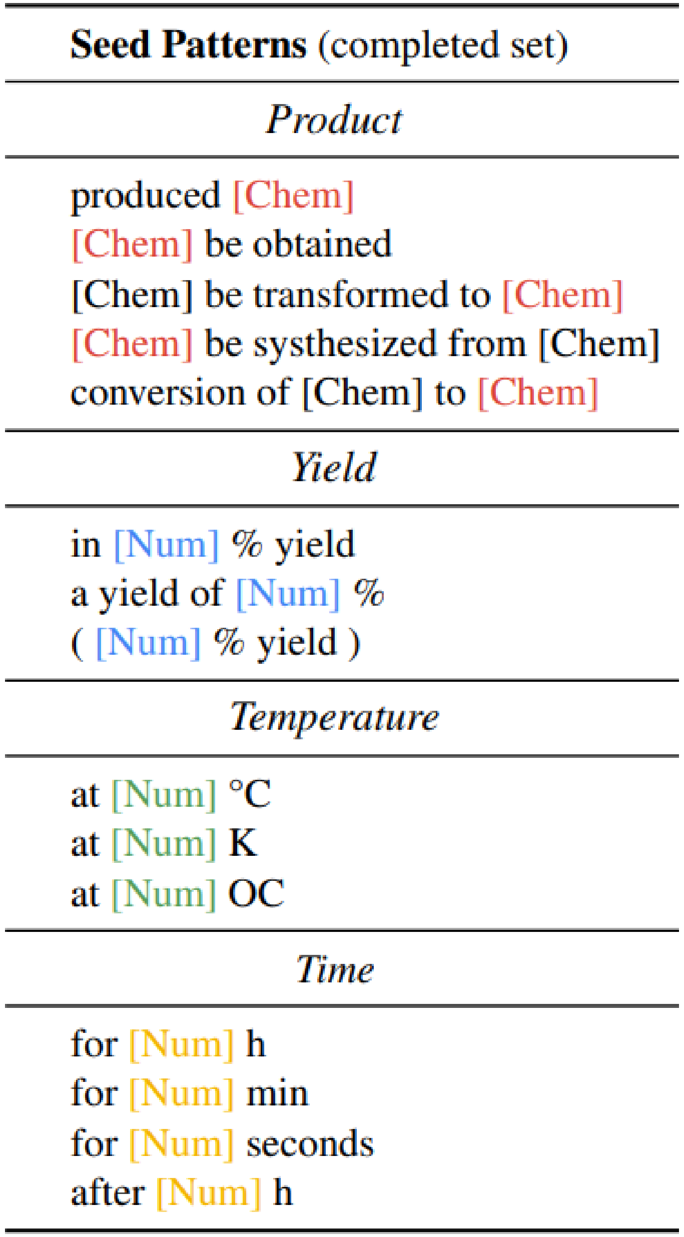
- Answer: bromide 4

How to train the QA model?

- If you already have some annotated data, fine-tune an LLM.
- If you do not have annotated data, start with some rules/patterns.

- Context: Bromolysis of the C(sp2)-Si bond of 3 with NBS produced bromide 4 as a colorless solid …
- Pattern: produced [Chem]
- Answer: bromide 4

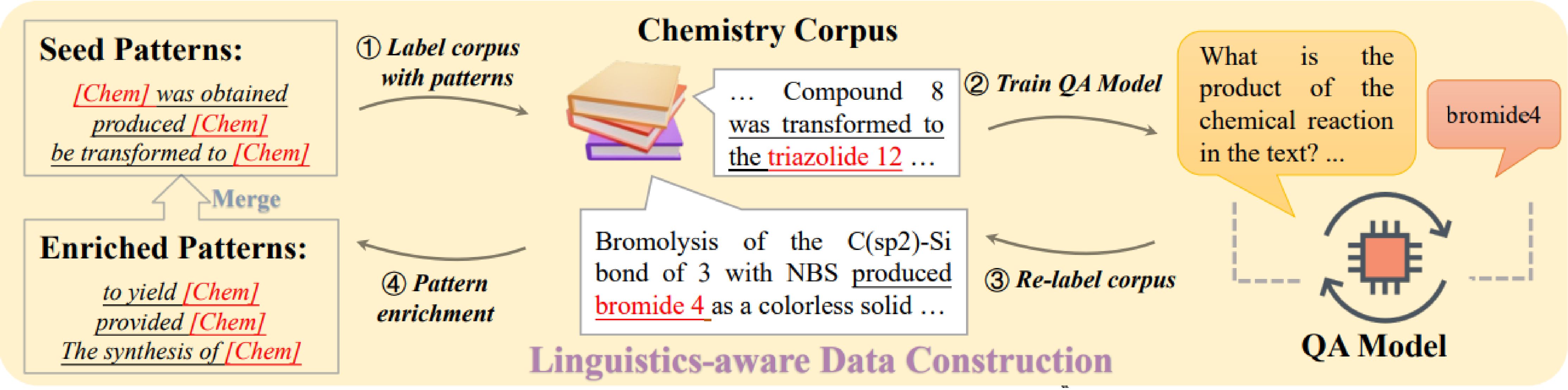
- Pattern-based extraction has high precision but low recall.



Seed Patterns (completed set)
Product
produced [Chem] [Chem] be obtained [Chem] be transformed to [Chem] [Chem] be systhesized from [Chem] conversion of [Chem] to [Chem]
Yield
in [Num] % yield a yield of [Num] % ([Num] % yield)
Temperature
at [Num] °C at [Num] K at [Num] OC
Time

for [Num] h
for [Num] min
for [Num] seconds
after [Num] h

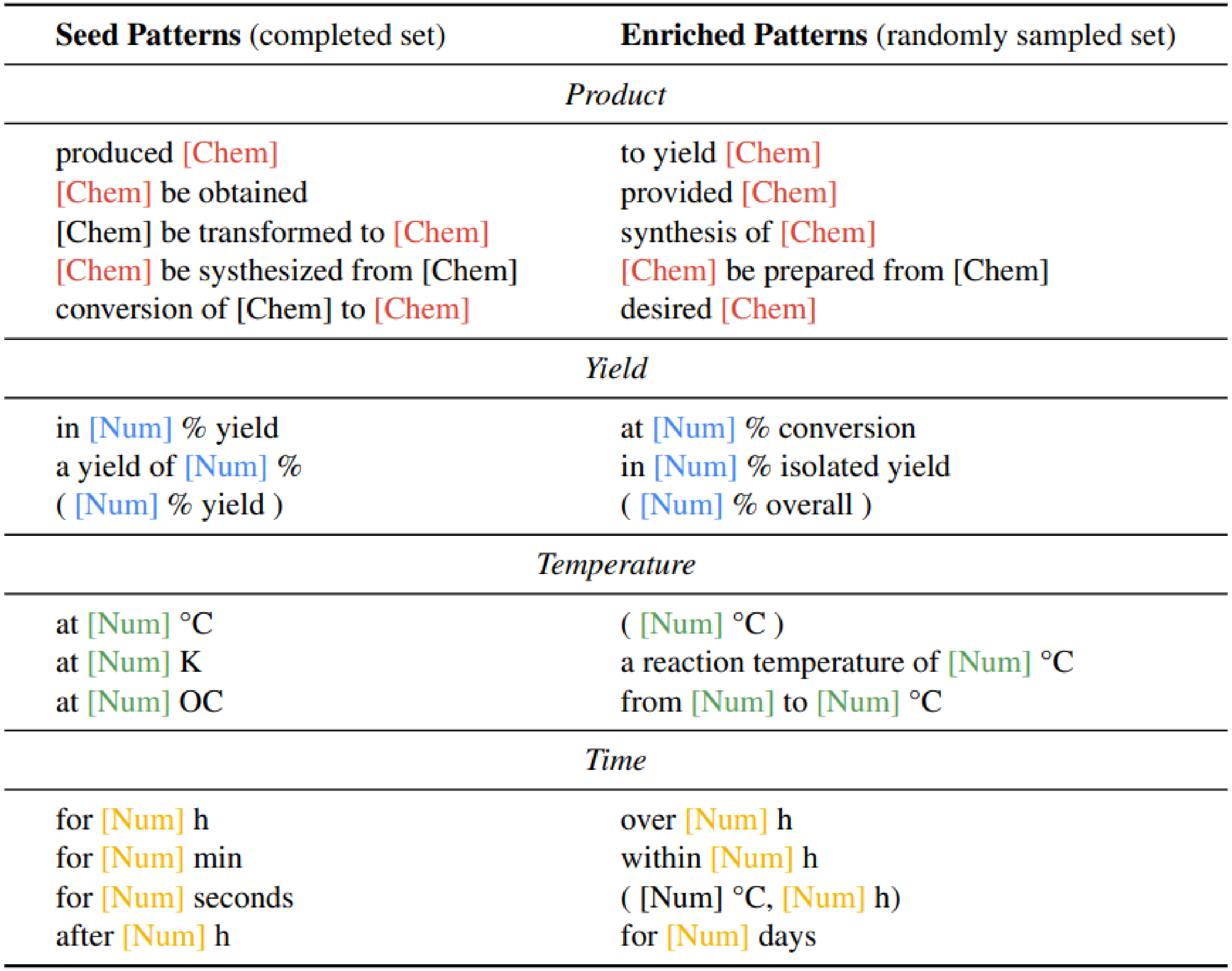
Pattern Enrichment


Seed Patterns:
[Chem] was obtained
produced [Chem]
be transformed to [Chem]
Merge
Enriched Patterns:
to yield [Chem]
provided [Chem]
The synthesis of [Chem]
① Label corpus with patterns
Chemistry Corpus
… Compound 8 was transformed to the triazolide 12 …
② Train QA Model
What is the product of the chemical reaction in the text? ...
bromide4
QA Model
③ Re-label corpus
Bromolysis of the C(sp2)-Si bond of 3 with NBS produced bromide 4 as a colorless solid …
④ Pattern enrichment
Linguistics-aware Data Construction

- Step 4.1: Get the local context (e.g., up to ±3 words) of new matches
  - New match: "… *to yield bromide 4 as a* …"
  - Candidate patterns: "*to yield [Chem]*","*yield [Chem] as*","*[Chem] as a*", …
- Step 4.2: Pick patterns that frequently appear

ReactIE: Enhancing Chemical Reaction Extraction with Weak Supervision. ACL 2023 Findings.

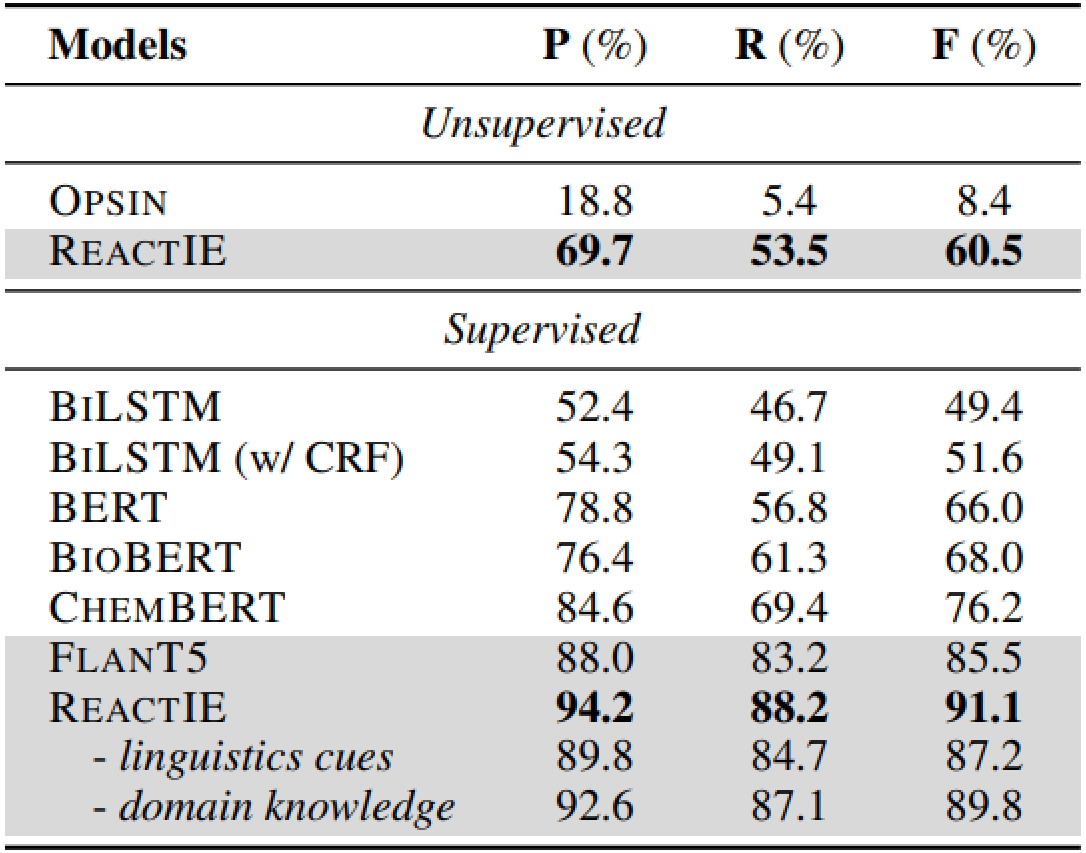
Enriched Patterns



Seed Patterns (completed set)	**Enriched Patterns** (randomly sampled set)
Product	
produced [Chem] [Chem] be obtained [Chem] be transformed to [Chem] [Chem] be systhesized from [Chem] conversion of [Chem] to [Chem]	to yield [Chem] provided [Chem] synthesis of [Chem] [Chem] be prepared from [Chem] desired [Chem]
Yield	
in [Num] % yield a yield of [Num] % ([Num] % yield)	at [Num] % conversion in [Num] % isolated yield ([Num] % overall)
Temperature	
at [Num] °C at [Num] K at [Num] OC	([Num] °C) a reaction temperature of [Num] °C from [Num] to [Num] °C
Time	

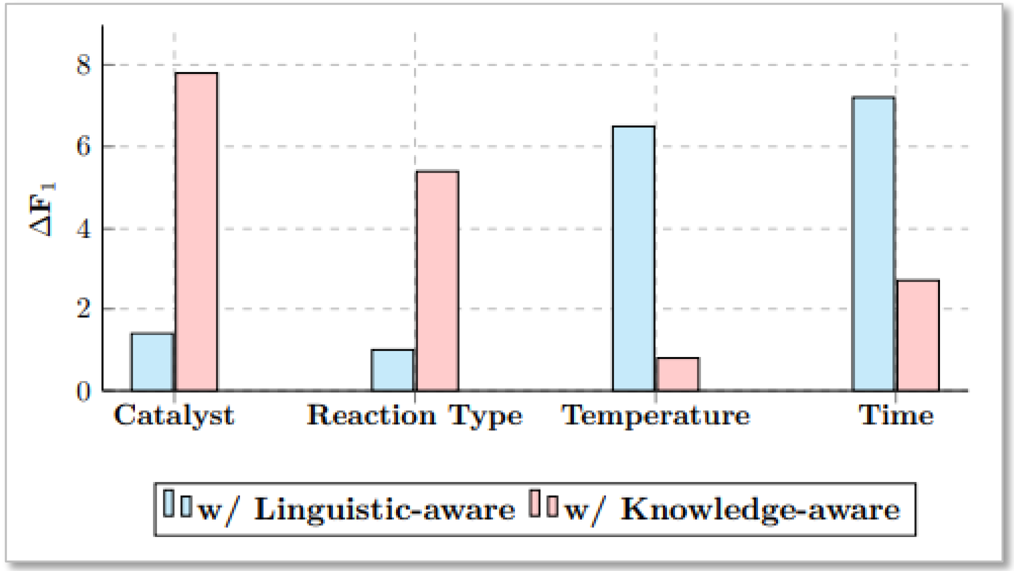
for [Num] h
for [Num] min
for [Num] seconds
after [Num] h
	over [Num] h
	within [Num] h
	([Num] °C, [Num] h)
	for [Num] days

Performance of ReactIE



Models	P (%)	R (%)	F (%)
	Unsupervised		
OPSIN	18.8	5.4	8.4
REACTIE	**69.7**	**53.5**	**60.5**
	Supervised		
BILSTM	52.4	46.7	49.4
BILSTM (w/ CRF)	54.3	49.1	51.6
BERT	78.8	56.8	66.0
BIOBERT	76.4	61.3	68.0
CHEMBERT	84.6	69.4	76.2
FLANT5	88.0	83.2	85.5
REACTIE	**94.2**	**88.2**	**91.1**
- *linguistics cues*	89.8	84.7	87.2
- *domain knowledge*	92.6	87.1	89.8

- Dataset: Reaction Corpus
  - https://github.com/jiangfeng1124/ChemRxnExtractor
- QA model: FLAN-T5



ReactIE: Enhancing Chemical Reaction Extraction with Weak Supervision. ACL 2023 Findings.

Take-Away Messages

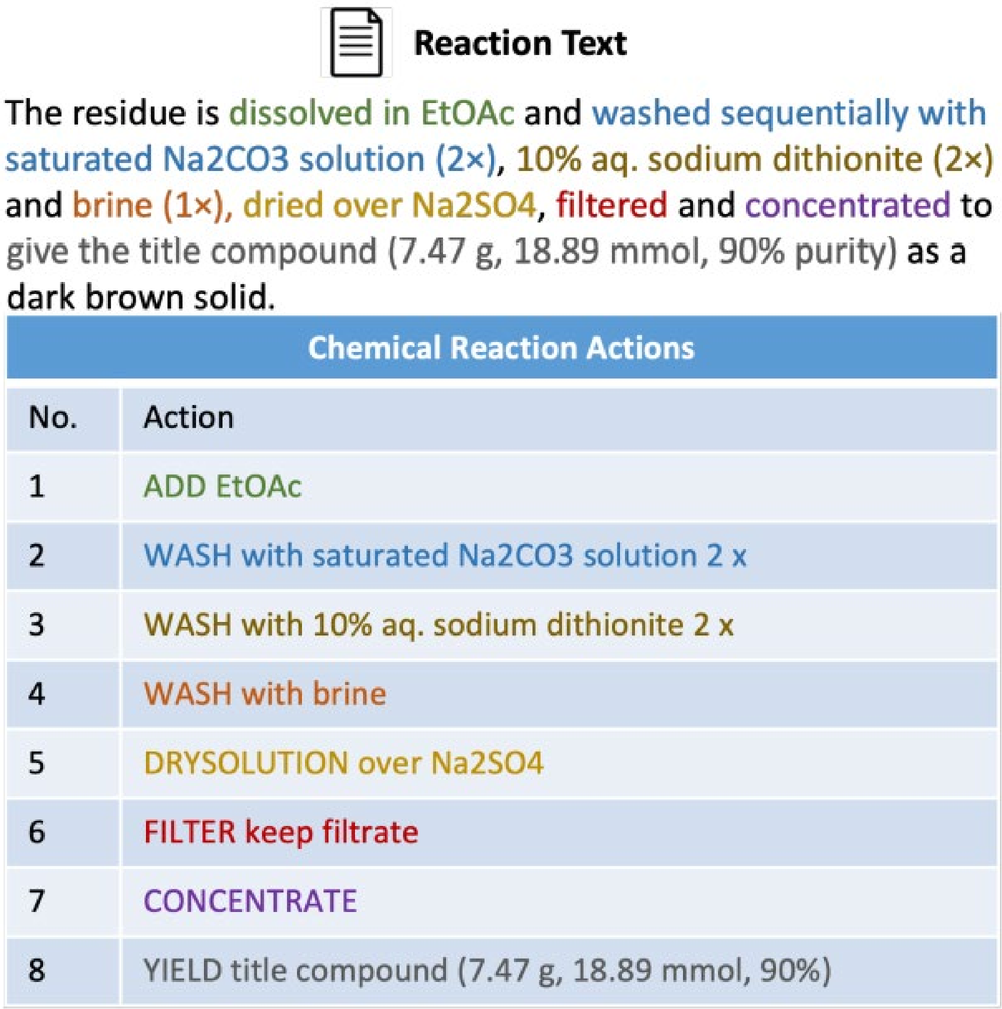
- Chemical reaction extraction (or more generally, (attribute, value) extraction), can be cast as a QA task.
- Linguistic patterns can be used to derive high-quality training data for extraction tasks, but pattern enrichment is needed to boost its recall. Also, patterns are more useful for certain attributes (e.g., temperature, time).

- Limitation:
    - There may be new attributes (e.g., experimental procedures) during inference time. The QA model is not trained on extracting such attributes at all. How to make the model generalizable to new attributes?
    - *Instruct and Extract: Instruction Tuning for On-Demand Information Extraction*. EMNLP 2023.

Agenda

- Fundamental Scientific Information Extraction Tasks
  - Named Entity Recognition: AIONER
  - Relation Extraction: SciER
- Advanced Scientific Information Extraction Tasks
  - Chemical Reaction Extraction: ReactIE
  - Action Extraction: ActionIE

Action Extraction



Reaction Text

The residue is dissolved in EtOAc and washed sequentially with saturated Na2CO3 solution (2×), 10% aq. sodium dithionite (2×) and brine (1×), dried over Na2SO4, filtered and concentrated to give the title compound (7.47 g, 18.89 mmol, 90% purity) as a dark brown solid.

Chemical Reaction Actions

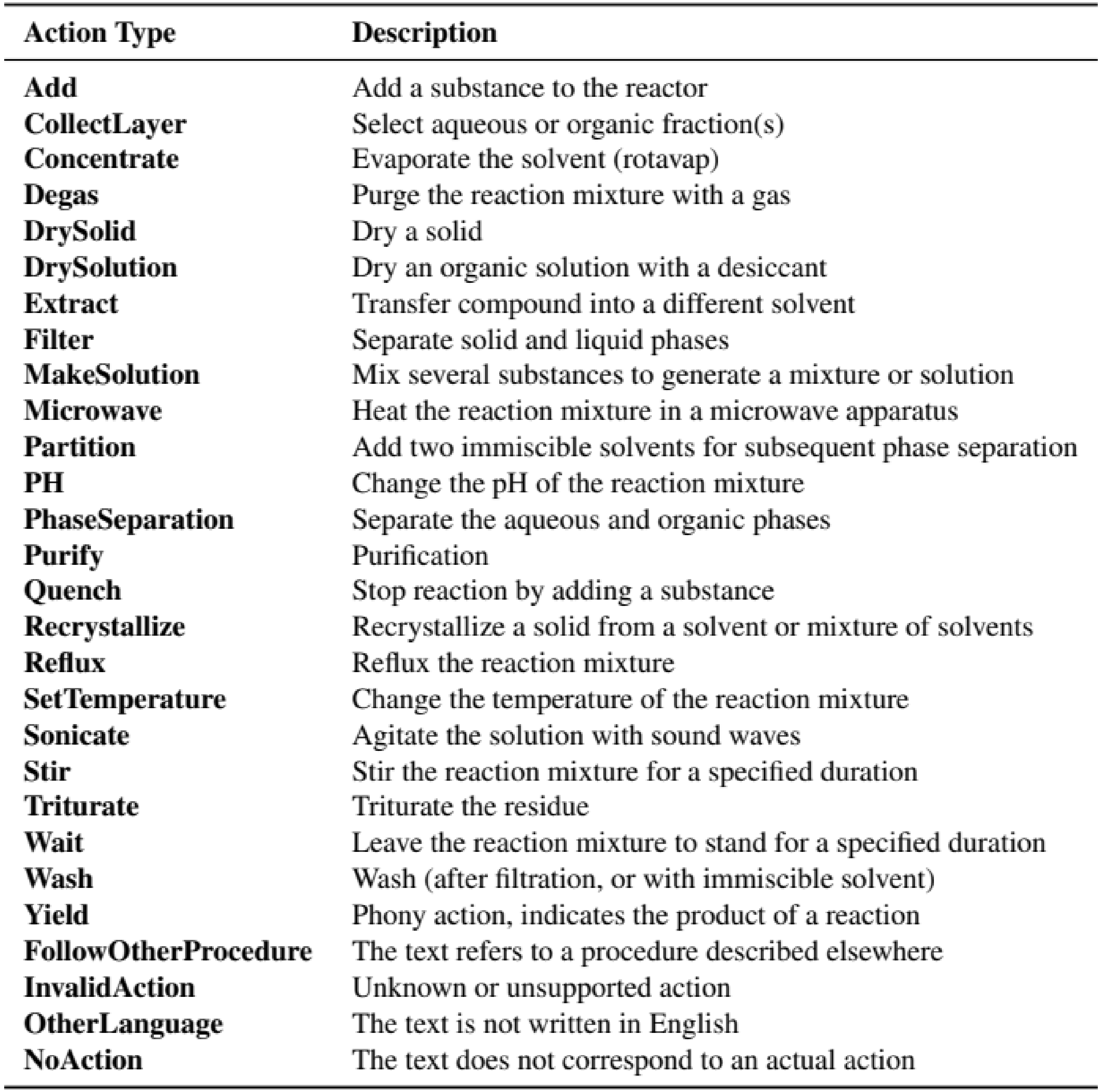
No.	Action
1	ADD EtOAc
2	WASH with saturated Na2CO3 solution 2 x
3	WASH with 10% aq. sodium dithionite 2 x
4	WASH with brine
5	DRYSOLUTION over Na2SO4
6	FILTER keep filtrate
7	CONCENTRATE
8	YIELD title compound (7.47 g, 18.89 mmol, 90%)

- Harder than reaction extraction
  - Need to follow a sequential order
  - The number of attributes in each action varies.
  - Even for the same action type, there may be missing attributes in different cases.
- Bear similarity with programming language!

ActionIE: Action Extraction from Scientific Literature with Programming Languages. ACL 2023 Findings.

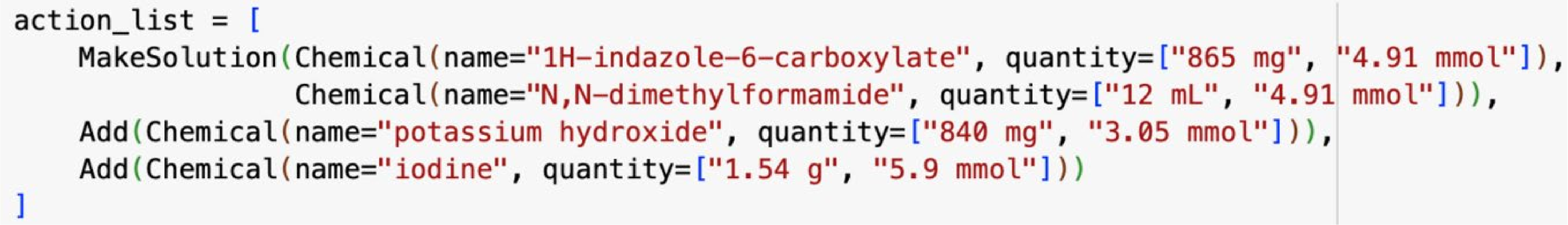
Action Extraction

- The action space is pre-defined.



Action Type	Description
Add	Add a substance to the reactor
CollectLayer	Select aqueous or organic fraction(s)
Concentrate	Evaporate the solvent (rotavap)
Degas	Purge the reaction mixture with a gas
DrySolid	Dry a solid
DrySolution	Dry an organic solution with a desiccant
Extract	Transfer compound into a different solvent
Filter	Separate solid and liquid phases
MakeSolution	Mix several substances to generate a mixture or solution
Microwave	Heat the reaction mixture in a microwave apparatus
Partition	Add two immiscible solvents for subsequent phase separation
PH	Change the pH of the reaction mixture
PhaseSeparation	Separate the aqueous and organic phases
Purify	Purification
Quench	Stop reaction by adding a substance
Recrystallize	Recrystallize a solid from a solvent or mixture of solvents
Reflux	Reflux the reaction mixture
SetTemperature	Change the temperature of the reaction mixture
Sonicate	Agitate the solution with sound waves
Stir	Stir the reaction mixture for a specified duration
Triturate	Triturate the residue
Wait	Leave the reaction mixture to stand for a specified duration
Wash	Wash (after filtration, or with immiscible solvent)
Yield	Phony action, indicates the product of a reaction
FollowOtherProcedure	The text refers to a procedure described elsewhere
InvalidAction	Unknown or unsupported action
OtherLanguage	The text is not written in English
NoAction	The text does not correspond to an actual action

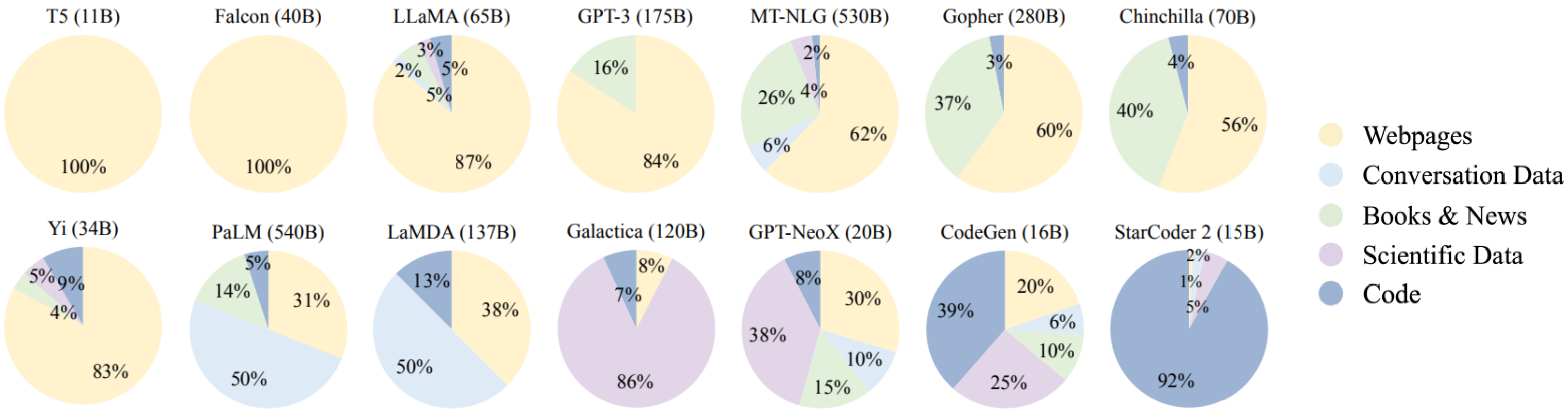
Using Programming Language to Describe Actions



```
action_list = [
    MakeSolution(Chemical(name="1H-indazole-6-carboxylate", quantity=["865 mg", "4.91 mmol"]),
                 Chemical(name="N,N-dimethylformamide", quantity=["12 mL", "4.91 mmol"])),
    Add(Chemical(name="potassium hydroxide", quantity=["840 mg", "3.05 mmol"])),
    Add(Chemical(name="iodine", quantity=["1.54 g", "5.9 mmol"]))
]
```

- A sequence of actions → A sequence of functions
- The order of actions matters. → The order of function calls in your code matters.
- The number of attributes in each action varies. → The number of arguments in each function varies.
- Even for the same action type, there may be missing attributes in different cases. → Even for the same function, there may be missing arguments (i.e., optional/default).

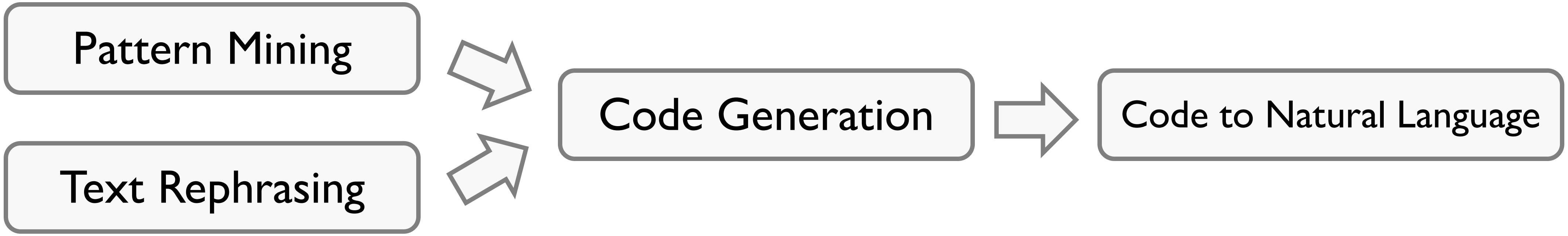
Why consider programming language/code?



- Many LLMs are pre-trained on massive code data, so they are powerful in code completion.
- Code completion has demonstrated its powerful in event structure prediction [1].

Code4Struct: Code Generation for Few-Shot Event Structure Prediction. ACL 2023.

The ActionIE Framework

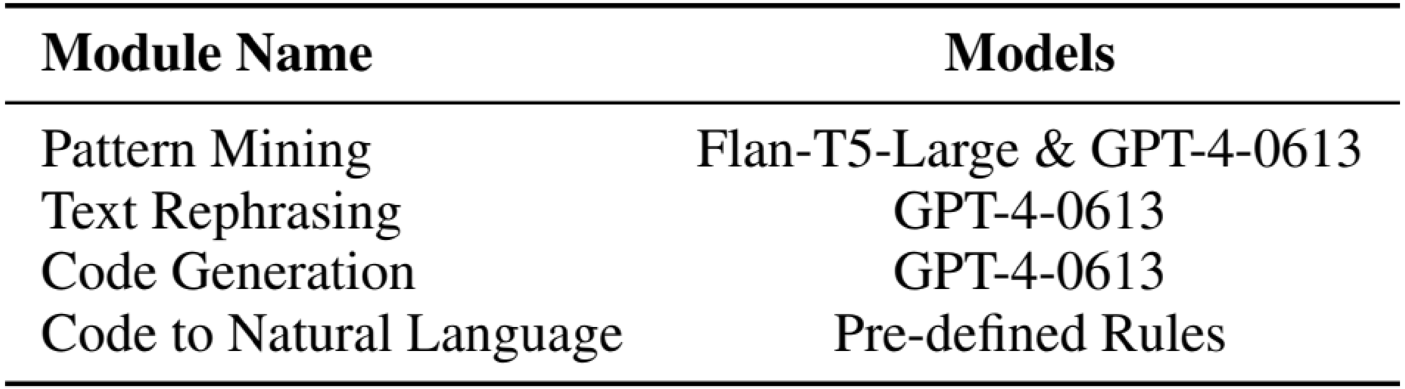


Pattern Mining

Text Rephrasing

Code Generation

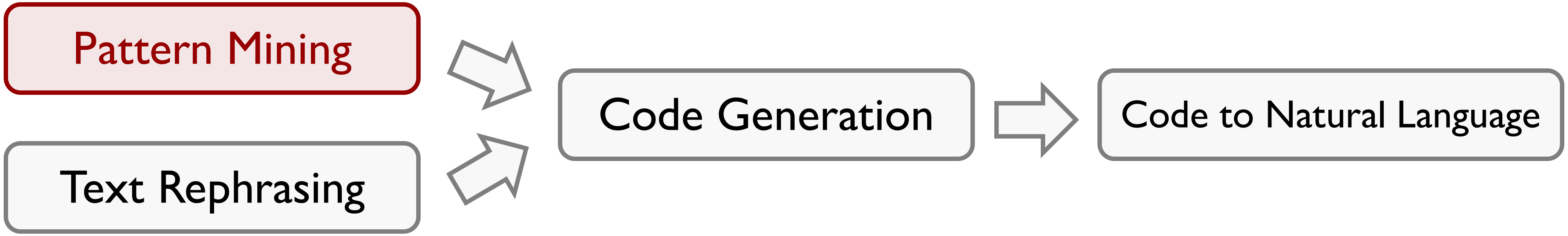
Code to Natural Language



Module Name	Models
Pattern Mining	Flan-T5-Large & GPT-4-0613
Text Rephrasing	GPT-4-0613
Code Generation	GPT-4-0613
Code to Natural Language	Pre-defined Rules

ActionIE: Action Extraction from Scientific Literature with Programming Languages. ACL 2023 Findings.

The ActionIE Framework



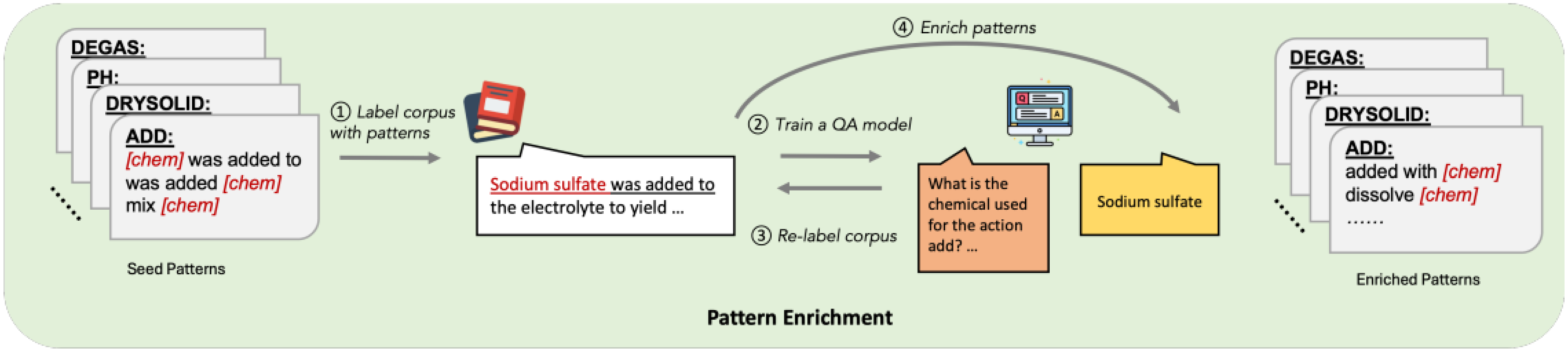
Pattern Mining

Text Rephrasing

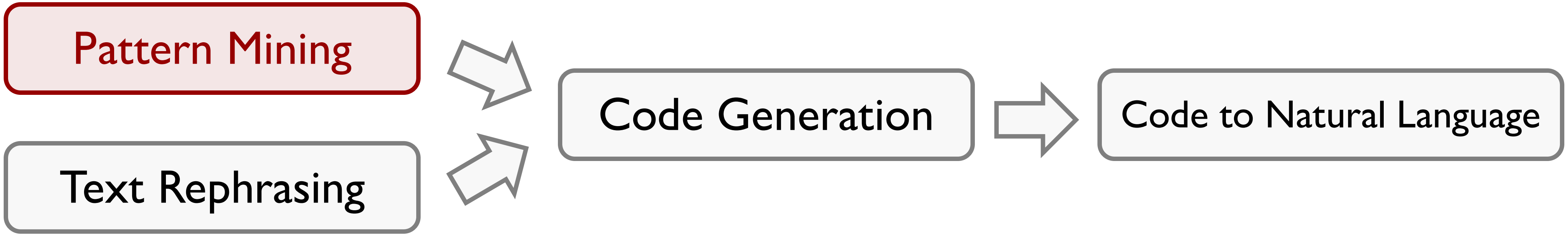
Code Generation

Code to Natural Language

- Similar to ReactIE, but the patterns here are describing actions



The ActionIE Framework



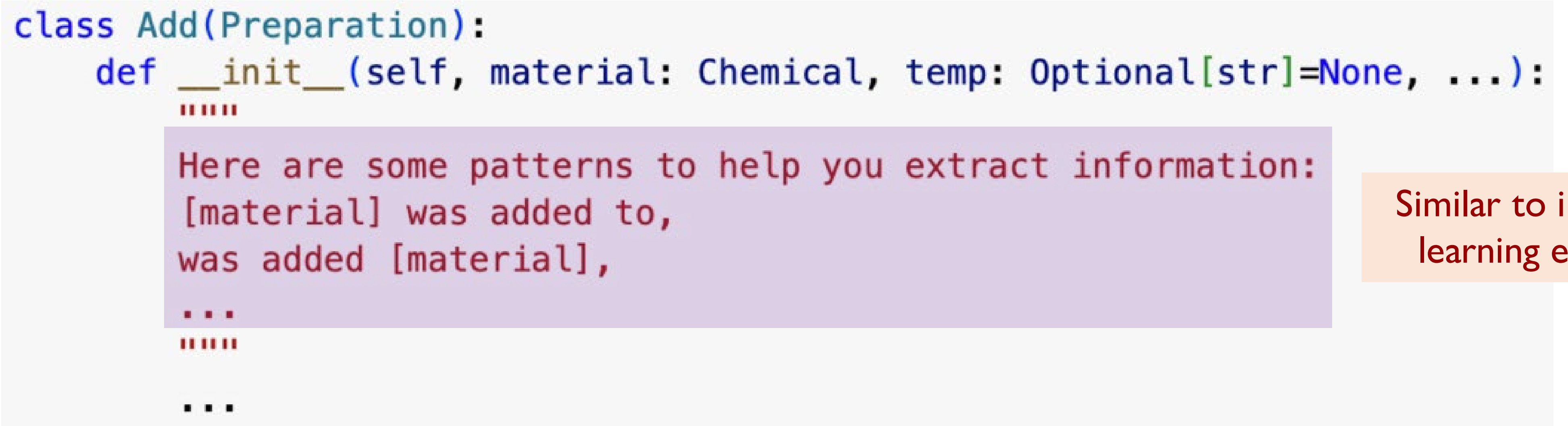
Pattern Mining

Text Rephrasing

Code Generation

Code to Natural Language

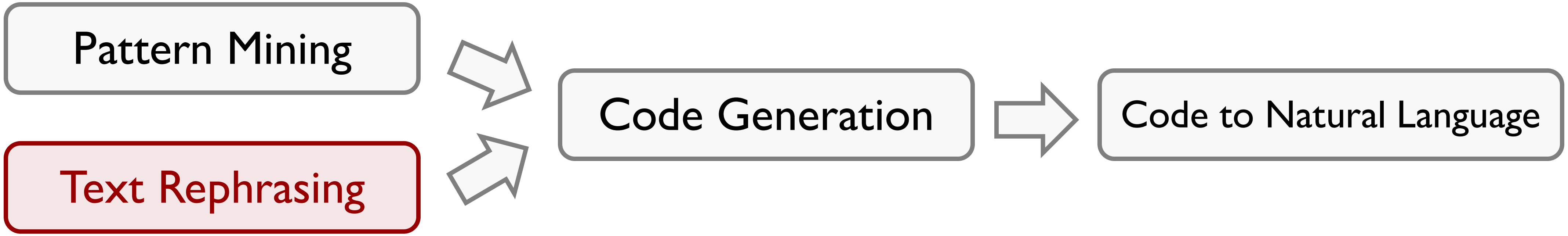
- The extracted patterns will be added into a function template as input to an LLM.



```
class Add(Preparation):
    def __init__(self, material: Chemical, temp: Optional[str]=None, ...):
        """
        Here are some patterns to help you extract information:
        [material] was added to,
        was added [material],
        ...
        """
        ...
```

Similar to in-context learning examples

The ActionIE Framework

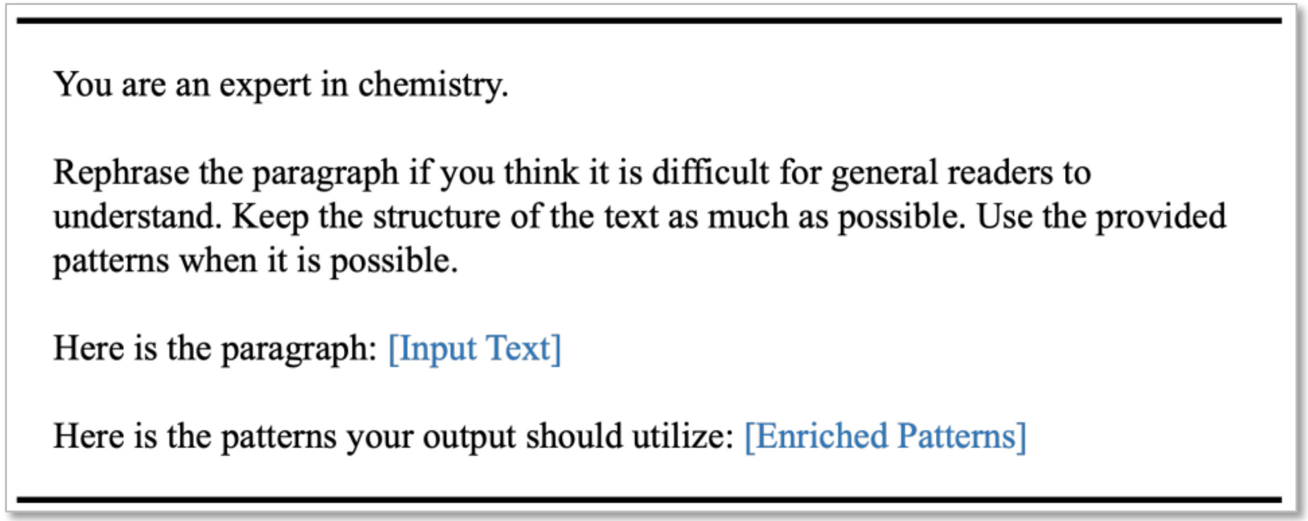


Pattern Mining

Text Rephrasing

Code Generation

Code to Natural Language



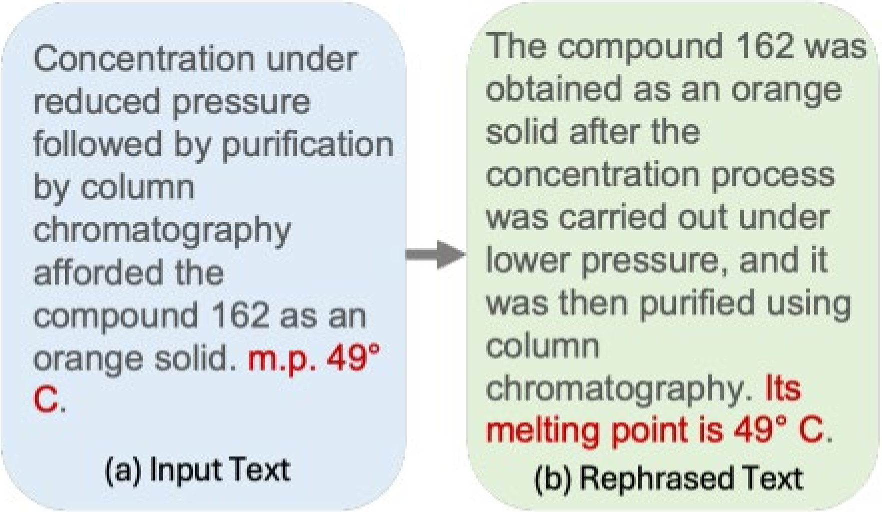
You are an expert in chemistry.

Rephrase the paragraph if you think it is difficult for general readers to understand. Keep the structure of the text as much as possible. Use the provided patterns when it is possible.

Here is the paragraph: [Input Text]

Here is the patterns your output should utilize: [Enriched Patterns]

Instruction



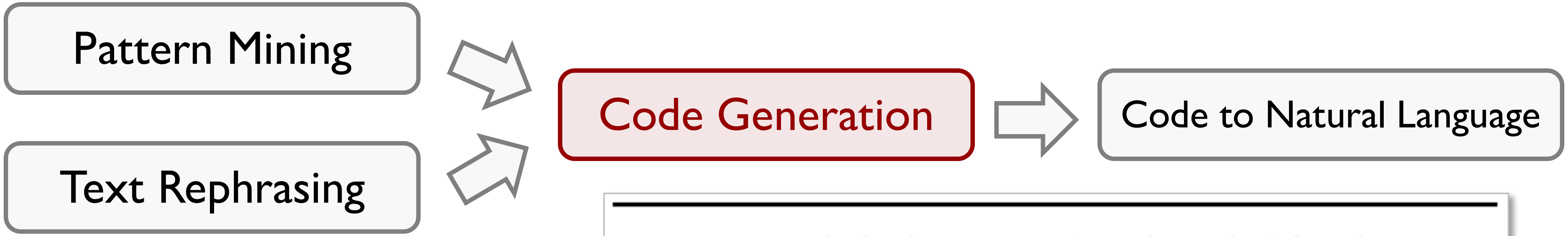
Concentration under reduced pressure followed by purification by column chromatography afforded the compound 162 as an orange solid. m.p. 49° C.

(a) Input Text

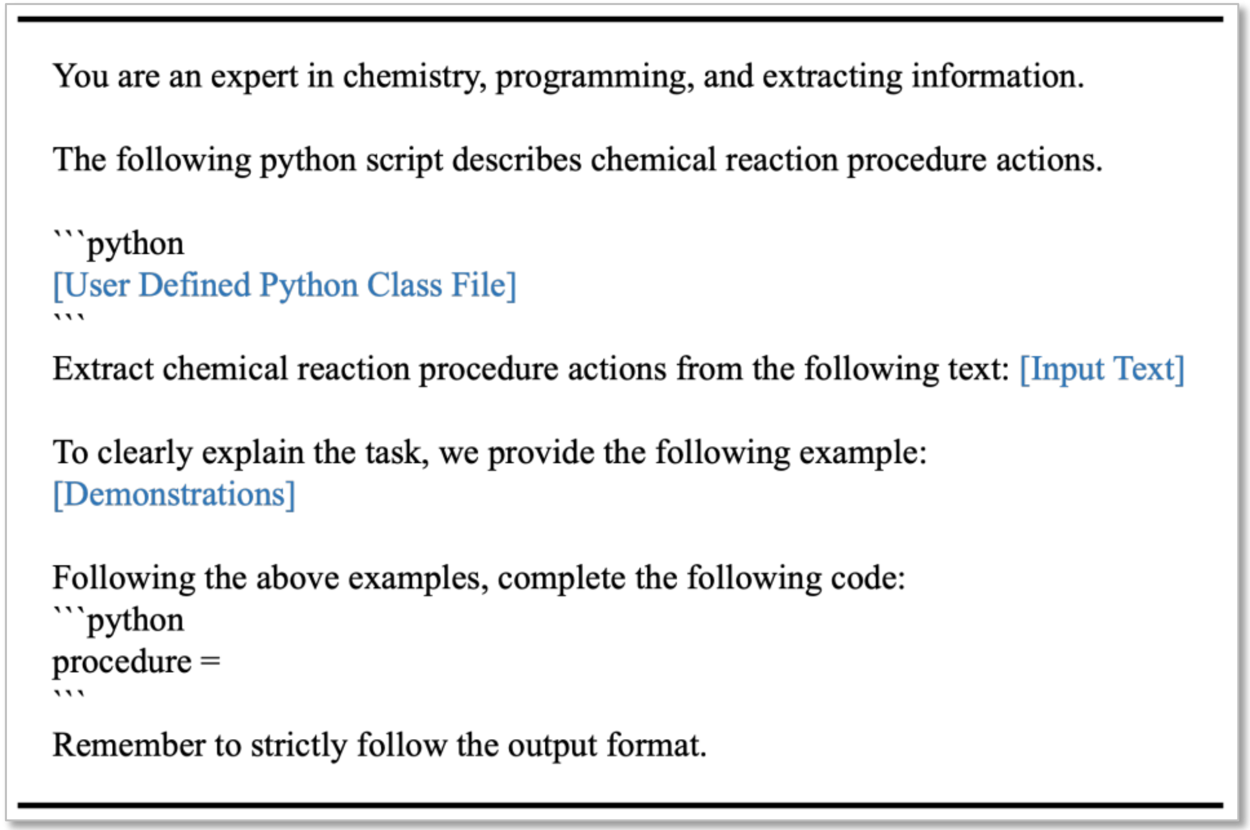
The compound 162 was obtained as an orange solid after the concentration process was carried out under lower pressure, and it was then purified using column chromatography. Its melting point is 49° C.

(b) Rephrased Text

The ActionIE Framework



Instruction



~~~~
You are an expert in chemistry, programming, and extracting information.

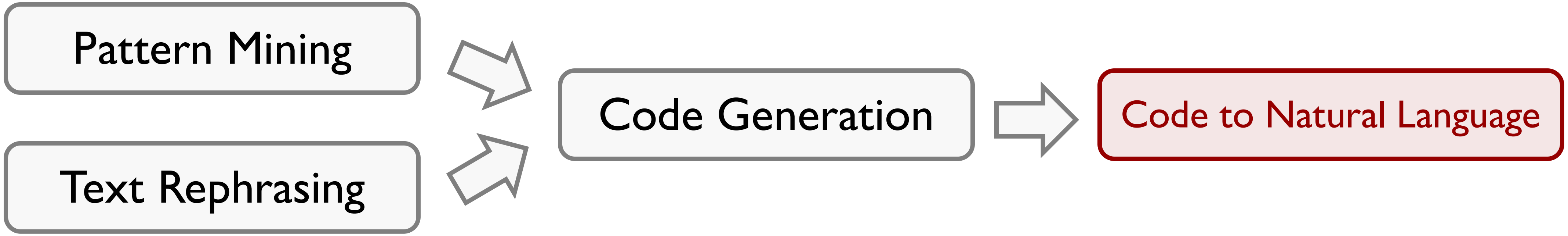
The following python script describes chemical reaction procedure actions.

```python
[User Defined Python Class File]
```
Extract chemical reaction procedure actions from the following text: [Input Text]

To clearly explain the task, we provide the following example:
[Demonstrations]

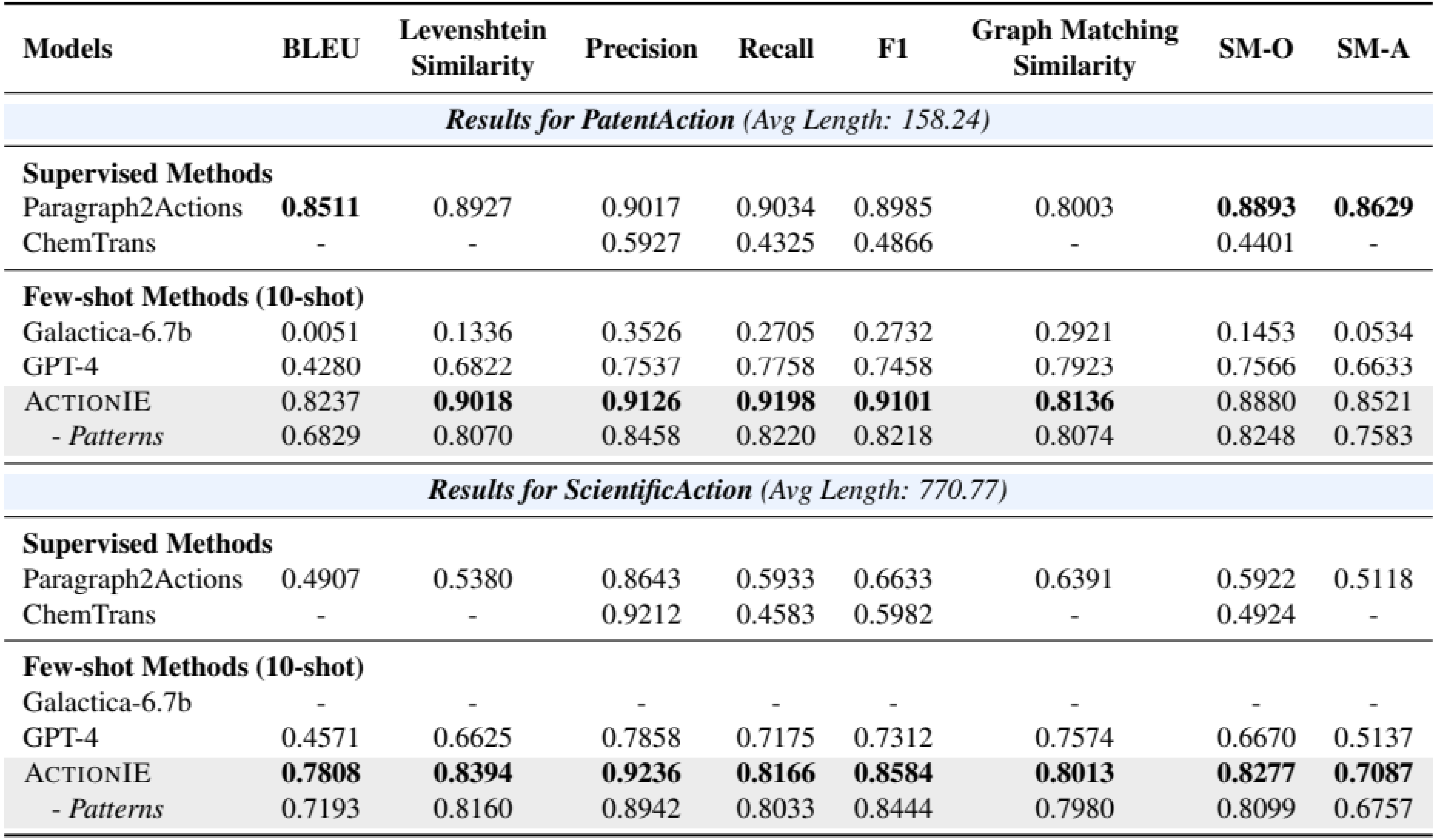
Following the above examples, complete the following code:
```python
procedure =
```
Remember to strictly follow the output format.
~~~~

The ActionIE Framework



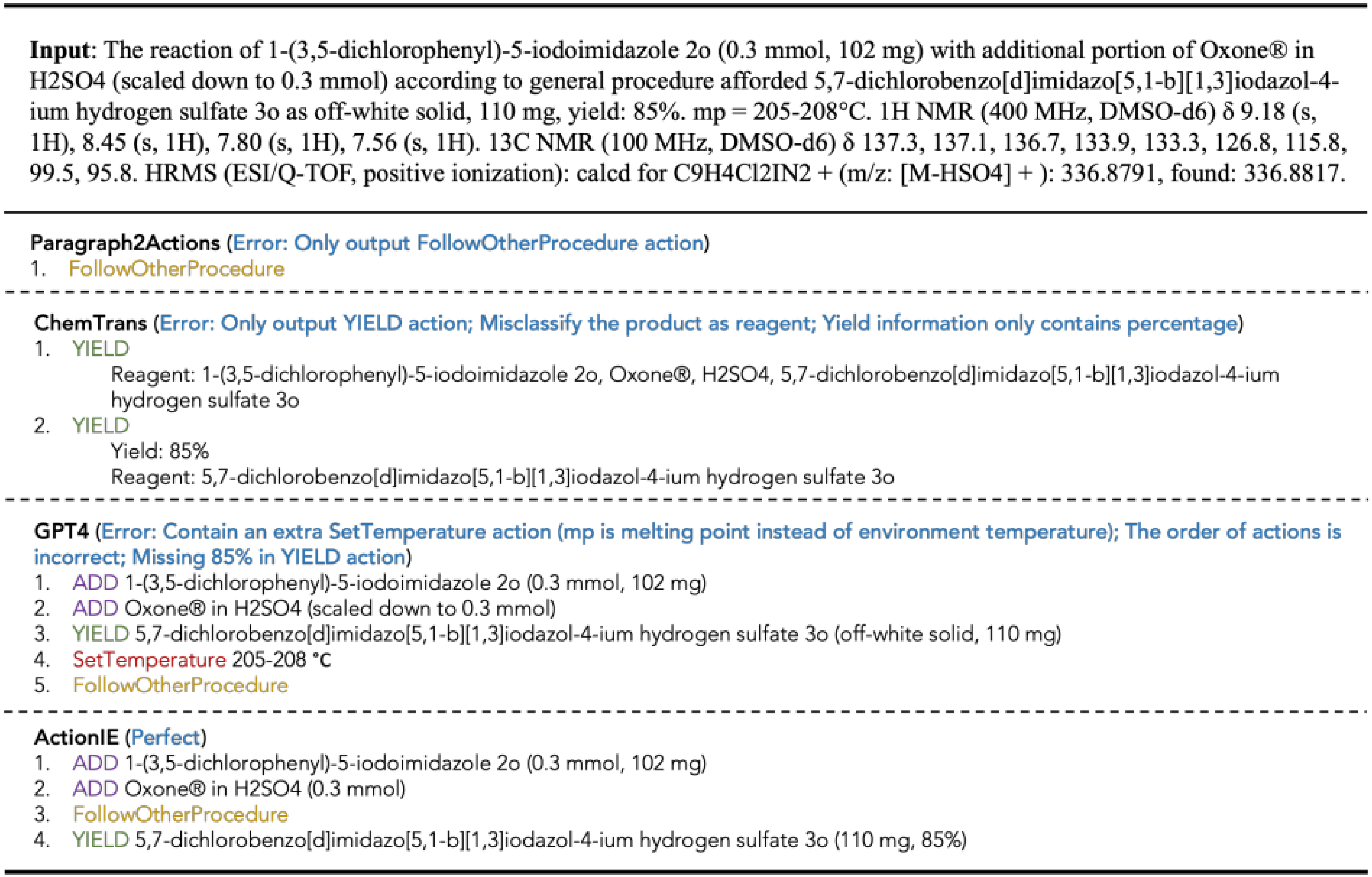
- Because the action space is pre-defined, just use action-specific templates to convert functions to natural language.
  - Function: Add(Chemical(name=“iodine”, quantity=[“1.54 g”, “5.9 mmol”]))
  - ADD template: ADD name (quantity[0], quantity[1])
  - Natural language: ADD iodine (1.54 g, 5.9 mmol)

Performance of ActionIE



Models	BLEU	Levenshtein Similarity	Precision	Recall	F1	Graph Matching Similarity	SM-O	SM-A
Results for PatentAction *(Avg Length: 158.24)*								
Supervised Methods								
Paragraph2Actions	**0.8511**	0.8927	0.9017	0.9034	0.8985	0.8003	**0.8893**	**0.8629**
ChemTrans	-	-	0.5927	0.4325	0.4866	-	0.4401	-
Few-shot Methods (10-shot)								
Galactica-6.7b	0.0051	0.1336	0.3526	0.2705	0.2732	0.2921	0.1453	0.0534
GPT-4	0.4280	0.6822	0.7537	0.7758	0.7458	0.7923	0.7566	0.6633
ActionIE	0.8237	**0.9018**	**0.9126**	**0.9198**	**0.9101**	**0.8136**	0.8880	0.8521
- *Patterns*	0.6829	0.8070	0.8458	0.8220	0.8218	0.8074	0.8248	0.7583
Results for ScientificAction *(Avg Length: 770.77)*								
Supervised Methods								
Paragraph2Actions	0.4907	0.5380	0.8643	0.5933	0.6633	0.6391	0.5922	0.5118
ChemTrans	-	-	0.9212	0.4583	0.5982	-	0.4924	-
Few-shot Methods (10-shot)								
Galactica-6.7b	-	-	-	-	-	-	-	-
GPT-4	0.4571	0.6625	0.7858	0.7175	0.7312	0.7574	0.6670	0.5137
ActionIE	**0.7808**	**0.8394**	**0.9236**	**0.8166**	**0.8584**	**0.8013**	**0.8277**	**0.7087**
- *Patterns*	0.7193	0.8160	0.8942	0.8033	0.8444	0.7980	0.8099	0.6757

Case Study



Input: The reaction of 1-(3,5-dichlorophenyl)-5-iodoimidazole 2o (0.3 mmol, 102 mg) with additional portion of Oxone® in H2SO4 (scaled down to 0.3 mmol) according to general procedure afforded 5,7-dichlorobenzo[d]imidazo[5,1-b][1,3]iodazol-4-ium hydrogen sulfate 3o as off-white solid, 110 mg, yield: 85%. mp = 205-208°C. 1H NMR (400 MHz, DMSO-d6) δ 9.18 (s, 1H), 8.45 (s, 1H), 7.80 (s, 1H), 7.56 (s, 1H). 13C NMR (100 MHz, DMSO-d6) δ 137.3, 137.1, 136.7, 133.9, 133.3, 126.8, 115.8, 99.5, 95.8. HRMS (ESI/Q-TOF, positive ionization): calcd for C9H4Cl2IN2 + (m/z: [M-HSO4] +): 336.8791, found: 336.8817.

Paragraph2Actions (Error: Only output FollowOtherProcedure action)

1. FollowOtherProcedure

ChemTrans (Error: Only output YIELD action; Misclassify the product as reagent; Yield information only contains percentage)

1. YIELD
   Reagent: 1-(3,5-dichlorophenyl)-5-iodoimidazole 2o, Oxone®, H2SO4, 5,7-dichlorobenzo[d]imidazo[5,1-b][1,3]iodazol-4-ium hydrogen sulfate 3o
2. YIELD
   Yield: 85%
   Reagent: 5,7-dichlorobenzo[d]imidazo[5,1-b][1,3]iodazol-4-ium hydrogen sulfate 3o

GPT4 (Error: Contain an extra SetTemperature action (mp is melting point instead of environment temperature); The order of actions is incorrect; Missing 85% in YIELD action)

1. ADD 1-(3,5-dichlorophenyl)-5-iodoimidazole 2o (0.3 mmol, 102 mg)
2. ADD Oxone® in H2SO4 (scaled down to 0.3 mmol)
3. YIELD 5,7-dichlorobenzo[d]imidazo[5,1-b][1,3]iodazol-4-ium hydrogen sulfate 3o (off-white solid, 110 mg)
4. SetTemperature 205-208 °C
5. FollowOtherProcedure

ActionIE (Perfect)

1. ADD 1-(3,5-dichlorophenyl)-5-iodoimidazole 2o (0.3 mmol, 102 mg)
2. ADD Oxone® in H2SO4 (0.3 mmol)
3. FollowOtherProcedure
4. YIELD 5,7-dichlorobenzo[d]imidazo[5,1-b][1,3]iodazol-4-ium hydrogen sulfate 3o (110 mg, 85%)

Take-Away Messages

- Programming language helps complex information extraction tasks (e.g., action extraction) because:
  - Many LLMs are pre-trained on massive code data
  - Function templates guide the structure of LLM generation

- Limitations
  - No strategies to handle missing values
    - In a function call, if an argument is not specified, the default value will be used.
    - In action extraction, the default value may or may not be global.
      - Common practice vs. details already specified in previous actions
  - No experiments of handling very long experiments (e.g., a full-text article describing total synthesis of a certain natural product)

Thank You!

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html