Exploring the versatility of LLMs for Relation extraction in underexplored biomedical areas

Maxime Delmas - 05/03/2024

idiap RESEARCH INSTITUTE

INFLAMM ALPS

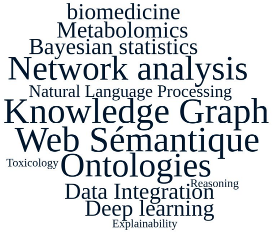
Background

PhD (2019 - 2022)

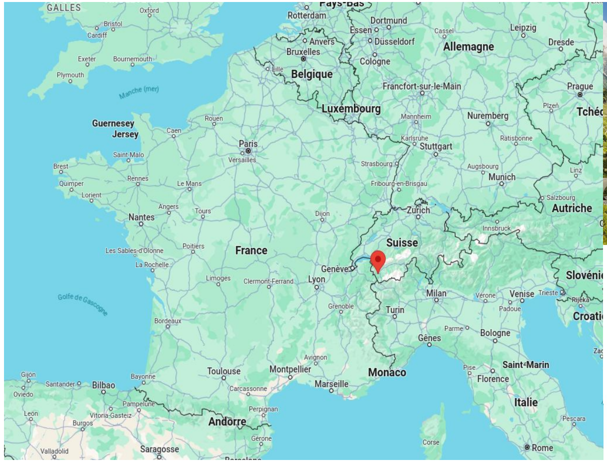
ToXalim
RESEARCH CENTRE IN FOOD TOXICOLOGY

INRAE

FORVM



Post doctoral position (2022 - now)

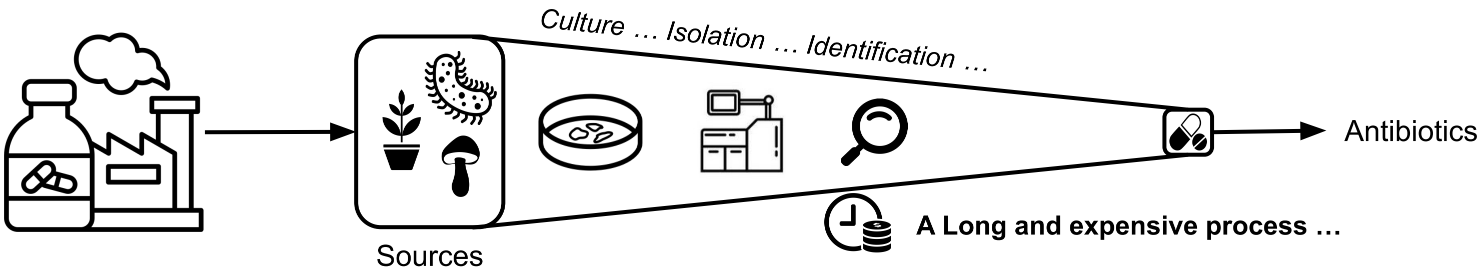


idiap
RESEARCH INSTITUTE

Neuro-Symbolic AI

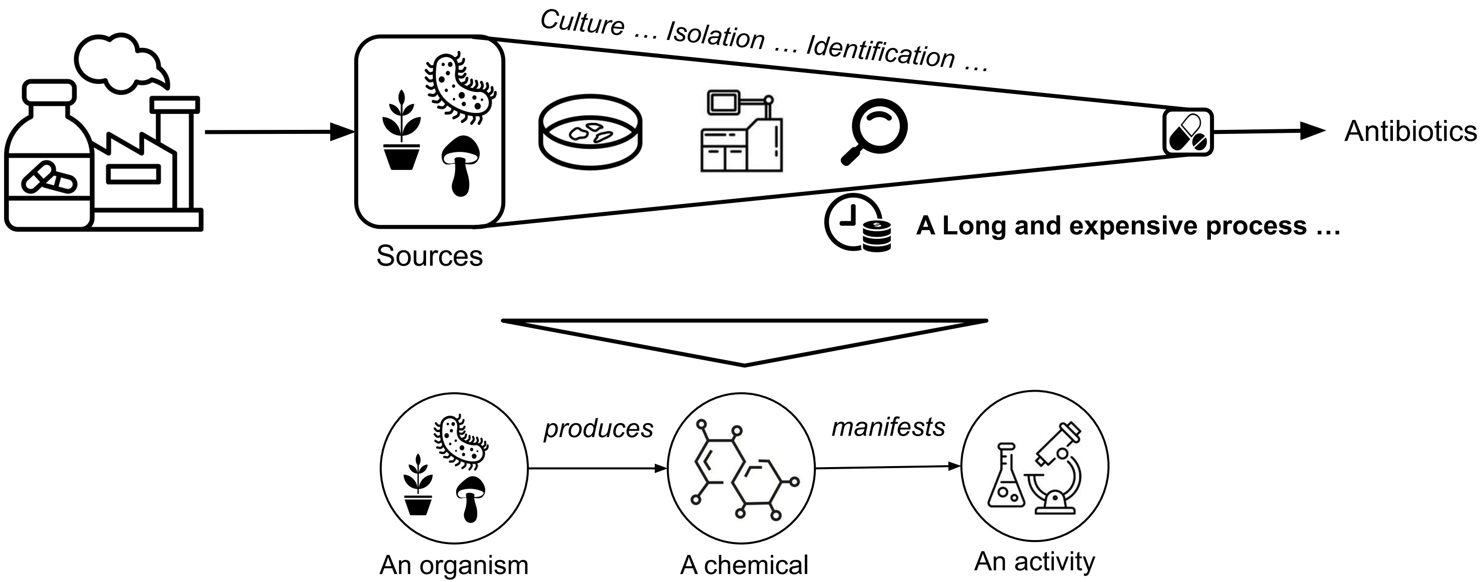
INFLAMM ALPS

Study context: avoiding rediscoveries

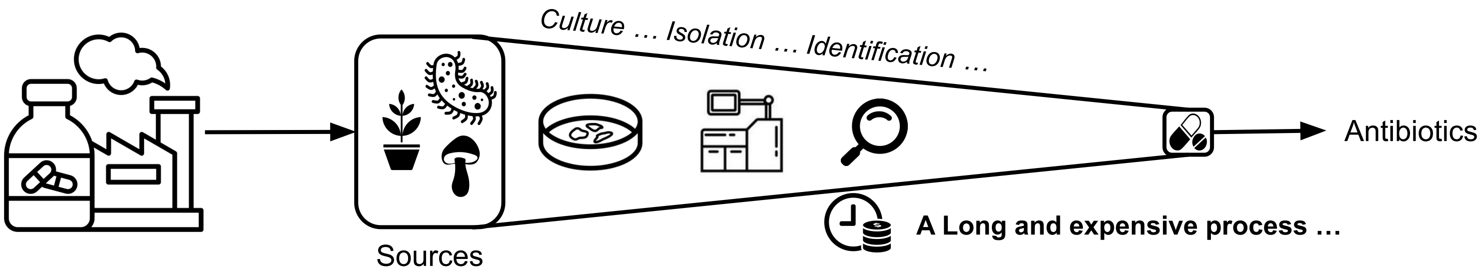


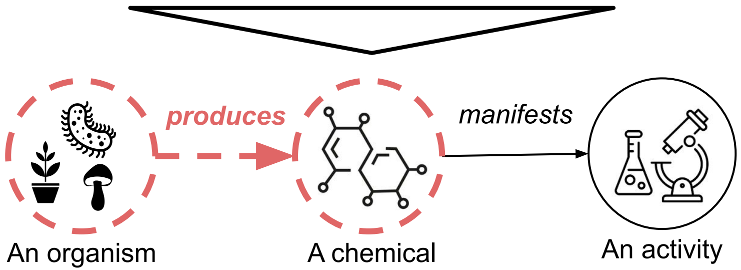
INFLAMALPS

Study context: avoiding rediscoveries

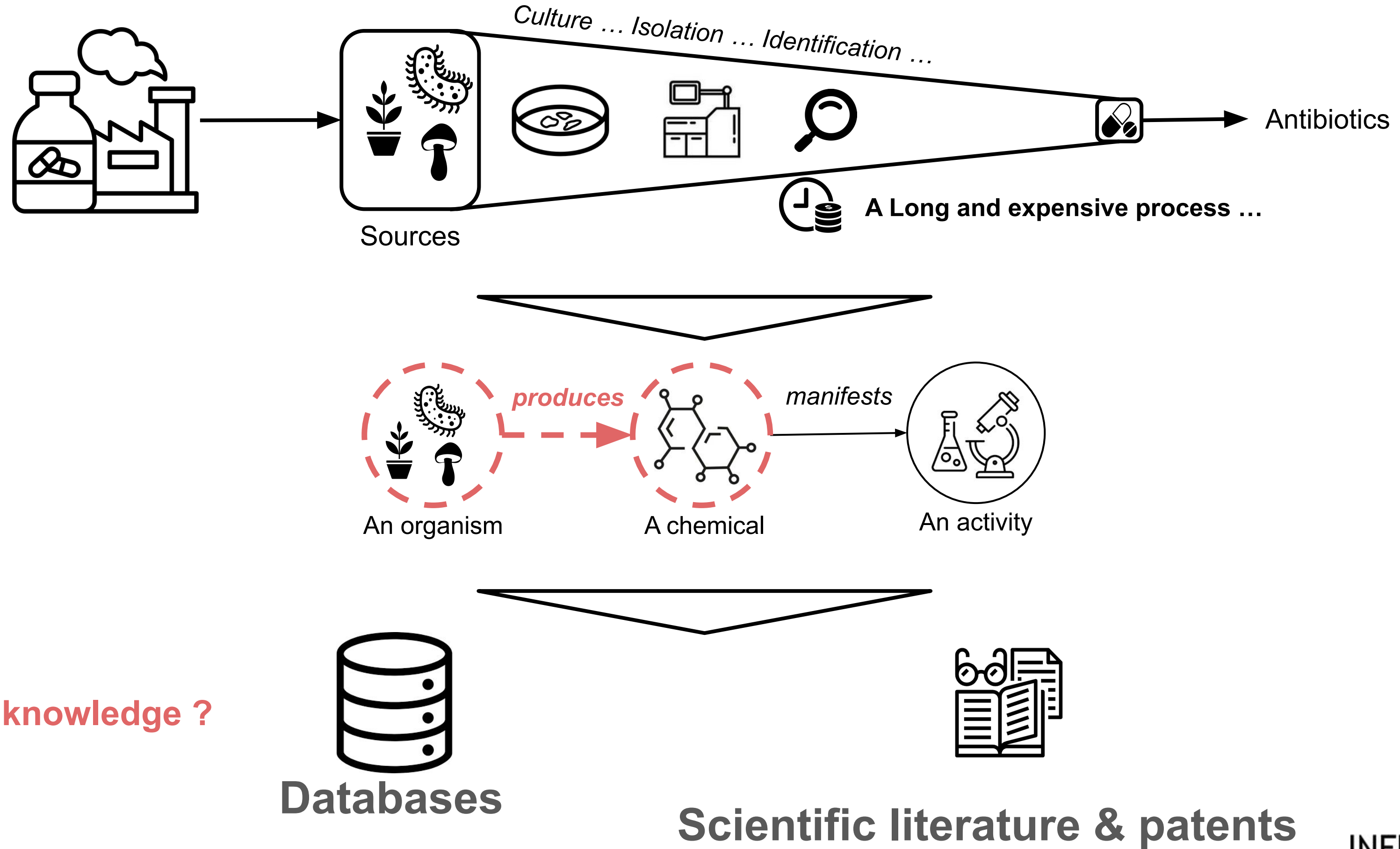


Study context: avoiding rediscoveries



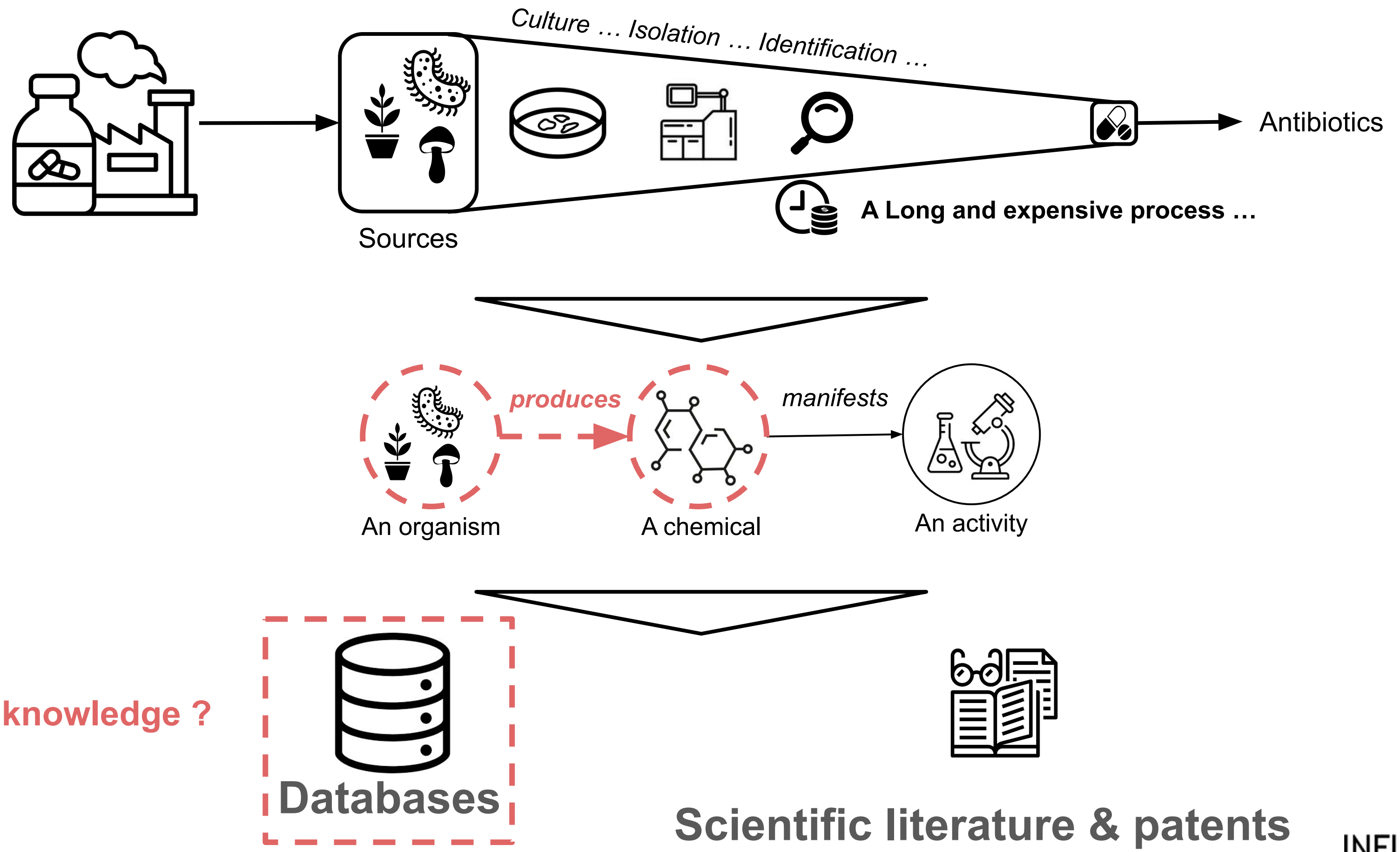


Study context: avoiding rediscoveries



Where is this knowledge ?

Study context: avoiding rediscoveries



Where is this knowledge ?

INFLAMMALPS

LOTUS: An Open Knowledge Base for natural products

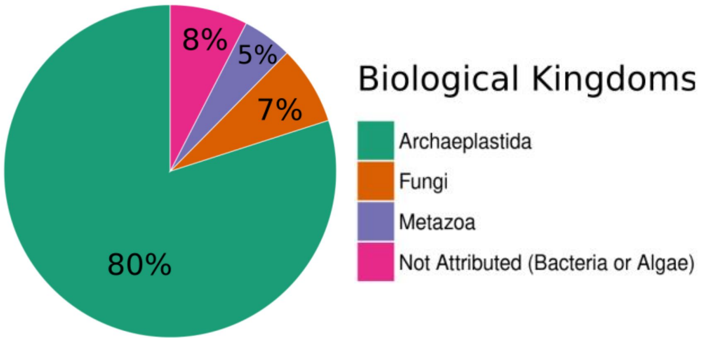
LOTUS

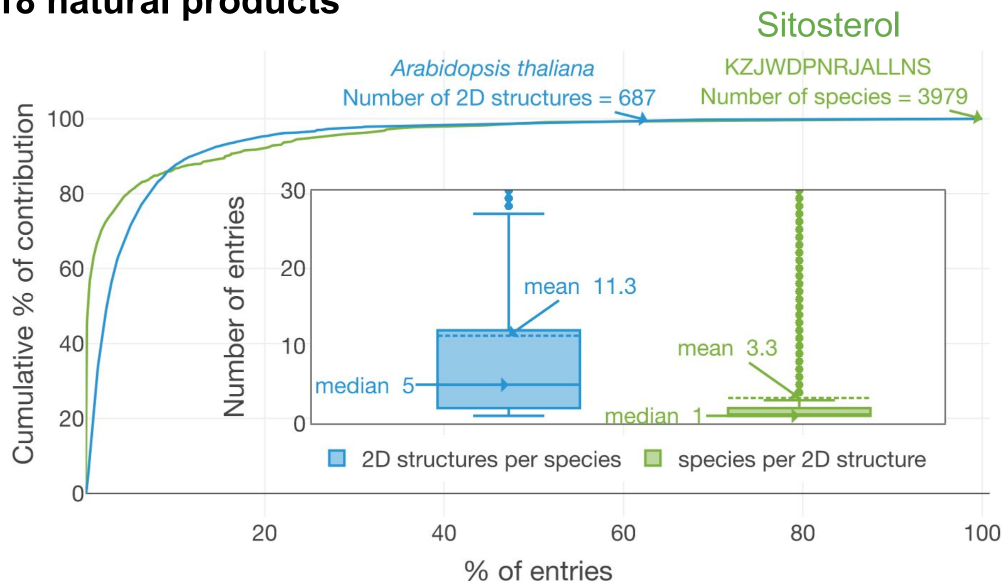
Data on > 276 000 natural products

A. Rutz et al., 'The LOTUS initiative for open knowledge management in natural products research', eLife, vol. 11, p. e70780, May 2022, doi: 10.7554/eLife.70780.

LOTUS: An Open Knowledge Base for natural products

Data on > 276,518 natural products

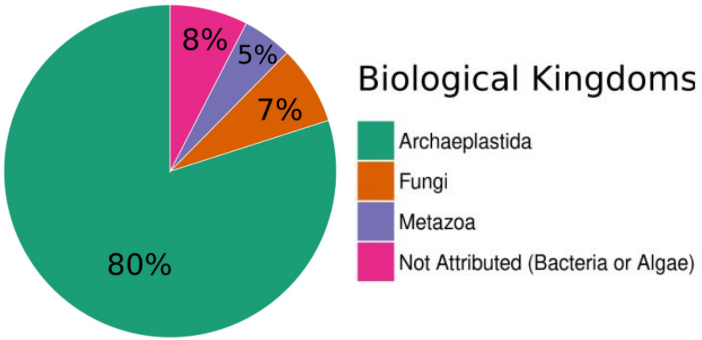


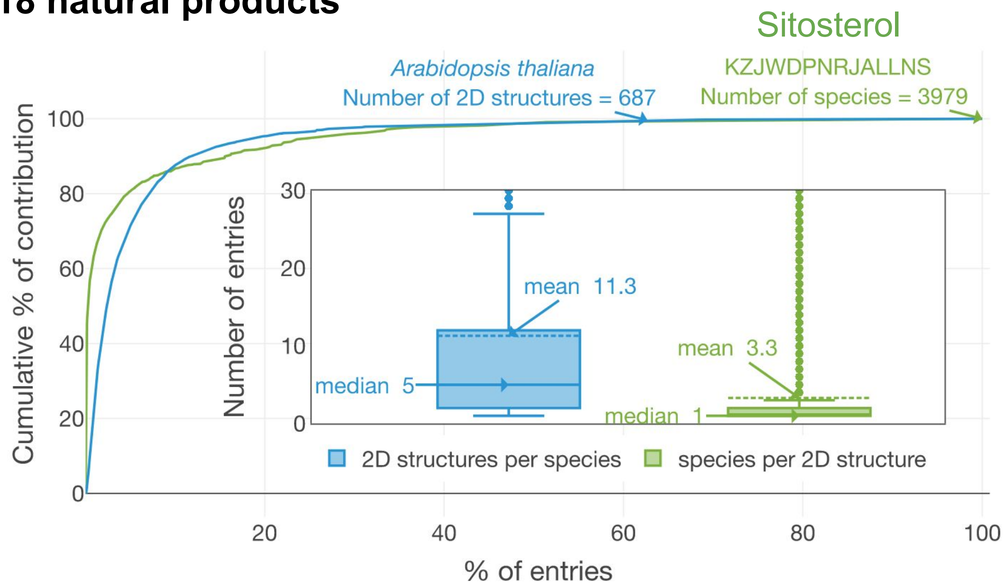


A. Rutz et al., 'The LOTUS initiative for open knowledge management in natural products research', eLife, vol. 11, p. e70780, May 2022, doi: 10.7554/eLife.70780.

LOTUS: An Open Knowledge Base for natural products

Data on > 276,518 natural products





Imbalance toward *model organisms, few data for more exotic* organisms

A. Rutz et al., 'The LOTUS initiative for open knowledge management in natural products research', eLife, vol. 11, p. e70780, May 2022, doi: 10.7554/eLife.70780.

Automatic extraction of relations from the literature

~~Databases~~ → **Scientific literature & patents**

Abstract

Ochratoxin A (OTA) is a very important mycotoxin, and its research is focused right now on the new findings of OTA, like being a complete carcinogen, information about OTA producers and new exposure sources of OTA. Citrinin (CIT) is another important mycotoxin, too, and its research turns towards nephrotoxicity. Both additive and synergistic effects have been described in combination with OTA. OTA is produced in foodstuffs by Aspergillus Section Circumdati (Aspergillus ochraceus, A. westerdijkiae, A. steynii) and Aspergillus Section Nigri (Aspergillus carbonarius, A. foetidus, A. lacticoffeatus, A. niger, A. sclerotioniger, A. tubingensis), mostly in subtropical and tropical areas. OTA is produced in foodstuffs by Penicillium verrucosum and P. nordicum, notably in temperate and colder zones. CIT is produced in foodstuffs by Monascus species (Monascus purpureus, M. ruber) and Penicillium species (Penicillium citrinum, P. expansum, P. radicicola, P. verrucosum). OTA was frequently found in foodstuffs of both plant origin (e.g., cereal products, coffee, vegetable, liquorice, raisins, wine) and animal origin (e.g., pork/poultry). CIT was also found in foodstuffs of vegetable origin (e.g., cereals, pomaceous fruits, black olive, roasted nuts, spices), food supplements based on rice fermented with red microfungi Monascus purpureus and in foodstuffs of animal origin (e.g., cheese).

Named Entity Recognition for Bio-entities

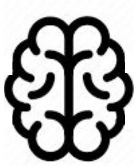
Extracted relations

Aspergillus ochraceus - Ochratoxin A
Aspergillus westerdijkiae - Ochratoxin A
Aspergillus steynii - Ochratoxin A
…
Monascus purpureus - Citrinin
Penicillium expansum - Citrinin

End-to-end NER / RE

How to build a Relation Extraction model - recipe

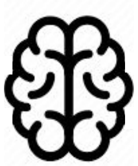
In the era of machine learning, efficient models are ***supervised***



the architecture
(the weights)

How to build a Relation Extraction model - recipe

In the era of machine learning, efficient models are ***supervised***



the architecture
(the weights)

+

Ochratoxin A (OTA) is a very important mycotoxin, and its research is focused right now on the new findings of OTA, like being a complete carcinogen, information about OTA producers and new exposure sources of OTA. Citrinin (CIT) is another important mycotoxin, too, and its research turns towards nephrotoxicity. Both additive and synergistic effects have been described in combination with OTA. OTA is produced in foodstuffs by Aspergillus Section Circumdati (Aspergillus ochraceus, A. westerdijkiae, A. steynii) and Aspergillus Section Nigri (Aspergillus carbonarius, A. foetidus, A. lacticoffeatus, A. niger, A. sclerotioniger, A. tubingensis), mostly in subtropical and tropical areas. OTA is produced in foodstuffs by Penicillium verrucosum and P. nordicum, notably in temperate and colder zones. CIT is produced in foodstuffs by Monascus species (Monascus purpureus, M. ruber) and Penicillium species (Penicillium citrinum, P. expansum, P. radicicola, P. verrucosum). OTA was frequently found in foodstuffs of both plant origin (e.g., cereal products, coffee, vegetable, liquorice, raisins, wine) and animal origin (e.g., pork/poultry). CIT was also found in foodstuffs of vegetable origin (e.g., cereals, pomaceous fruits, black olive, roasted nuts, spices), food supplements based on rice fermented with red microfungi Monascus purpureus and in foodstuffs of animal origin (e.g., cheese).

+

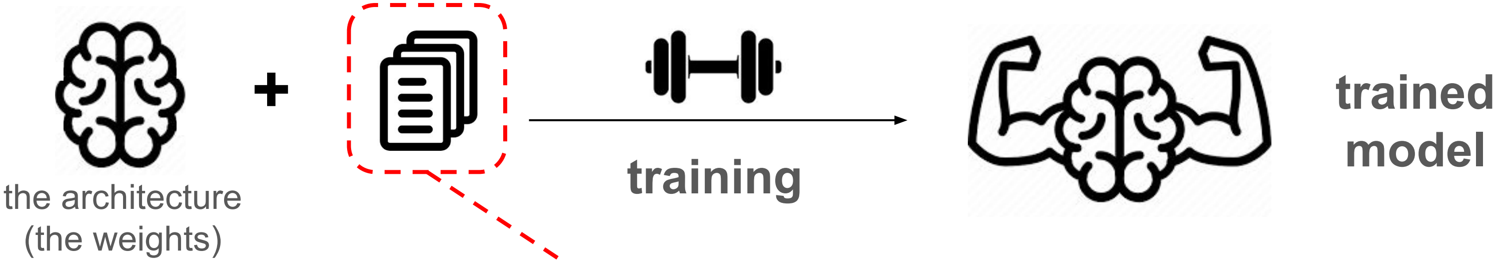
The data 🙏

Expected output labels

Aspergillus ochraceus - Ochratoxin A
Aspergillus westerdijkiae - Ochratoxin A
Aspergillus steynii - Ochratoxin A
…
Monascus purpureus - Citrinin
Penicillium expansum - Citrinin

How to build a Relation Extraction model - recipe

In the era of machine learning, efficient models are ***supervised***



Ochratoxin A (OTA) is a very important mycotoxin, and its research is focused right now on the new findings of OTA, like being a complete carcinogen, information about OTA producers and new exposure sources of OTA. Citrinin (CIT) is another important mycotoxin, too, and its research turns towards nephrotoxicity. Both additive and synergistic effects have been described in combination with OTA. OTA is produced in foodstuffs by Aspergillus Section Circumdati (Aspergillus ochraceus, A. westerdijkiae, A. steynii) and Aspergillus Section Nigri (Aspergillus carbonarius, A. foetidus, A. lacticoffeatus, A. niger, A. sclerotioniger, A. tubingensis), mostly in subtropical and tropical areas. OTA is produced in foodstuffs by Penicillium verrucosum and P. nordicum, notably in temperate and colder zones. CIT is produced in foodstuffs by Monascus species (Monascus purpureus, M. ruber) and Penicillium species (Penicillium citrinum, P. expansum, P. radicicola, P. verrucosum). OTA was frequently found in foodstuffs of both plant origin (e.g., cereal products, coffee, vegetable, liquorice, raisins, wine) and animal origin (e.g., pork/poultry). CIT was also found in foodstuffs of vegetable origin (e.g., cereals, pomaceous fruits, black olive, roasted nuts, spices), food supplements based on rice fermented with red microfungi Monascus purpureus and in foodstuffs of animal origin (e.g., cheese).

+

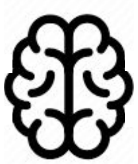
The data 🙏

Expected output labels

Aspergillus ochraceus - Ochratoxin A
Aspergillus westerdijkiae - Ochratoxin A
Aspergillus steynii - Ochratoxin A
…
Monascus purpureus - Citrinin
Penicillium expansum - Citrinin

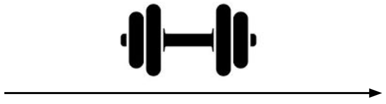
How to build a Relation Extraction model - recipe

In the era of machine learning, efficient models are ***supervised***

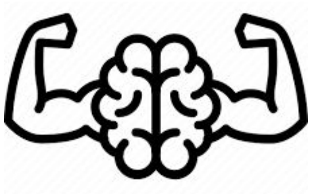


the architecture
(the weights)

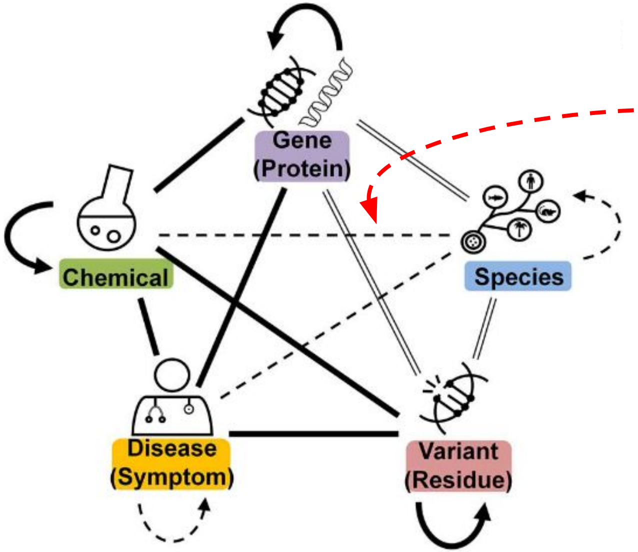
+



training



trained model



Ochratoxin A (OTA) is a very important mycotoxin, and its research is focused right now on the new findings of OTA, like being a complete carcinogen, information about OTA producers and new exposure sources of OTA. Citrinin (CIT) is another important mycotoxin, too, and its research turns towards nephrotoxicity. Both additive and synergistic effects have been described in combination with OTA. OTA is produced in foodstuffs by Aspergillus Section Circumdati (Aspergillus ochraceus, A. westerdijkiae, A. steynii) and Aspergillus Section Nigri (Aspergillus carbonarius, A. foetidus, A. lacticoffeatus, A. niger, A. sclerotioniger, A. tubingensis), mostly in subtropical and tropical areas. OTA is produced in foodstuffs by Penicillium verrucosum and P. nordicum, notably in temperate and colder zones. CIT is produced in foodstuffs by Monascus species (Monascus purpureus, M. ruber) and Penicillium species (Penicillium citrinum, P. expansum, P. radicicola, P. verrucosum). OTA was frequently found in foodstuffs of both plant origin (e.g., cereal products, coffee, vegetable, liquorice, raisins, wine) and animal origin (e.g., pork/poultry). CIT was also found in foodstuffs of vegetable origin (e.g., cereals, pomaceous fruits, black olive, roasted nuts, spices), food supplements based on rice fermented with red microfungi Monascus purpureus and in foodstuffs of animal origin (e.g., cheese).

+

The data 🙏

Expected output labels

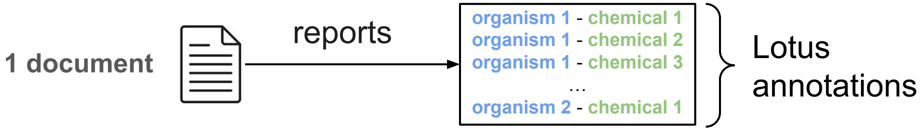
Aspergillus ochraceus - Ochratoxin A
Aspergillus westerdijkiae - Ochratoxin A
Aspergillus steynii - Ochratoxin A
…
Monascus purpureus - Citrinin
Penicillium expansum - Citrinin

But there is no curated datasets …

L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, and Z. Lu, 'BioRED: a rich biomedical relation extraction dataset', Briefings in Bioinformatics, vol. 23, no. 5, p. bbac282, Sep. 2022, doi: 10.1093/bib/bbac282.

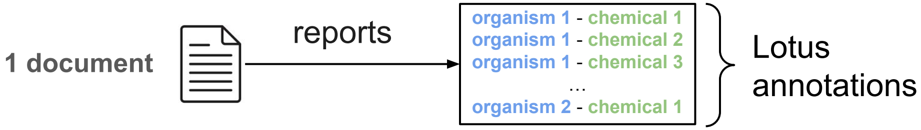
But what about LOTUS ?

LOTUS Harmonization / Processing / Validation / Dissemination



But what about LOTUS ?

LOTUS Harmonization / Processing / Validation / Dissemination



5-hydroxytryptamine-derived alkaloids from two marine sponges of the genus Hyrtios.

Salmoun M, Devijver C ... van Soest RW • J. Nat. Prod.

Add to Collection   BiocXML

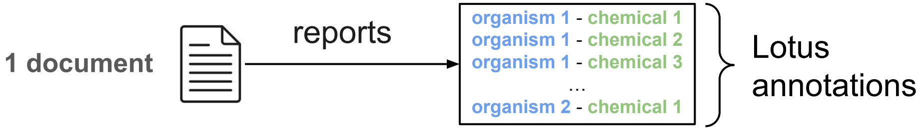
Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of their spectral properties. H. erectus contained hyrtiosulawesine (4), a new beta-carboline alkaloid, together with the already known alkaloids 5-hydroxyindole-3-carbaldehyde (1), hyrtiosin B (2), and 5-hydroxy-3-(2-hydroxyethyl)indole (3). H. reticulatus contained the novel derivative 1,6-dihydroxy-1,2,3,4-tetrahydro-beta-carboline (11) together with serotonin (5), 6-hydroxy-1-methyl-1,2,3,4-tetrahydro-beta-carboline (7), and 6-hydroxy-3,4-dihydro-1-oxo-beta-carboline (9).

- Hyrtios erectus - Hyrtiosulawesine ✓
- Hyrtios erectus - 5-hydroxy-1H-indole-3-carbaldehyde ✗
- Hyrtios erectus - 1,2-bis(5-hydroxy-1H-indol-3-yl)ethane-1,2-dione ✗
- Hyrtios erectus - 5-Hydroxytryptophol ✗
- Hyrtios reticulatus - (1R)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole-1,6-diol ✗
- Hyrtios reticulatus - Serotonin ✓
- Hyrtios reticulatus - (1R)-1-methyl-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indol-6-ol ✗
- Hyrtios reticulatus - 2,3,4,9-Tetrahydro-6-hydroxy-1H-pyrido(3,4-b)indol-1-one ✗
- Hyrtios reticulatus - (S)-6-Hydroxytetrahydroharman ✗

Discrepancies between text and expected output labels

But what about LOTUS ?

LOTUS **Harmonization / Processing / Validation / Dissemination**



5-hydroxytryptamine-derived alkaloids from two marine sponges of the genus Hyrtios.

Salmoun M, Devijver C ... van Soest RW • J. Nat. Prod.

Add to Collection BiocXML

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of their spectral properties. H. erectus contained hyrtiosulawesine (4), a new beta-carboline alkaloid, together with the already known alkaloids 5-hydroxyindole-3-carbaldehyde (1), hyrtiosin B (2), and 5-hydroxy-3-(2-hydroxyethyl)indole (3). H. reticulatus contained the novel derivative 1,6-dihydroxy-1,2,3,4-tetrahydro-beta-carboline (11) together with serotonin (5), 6-hydroxy-1-methyl-1,2,3,4-tetrahydro-beta-carboline (7), and 6-hydroxy-3,4-dihydro-1-oxo-beta-carboline (9).

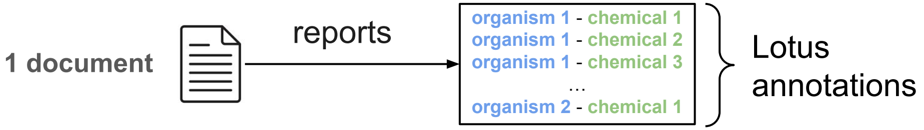
- Hyrtios erectus - Hyrtiosulawesine ✓
- Hyrtios erectus - 5-hydroxy-1H-indole-3-carbaldehyde ✗
- Hyrtios erectus - 1,2-bis(5-hydroxy-1H-indol-3-yl)ethane-1,2-dione ✗
- Hyrtios erectus - 5-Hydroxytryptophol ✗
- Hyrtios reticulatus - (1R)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole-1,6-diol ✗
- Hyrtios reticulatus - Serotonin ✓
- Hyrtios reticulatus - (1R)-1-methyl-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indol-6-ol ✗
- Hyrtios reticulatus - 2,3,4,9-Tetrahydro-6-hydroxy-1H-pyrido(3,4-b)indol-1-one ✗
- Hyrtios reticulatus - (S)-6-Hydroxytetrahydroharman ✗

Discrepancies between text and expected output labels

Not design for building NLP related datasets

But what about LOTUS ?

LOTUS Harmonization / Processing / Validation / Dissemination



5-hydroxytryptamine-derived alkaloids from two marine sponges of the genus Hyrtios.

Salmoun M, Devijver C ... van Soest RW • J. Nat. Prod.

Add to Collection   BiocXML

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of their spectral properties. H. erectus contained hyrtiosulawesine (4), a new beta-carboline alkaloid, together with the already known alkaloids 5-hydroxyindole-3-carbaldehyde (1), hyrtiosin B (2), and 5-hydroxy-3-(2-hydroxyethyl)indole (3). H. reticulatus contained the novel derivative 1,6-dihydroxy-1,2,3,4-tetrahydro-beta-carboline (11) together with serotonin (5), 6-hydroxy-1-methyl-1,2,3,4-tetrahydro-beta-carboline (7), and 6-hydroxy-3,4-dihydro-1-oxo-beta-carboline (9).

- Hyrtios erectus - Hyrtiosulawesine ✓
- Hyrtios erectus - 5-hydroxy-1H-indole-3-carbaldehyde ✗
- Hyrtios erectus - 1,2-bis(5-hydroxy-1H-indol-3-yl)ethane-1,2-dione ✗
- Hyrtios erectus - 5-Hydroxytryptophol ✗
- Hyrtios reticulatus - (1R)-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indole-1,6-diol ✗
- Hyrtios reticulatus - Serotonin ✓
- Hyrtios reticulatus - (1R)-1-methyl-2,3,4,9-tetrahydro-1H-pyrido[3,4-b]indol-6-ol ✗
- Hyrtios reticulatus - 2,3,4,9-Tetrahydro-6-hydroxy-1H-pyrido(3,4-b)indol-1-one ✗
- Hyrtios reticulatus - (S)-6-Hydroxytetrahydroharman ✗

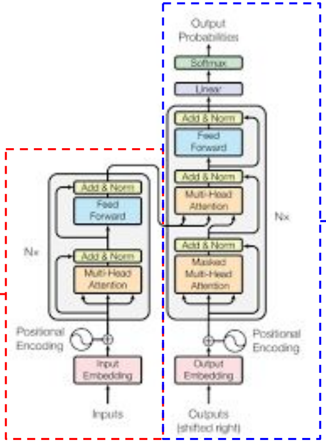
Discrepancies between text and expected output labels

Not design for building NLP related datasets

Can LLM help us in this context ?

What are Large Language Models ?

Transformers



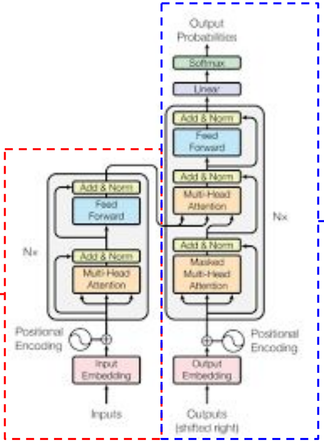
encoder

decoder

Attention is **all you need**
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent … **We** implement this inside of scaled dot-product **attention** by masking out (setting to −∞) …
☆ Enregistrer 99 Citer Cité 110783 fois Autres articles Les 62 versions

What are Large Language Models ?

Transformers



decoder

encoder

Representation

Attention is **all you need**
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent … **We** implement this inside of scaled dot-product **attention** by masking out (setting to −∞) …
☆ Enregistrer 99 Citer Cité 110783 fois Autres articles Les 62 versions »

Bert: Pre-training of deep **bidirectional** transformers for language understanding
J De…
… B…
2018…
☆ E…

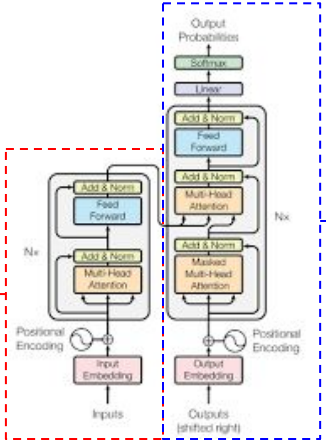
Electra: **Pre-training** text encoders as discriminators rather than generators
K Clark, M…
… back-p…
… gains f…
☆ Enregi…

Roberta: A robustly **optimized bert** pretraining approach
Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen… - arXiv preprint arXiv …, 2019 - arxiv.org
… We find that **BERT** was significantly undertrained and propose an improved … **BERT** models, which we call **RoBERTa**, that can match or exceed the performance of all of the post-**BERT** …
☆ Enregistrer 99 Citer Cité 9454 fois Autres articles Les 6 versions »

What are Large Language Models ?

Transformers

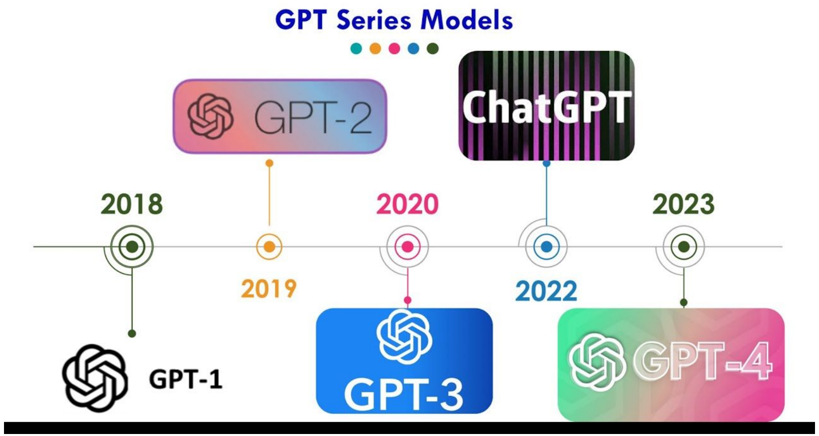
Attention is all you need
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent
… **We** implement this inside of scaled dot-product **attention** by masking out (setting to −∞) …
☆ Enregistrer 99 Citer Cité 110783 fois Autres articles Les 62 versions



decoder → Generation

Representation ← encoder

Generative Pre-training Transformerc



Bert: Pre-training of deep **bidirectional** transformers for language understanding
J De…
… B…
2018…
☆ E…

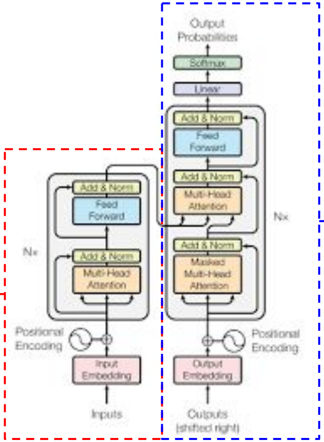
Electra: Pre-training text encoders as discriminators rather than generators
K Clark, M…
… back-p…
… gains f…
☆ Enregi…

Roberta: A robustly **optimized bert** pretraining approach
Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen… - arXiv preprint arXiv …, 2019 - arxiv.org
… We find that **BERT** was significantly undertrained and propose an improved … **BERT** models, which we call **RoBERTa**, that can match or exceed the performance of all of the post-**BERT** …
☆ Enregistrer 99 Citer Cité 9454 fois Autres articles Les 6 versions

What are Large Language Models ?

Transformers

Attention is all you need
A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc
… to attend to **all** positions in the decoder up to and including that position. **We need** to prevent
… **We** implement this inside of scaled dot-product **attention** by masking out (setting to −∞) …
☆ Enregistrer 99 Citer Cité 110783 fois Autres articles Les 62 versions

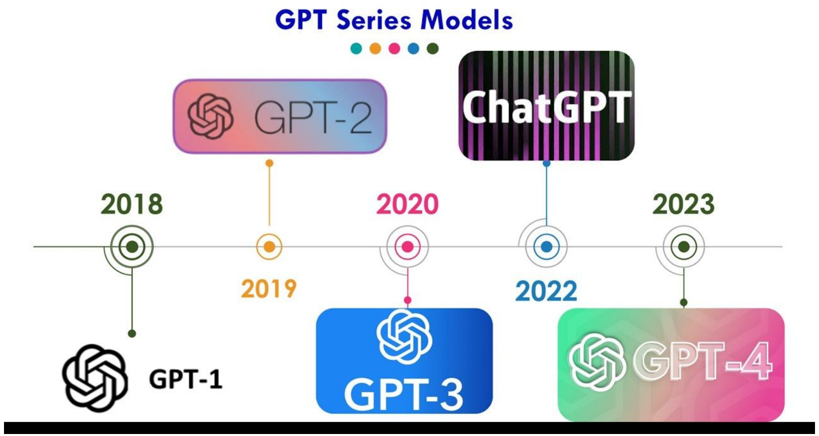


Representation ← encoder

decoder → Generation

Generative Pre-training Transformerc

T5



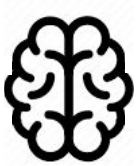
Bert: Pre-training of deep **bidirectional** transformers for language understanding
J De…
… B…
2018…
☆ E…

Electra: Pre-training text encoders as discriminators rather than generators
K Clark, M…
… back-p…
… gains f…
☆ Enregi…

Roberta: A robustly **optimized bert** pretraining approach
Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen… - arXiv preprint arXiv …, 2019 - arxiv.org
… We find that **BERT** was significantly undertrained and propose an improved … **BERT** models, which we call **RoBERTa**, that can match or exceed the performance of all of the post-**BERT** …
☆ Enregistrer 99 Citer Cité 9454 fois Autres articles Les 6 versions

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



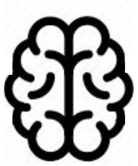
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



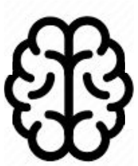
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



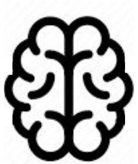
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



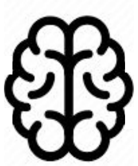
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



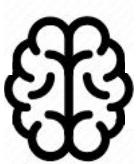
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



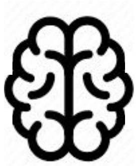
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



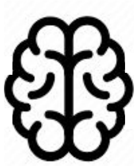
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



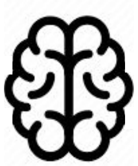
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



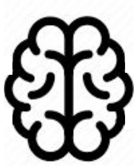
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of their

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



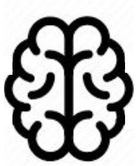
the architecture
(the weights - Billions)

Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of their spectral

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



the architecture
(the weights - Billions)

+

ALL available text

Task: Predict the next word

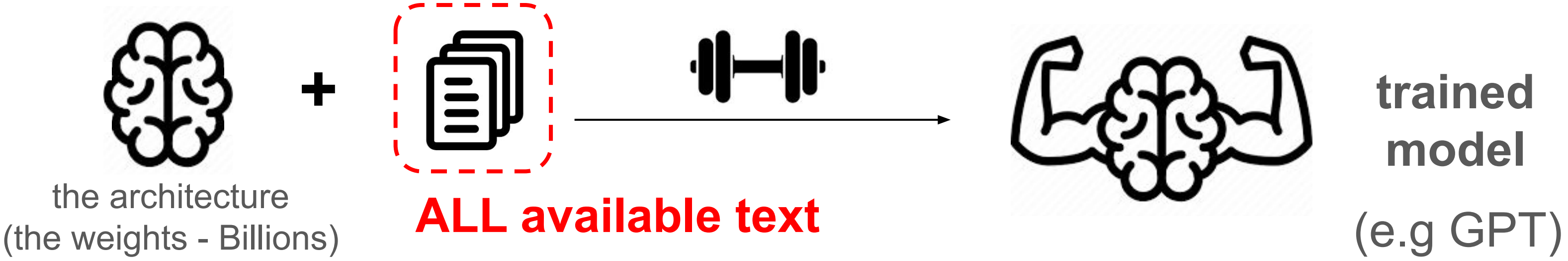
Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of their spectral properties ….

↓

Learn a **representation** of the text

Pre-trained / Fundation models

GPT: Generative **Pre-trained** Models



Task: Predict the next word

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of their spectral properties ….

↓

Learn a **representation** of the text

How Can We Know What Language Models Know?

Zhengbao Jiang[1*] Frank F. Xu[1*] Jun Araki[2] Graham Neubig[1]
Language Technologies Institute, Carnegie Mellon University[1]
Bosch Research North America[2]
{zhengbaj,fangzhex,gneubig}@cs.cmu.edu jun.araki@us.bosch.com

Evaluating Open-Domain Question Answering in the Era of Large Language Models

Ehsan Kamalloo[◇♣] Nouha Dziri[♠] Charles L. A. Clarke[♣] Davood Rafiei[◇]

Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases

Boxi Cao[1,3], Hongyu Lin[1*], Xianpei Han[1,2*], Le Sun[1,2]
Lingyong Yan[1,3], Meng Liao[4], Tong Xue[4], Jin Xu[4]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Data Quality Team, WeChat, Tencent Inc., China
{boxi2020,hongyu,xianpei,sunle,lingyong2014}@iscas.ac.cn
{maricoliao,xavierxue,jinxxu}@tencent.com

KoLA: Carefully Benchmarking World Knowledge of Large Language Models

Jifan Yu*, Xiaozhi Wang*, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, Juanzi Li[†]
Tsinghua University
Beijing, China 100084
kola-benchmark@googlegroups.com

Text as a projection of the world: real knowledge ?

Evaluating Open-QA Evaluation

Cunxiang Wang[1*], Sirui Cheng[2*], Qipeng Guo[3], Yuanhao Yue[4], Bowen Ding[1], Zhikun Xu[4], Yidong Wang[1], Xiangkun Hu[3], Zheng Zhang[3], and Yue Zhang[1†]
[1]School of Engineering, Westlake University, China
[2]Northeastern University, China; [3]Amazon AWS AI; [4]Fudan University, China
{wangcunxiang, zhangyue}@westlake.edu.cn

The Internal State of an LLM Knows When It's Lying

Amos Azaria
School of Computer Science,
Ariel University, Israel

Tom Mitchell
Machine Learning Dept.,
Carnegie Mellon University, Pittsburgh, PA

Evaluating Open-QA Evaluation

Cunxiang Wang[1*], Sirui Cheng[2*], Qipeng Guo[3], Yuanhao Yue[4], Bowen Ding[1], Zhikun Xu[4], Yidong Wang[1], Xiangkun Hu[3], Zheng Zhang[3], and Yue Zhang[1†]
[1]School of Engineering, Westlake University, China
[2]Northeastern University, China; [3]Amazon AWS AI; [4]Fudan University, China
{wangcunxiang, zhangyue}@westlake.edu.cn

Statistical Knowledge Assessment for Large Language Models

Qingxiu Dong[1], Jingjing Xu[2], Lingpeng Kong[3], Zhifang Sui[1] and Lei Li[4]
[1] National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2] Shanghai AI Lab [3] The University of Hong Kong [4] Carnegie Mellon University
dqx@stu.pku.edu.cn, {jingjingxu, szf}@pku.edu.cn, lpk@cs.hku.hk, leili@cs.cmu.ed

Measuring and Modifying Factual Knowledge in Large Language Models

Pouya Pezeshkpour
Megagon Labs
pouya@megagon.ai

Language Models as Knowledge Bases?

Fabio Petroni[1] Tim Rocktäschel[1,2] Patrick Lewis[1,2] Anton Bakhtin[1]
Yuxiang Wu[1,2] Alexander H. Miller[1] Sebastian Riedel[1,2]
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com

LLM as Knowledge bases ?

PubMed ***> 35 million citations*** ▷ *Hard to process …*

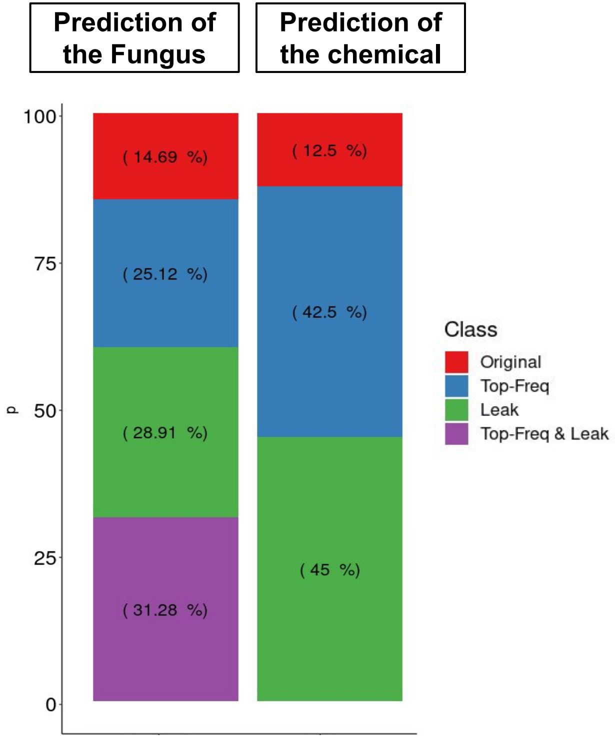
*But, this literature has already been “**digested”** during pre-training !*

LLM as Knowledge bases ?

Pub Med ***> 35 million citations*** → *Hard to process …*

*But, this literature has already been “**digested”** during pre-training !*

- Short answer: **No …**
- **Low global retrieval performances**
- **Correct predictions are mostly leaks**
  E.g: *Monascus Pilosus produces monascin*
- ***Some predictions are generics***
  E.g: Aspergillus Niger or *Ergosterol*



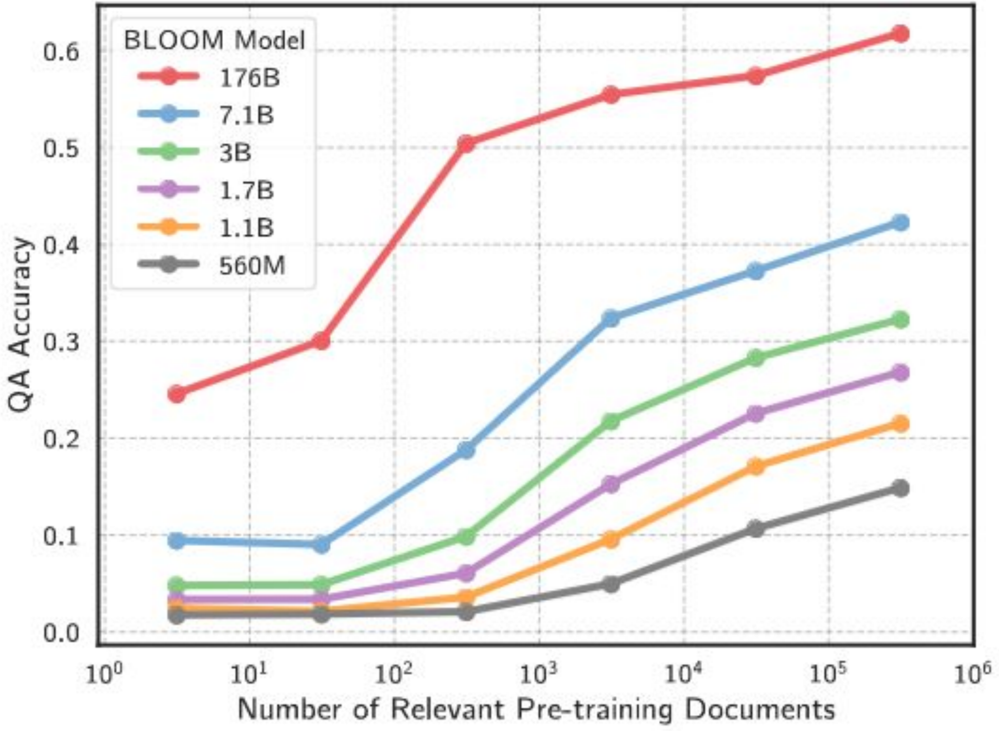
M. Wysocka, O. Wysocki, M. Delmas, V. Mutel, and A. Freitas, ‘Large Language Models, scientific knowledge and factuality: A systematic analysis in antibiotic discovery’. arXiv, Dec. 05, 2023. doi: 10.48550/arXiv.2305.17819.

LLM as Knowledge bases ?

PubMed ***> 35 million citations*** 〉 *Hard to process …*

*But, this literature has already been "**digested"** during pre-training !*

- Short answer: **No …**
- **Low global retrieval performances**
- **Correct predictions are mostly leaks**
  E.g: *Monascus Pilosus produces monascin*
- ***Some predictions are generics***
  E.g: Aspergillus Niger or *Ergosterol*

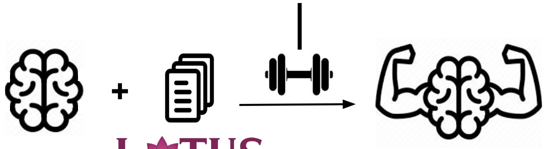


Kandpal, N., Deng, H., Roberts, A., Wallace, E., Raffel, C., 2023. Large Language Models Struggle to Learn Long-Tail Knowledge.

We need to extract the relation from the text !

We need to extract the relation from the text !

Supervised

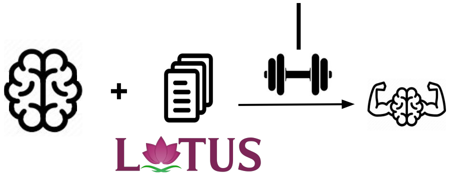


LOTUS

text ≠ output labels

We need to extract the relation from the text !

Supervised



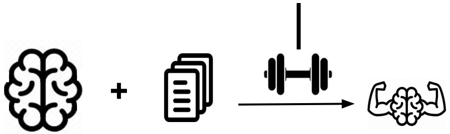
text ≠ output labels

Manually curated test set (GME-sampler)

	precision	recall	f1
Seq2rel	**47.3**	5.8	10.4
GPT-2	44.8	21.7	29.3
BioGPT	42.2	**26.5**	**32.5**

We need to extract the relation from the text !

Supervised



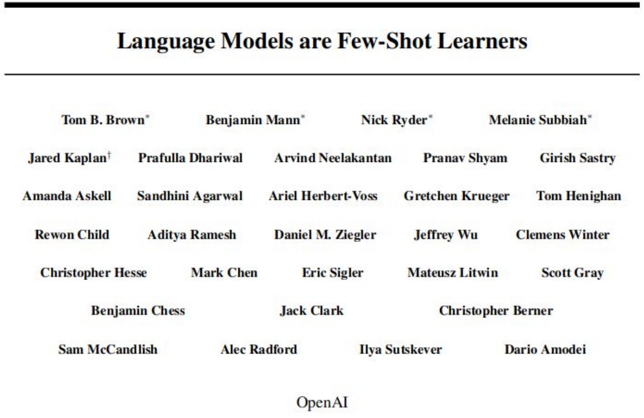
LOTUS

text ≠ output labels

Manually curated test set (GME-sampler)

	precision	recall	f1
Seq2rel	**47.3**	5.8	10.4
GPT-2	44.8	21.7	29.3
BioGPT	42.2	**26.5**	**32.5**

(Weakly) supervised



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

"The model is conditioned on a natural language instruction and/or a few demonstrations of the task and is then expected to complete further instances of the task simply by predicting what comes next."

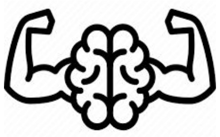
LLM are Few shot learners

Use the representation learned during pre-training

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: LLM completing ….

T. B. Brown et al., 'Language Models are Few-Shot Learners'. 2020.

LLM are Few shot learners



Use the representation learned during pre-training

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

T. B. Brown et al., 'Language Models are Few-Shot Learners'. 2020.

LLM are Few shot learners

Instruction

The task is to extract relations between organisms and chemicals from the input text.

Demonstrations → Archetypal sentences

INPUT: The antimicrobially active EtOH extracts of Maytenus heterophylla yielded a new dihydroagarofuran alkaloid, 1beta-acetoxy-9alpha-benzoyloxy-dihydroagarofuran, together with the known compounds beta-amyrin, maytenfolic acid, …

OUTPUT: Maytenus heterophylla produces 1beta-acetoxy-9alpha-benzoyloxy-dihydroagarofuran. Maytenus heterophylla produces beta-amyrin. Maytenus heterophylla produces maytenfolic acid.

INPUT: Ten new ergosteroids, gloeophyllins A-J (1-10), have been isolated from the solid cultures of Gloeophyllum abietinum.
OUTPUT: Gloeophyllum abietinum produces gloeophyllin A. Gloeophyllum abietinum produces gloeophyllin B. Gloeophyllum abietinum produces gloeophyllin C. Gloeophyllum abietinum produces gloeophyllin D. Gloeophyllum abietinum produces gloeophyllin I. Gloeophyllum abietinum produces gloeophyllin J.

INPUT: The present work describes the isolation of the cyclic peptides geodiamolides A, B, H and I (1-4) from G. corticostylifera and their anti-proliferative effects against sea urchin eggs and human breast cancer cell lineages.
OUTPUT: G. corticostylifera produces geodiamolide A. G. corticostylifera produces geodiamolide B […]

INPUT: Four new cyclic peptides, patellamide G (2) and ulithiacyclamides E-G (3-5), along with the known patellamides A-C (6-8) and ulithiacyclamide B (9), were isolated from the ascidian Lissoclinum patella collected in Pohnpei, Federated States of Micronesia.
OUTPUT: Lissoclinum patella produces patellamide G. Lissoclinum patella produces ulithiacyclamide E. Lissoclinum patella produces ulithiacyclamide F. Lissoclinum patella produces ulithiacyclamide B.

INPUT: Chemical investigation of Trogopterus faeces has led to the isolation of seven flavonoids. Their structures were elucidated by chemical and spectral analyses. In an anticoagulative assay, three kaempferol coumaroyl rhamnosides had significant antithrombin activity. This is the first report on the occurrence of flavonoid glycosides in Trogopterus faeces.
OUTPUT: Trogopterus faeces produces flavonoids. Trogopterus faeces produces kaempferol coumaroyl rhamnosides. Trogopterus faeces produces flavonoid glycosides.

New Instance → **INPUT: ** Abstract ****

To fill → **OUTPUT: `[LLM completing … ]`**

LLM are Few shot learners

Instruction **The task is to extract relations between organisms and chemicals from the input text.**

Demonstrations → Archetypal sentences

INPUT: The antimicrobially active EtOH extracts of Maytenus heterophylla yielded a new dihydroagarofuran alkaloid, 1beta-acetoxy-9alpha-benzoyloxy-dihydroagarofuran, together with the known compounds beta-amyrin, maytenfolic acid, …

OUTPUT: Maytenus heterophylla produces 1beta-acetoxy-9alpha-benzoyloxy-dihydroagarofuran. Maytenus heterophylla produces beta-amyrin. Maytenus heterophylla produces maytenfolic acid.

INPUT: Ten new ergosteroids, gloeophyllins A-J (1-10), have been isolated from the solid cultures of Gloeophyllum abietinum.
OUTPUT: Gloeophyllum abietinum produces gloeophyllin A. Gloeophyllum abietinum produces gloeophyllin B. Gloeophyllum abietinum produces gloeophyllin C. Gloeophyllum abietinum produces gloeophyllin D. Gloeophyllum abietinum produces gloeophyllin I. Gloeophyllum abietinum produces gloeophyllin J.

INPUT: The present work describes the isolation of the cyclic peptides geodiamolides A, B, H and I (1-4) from G. corticostylifera and their anti-proliferative effects against sea urchin eggs and human breast cancer cell lineages.
OUTPUT: G. corticostylifera produces geodiamolide A. G. corticostylifera produces geodiamolide B […]

INPUT: Four new cyclic peptides, patellamide G (2) and ulithiacyclamides E-G (3-5), along with the known patellamides A-C (6-8) and ulithiacyclamide B (9), were isolated from the ascidian Lissoclinum patella collected in Pohnpei, Federated States of Micronesia.
OUTPUT: Lissoclinum patella produces patellamide G. Lissoclinum patella produces ulithiacyclamide E. Lissoclinum patella produces ulithiacyclamide F. Lissoclinum patella produces ulithiacyclamide B.

INPUT: Chemical investigation of Trogopterus faeces has led to the isolation of seven flavonoids. Their structures were elucidated by chemical and spectral analyses. In an anticoagulative assay, three kaempferol coumaroyl rhamnosides had significant antithrombin activity. This is the first report on the occurrence of flavonoid glycosides in Trogopterus faeces.
OUTPUT: Trogopterus faeces produces flavonoids. Trogopterus faeces produces kaempferol coumaroyl rhamnosides. Trogopterus faeces produces flavonoid glycosides.

New Instance → **INPUT:** *Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of their spectral properties. H. erectus contained hyrtiosulawesine (4), a new beta-carboline alkaloid, together with the already known alkaloids 5-hydroxyindole-3-carbaldehyde (1), hyrtiosin B (2), and 5-hydroxy-3-(2-hydroxyethyl)indole (3). H. reticulatus contained the novel derivative 1,6-dihydroxy-1,2,3,4-tetrahydro-beta-carboline (11) together with serotonin (5), 6-hydroxy-1-methyl-1,2,3,4-tetrahydro-beta-carboline (7), and 6-hydroxy-3,4-dihydro-1-oxo-beta-carboline (9).*

To fill → **OUTPUT:** Hyrtios erectus produces hyrtiosulawesine, Hyrtios erectus produces 5-hydroxyindole-3-carbaldehyde. Hyrtios erectus produces hyrtiosin B

We need to extract the relation from the text !

Supervised

LOTUS

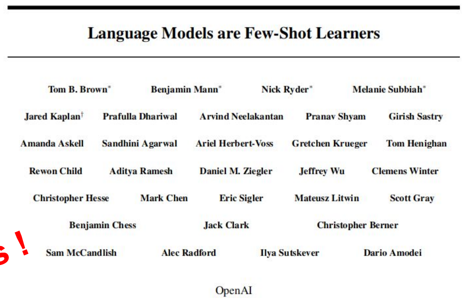
text ≠ output labels

Manually curated test set (GME-sampler)

	precision	recall	f1
Seq2rel	**47.3**	5.8	10.4
GPT-2	44.8	21.7	29.3
BioGPT	42.2	**26.5**	**32.5**

Competitive ↔ *with only 5 examples*

(Weakly) supervised



Language Models are Few-Shot Learners

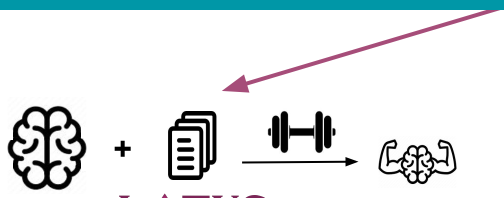
Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

Only Open Models !

LLM	precision	recall	f1
Llama-7B	27.0	9.04	13.55
Llama-13B	35.64	23.64	28.49
Llama-30B	38.51	<u>23.24</u>	28.99
Llama-65B	<u>40.16</u>	22.97	<u>29.23</u>
Alpaca-7B	15.14	2.21	5.86
Vicuna-13B	38.4	20.43	26.48

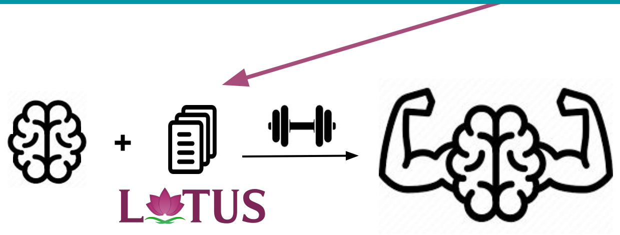
The **data** is the main **bottleneck**



LOTUS

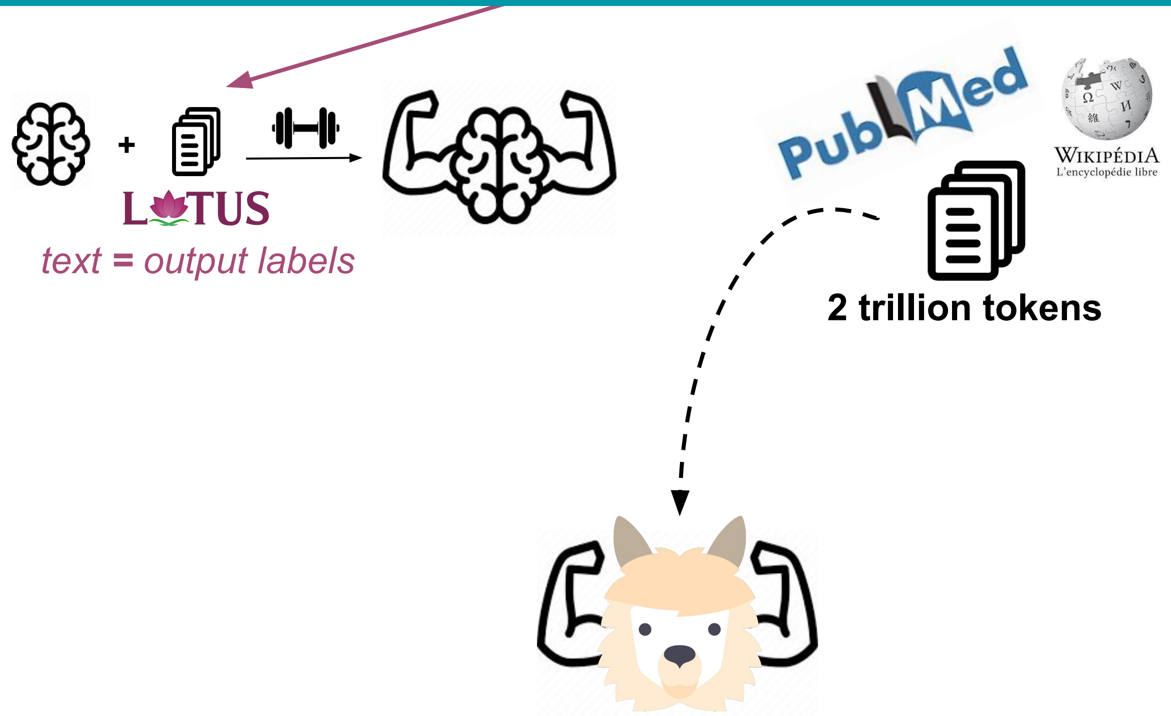
text ≠ output labels

The **data** is the main **bottleneck**

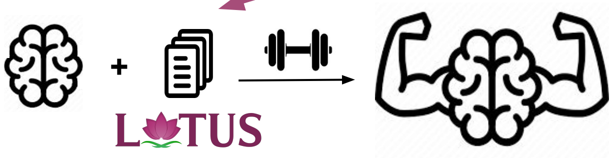


text = output labels

The **data** is the main **bottleneck**



The **data** is the main **bottleneck**



LOTUS

text = output labels

PubMed

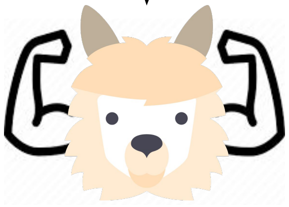
Wikipédia
L'encyclopédie libre

2 trillion tokens

input text

Ochratoxin A (OTA) is a very important mycotoxin, and its research is focused right now on the new findings of OTA, like being a complete carcinogen, information about OTA producers and new exposure sources of OTA. Citrinin (CIT) is another important mycotoxin, too, and its research turns towards nephrotoxicity. Both additive and synergistic effects have been described in combination with OTA. OTA is produced in foodstuffs by Aspergillus Section Circumdati (Aspergillus ochraceus, A. westerdijkiae, A. steynii) and Aspergillus Section Nigri (Aspergillus carbonarius, A. foetidus, A. lacticoffeatus, A. niger, A. sclerotioniger, A. tubingensis), mostly in subtropical and tropical areas. OTA is produced in foodstuffs by Penicillium verrucosum and P. nordicum, notably in temperate and colder zones. CIT is produced in foodstuffs by Monascus species (Monascus purpureus, M. ruber) and Penicillium species (Penicillium citrinum, P. expansum, P. radicicola, P. verrucosum). OTA was frequently found in foodstuffs of both plant origin (e.g., cereal products, coffee, vegetable, liquorice, raisins, wine) and animal origin (e.g., pork/poultry). CIT was also found in foodstuffs of vegetable origin (e.g., cereals, pomaceous fruits, black olive, roasted nuts, spices), food supplements based on rice fermented with red microfungi Monascus purpureus and in foodstuffs of animal origin (e.g., cheese).

task

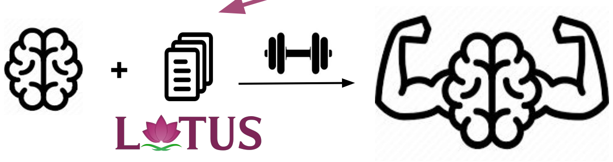


hard

Expected relations

Aspergillus ochraceus - Ochratoxin A
Aspergillus westerdijkiae - Ochratoxin A
Aspergillus steynii - Ochratoxin A
...
Monascus purpureus - Citrinin
Penicillium expansum - Citrinin

The **data** is the main **bottleneck**



LOTUS

text ≠ output labels

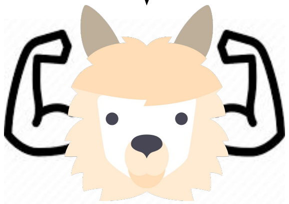
PubMed

Wikipédia
L'encyclopédie libre

2 trillion tokens

input text

Ochratoxin A (OTA) is a very important mycotoxin, and its research is focused right now on the new findings of OTA, like being a complete carcinogen, information about OTA producers and new exposure sources of OTA. Citrinin (CIT) is another important mycotoxin, too, and its research turns towards nephrotoxicity. Both additive and synergistic effects have been described in combination with OTA. OTA is produced in foodstuffs by Aspergillus Section Circumdati (Aspergillus ochraceus, A. westerdijkiae, A. steynii) and Aspergillus Section Nigri (Aspergillus carbonarius, A. foetidus, A. lacticoffeatus, A. niger, A. sclerotioniger, A. tubingensis), mostly in subtropical and tropical areas. OTA is produced in foodstuffs by Penicillium verrucosum and P. nordicum, notably in temperate and colder zones. CIT is produced in foodstuffs by Monascus species (Monascus purpureus, M. ruber) and Penicillium species (Penicillium citrinum, P. expansum, P. radicicola, P. verrucosum). OTA was frequently found in foodstuffs of both plant origin (e.g., cereal products, coffee, vegetable, liquorice, raisins, wine) and animal origin (e.g., pork/poultry). CIT was also found in foodstuffs of vegetable origin (e.g., cereals, pomaceous fruits, black olive, roasted nuts, spices), food supplements based on rice fermented with red microfungi Monascus purpureus and in foodstuffs of animal origin (e.g., cheese).



Expected relations

Aspergillus ochraceus - Ochratoxin A
Aspergillus westerdijkiae - Ochratoxin A
Aspergillus steynii - Ochratoxin A
…
Monascus purpureus - Citrinin
Penicillium expansum - Citrinin

What about reversing the task ?

Creating controlled synthetic input text from the expected relations

What is inside a PubMed entry ?

What do we need ?

> J Nat Prod. 2002 Aug;65(8):1173-6. doi: 10.1021/np020009+.

5-hydroxytryptamine-derived alkaloids from two marine sponges of the genus Hyrtios

← *A title*

Mostafa Salmoun [1], Christine Devijver, Désiré Daloze, Jean-Claude Braekman, Rob W M van Soest

Affiliations + expand

PMID: 12193025 DOI: 10.1021/np020009+

Abstract

Indonesian specimens of the marine sponges Hyrtios erectus and H. reticulatus were found to contain 5-hydroxytryptamine-derived alkaloids. Their structures were determined on the basis of their spectral properties. H. erectus contained hyrtiosulawesine (4), a new beta-carboline alkaloid, together with the already known alkaloids 5-hydroxyindole-3-carbaldehyde (1), hyrtiosin B (2), and 5-hydroxy-3-(2-hydroxyethyl)indole (3). H. reticulatus contained the novel derivative 1,6-dihydroxy-1,2,3,4-tetrahydro-beta-carboline (11) together with serotonin (5), 6-hydroxy-1-methyl-1,2,3,4-tetrahydro-beta-carboline (7), and 6-hydroxy-3,4-dihydro-1-oxo-beta-carboline (9).

← *The abstract (What we want to generate)*

MeSH terms

- Animals
- Chromatography, Thin Layer
- Indole Alkaloids / chemistry
- Indole Alkaloids / isolation & purification*
- Indonesia
- Molecular Structure
- Nuclear Magnetic Resonance, Biomolecular
- Porifera / chemistry*
- Serotonin / analogs & derivatives*
- Serotonin / chemistry
- Serotonin / isolation & purification*
- Spectrophotometry, Ultraviolet
- Stereoisomerism

← *Some keywords / keyphrases*

What is inside a PubMed entry ?

What do we need ?

> J Nat Prod. 2002 Aug;65(8):1173-6. doi: 10.1021/np020009+.

5-hydroxytryptamine-derived alkaloids from two marine sponges of the genus Hyrtios

← *A title*

Mostafa Salmoun [1], Christine Devijver, Désiré Daloze, Jean-Claude Braekman, Rob W M van Soest

Affiliations + expand

PMID: 12193025 DOI: 10.1021/np020009+

Abstract

Given a title, some keywords and the main findings (expected relations), create a scientific abstract

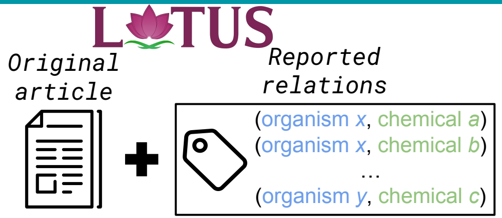
← *The abstract (What we want to generate)*

MeSH terms

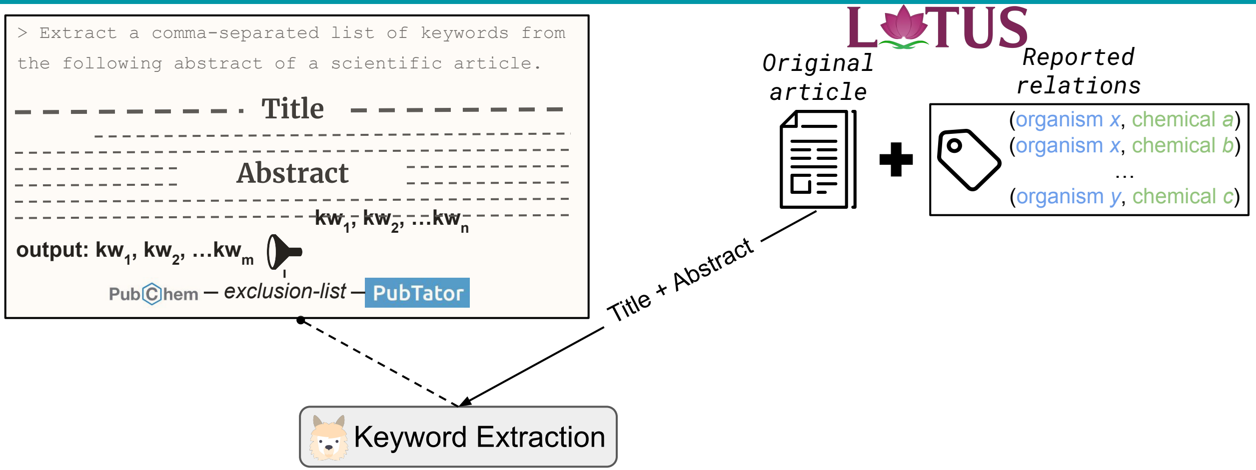
- Animals
- Chromatography, Thin Layer
- Indole Alkaloids / chemistry
- Indole Alkaloids / isolation & purification*
- Indonesia
- Molecular Structure
- Nuclear Magnetic Resonance, Biomolecular
- Porifera / chemistry*
- Serotonin / analogs & derivatives*
- Serotonin / chemistry
- Serotonin / isolation & purification*
- Spectrophotometry, Ultraviolet
- Stereoisomerism

← *Some keywords / keyphrases*

Synthetic generation

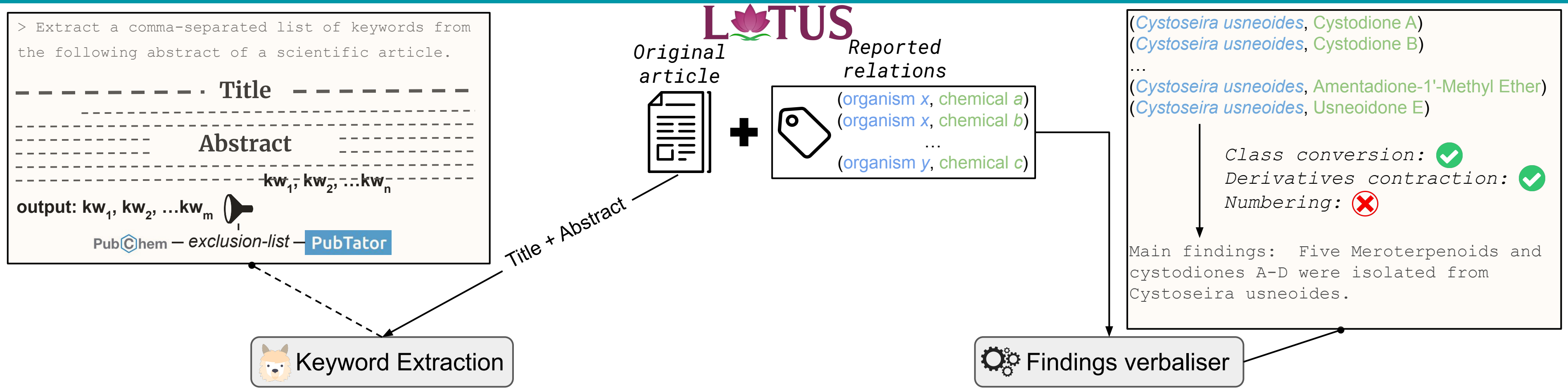


Synthetic generation

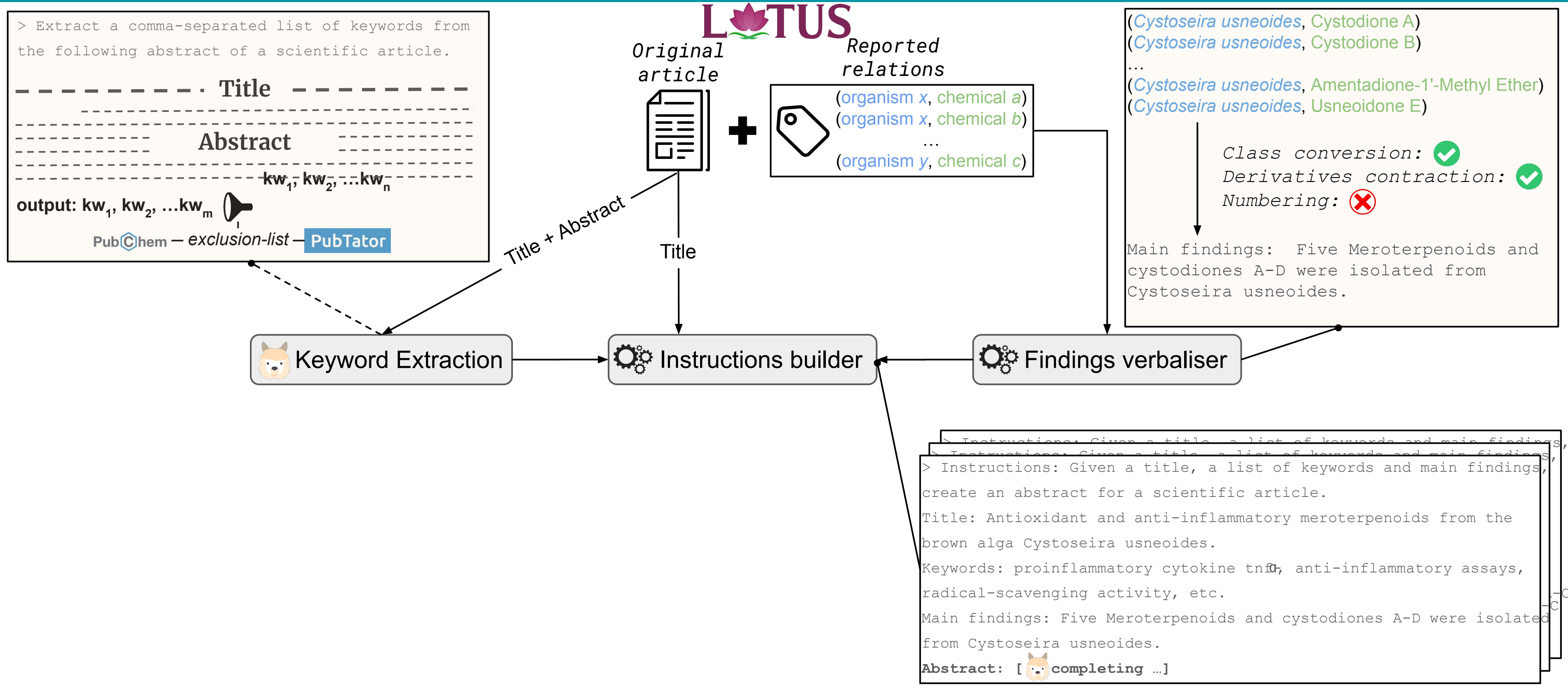


Not all articles have MeSH

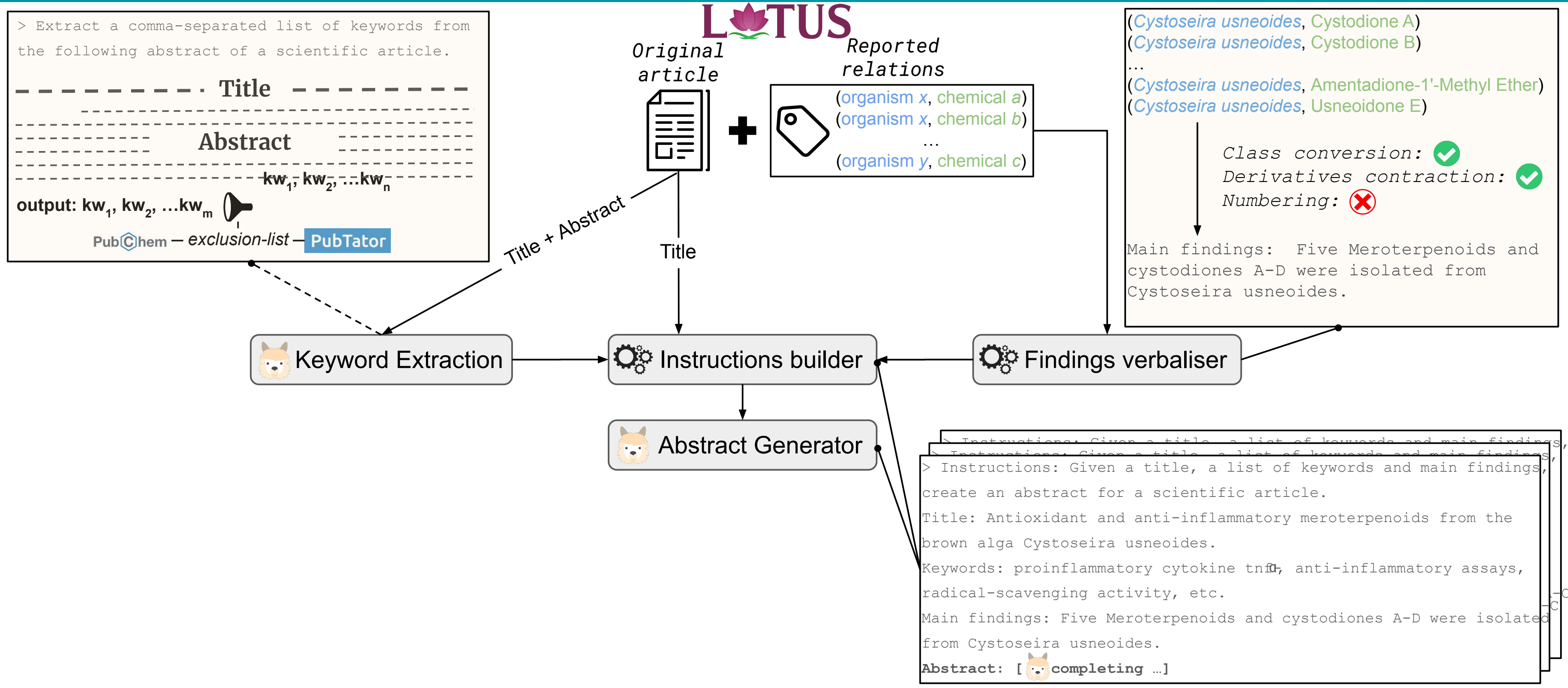
Synthetic generation



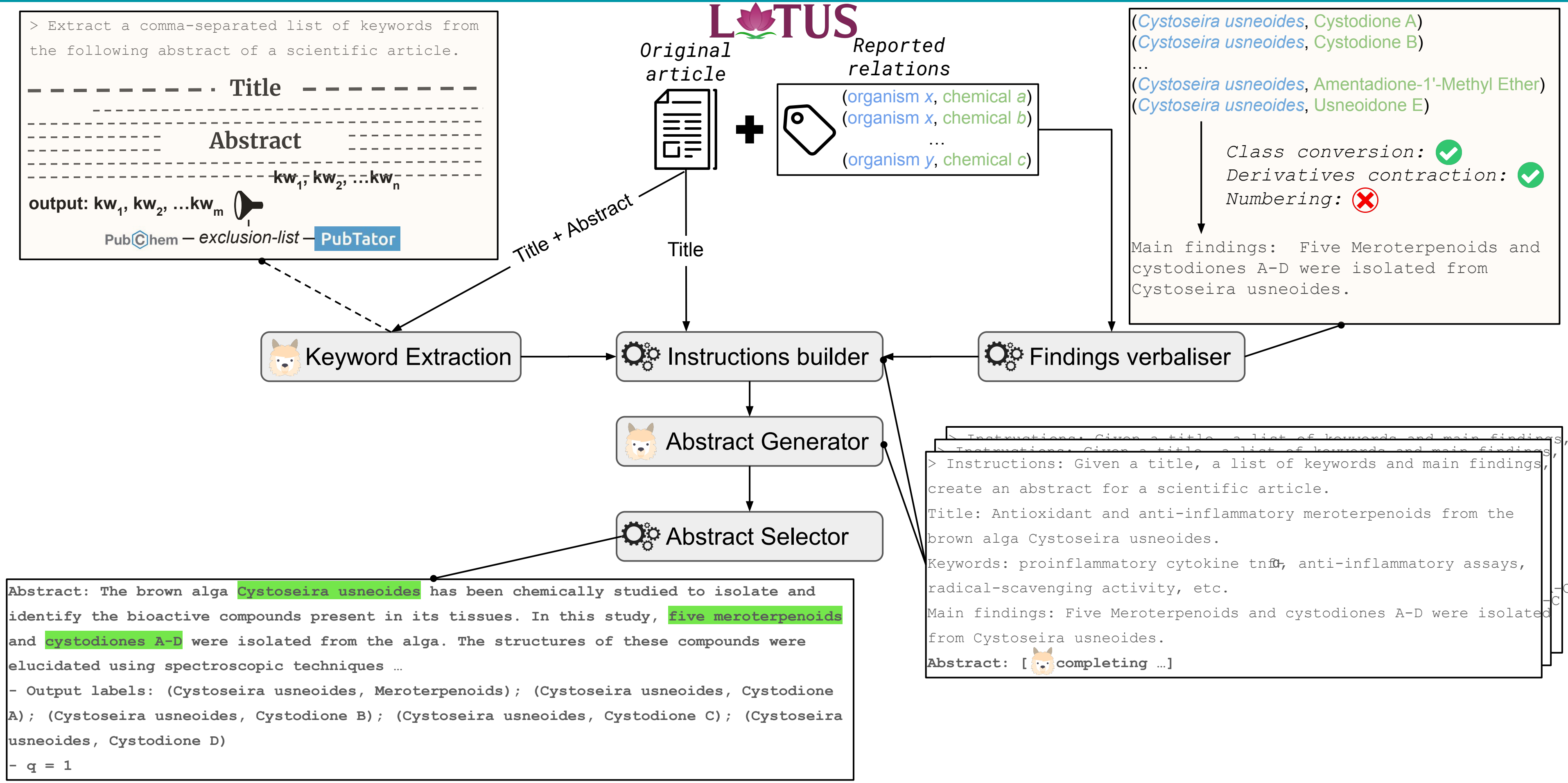
Synthetic generation



Synthetic generation



Synthetic generation



Which one is synthetic ?

> J Nat Prod. 2009 Jul;72(7):1361-3. doi: 10.1021/np900181q.

Salvinorins J from Salvia divinorum: mutarotation in the neoclerodane system

Lukasz M Kutrzeba [1], Daneel Ferreira, Jordan K Zjawiony

Affiliations + expand

Abstract

A search for biosynthetic precursors of salvinorin A (1) led to the isolation of a new neoclerodane diterpenoid hemiacetal mixture, salvinorins J (2), from the chloroform extract of Salvia divinorum. A leaf surface extraction method was used on S. divinorum, affording a chlorophyll-free extract containing predominantly neoclerodane diterpenoids, including the new salvinorins J (2) and 14 known analogues. Salvinorins J (2) represent an example of a neoclerodane hemiacetal (lactol) susceptible to mutarotation with the formation of an equilibrium mixture of C-17 epimers.

> J Nat Prod. 2009 Jul;72(7):1361-3. doi: 10.1021/np900181q.

Salvinorins J from Salvia divinorum: mutarotation in the neoclerodane system

Lukasz M Kutrzeba [1], Daneel Ferreira, Jordan K Zjawiony

Affiliations + expand

Abstract

The neoclerodane system of Salvia divinorum has been found to undergo mutarotation, resulting in the formation of c-17 epimers. This was confirmed through the isolation and identification of several compounds from the plant using a leaf surface extraction method. The compounds isolated include Salvinorin A (1), Salvinorin B (2), Salvinorins F-H (3-5), (-)-Hardwickiic acid (6) and divinatorins A-C (7-9). These findings provide new insights into the chemical diversity of Salvia divinorum and highlight the importance of considering mutarotation in the study of neoclerodane systems.

Which one is synthetic ?

Real

J Nat Prod. 2009 Jul;72(7):1361-3. doi: 10.1021/np900181q.

Salvinorins J from Salvia divinorum: mutarotation in the neoclerodane system

Lukasz M Kutzreba [1], Daneel Ferreira, Jordan K Zjawiony

Affiliations + expand

Abstract

A search for biosynthetic precursors of salvinorin A (1) led to the isolation of a new neoclerodane diterpenoid hemiacetal mixture, salvinorins J (2), from the chloroform extract of Salvia divinorum. A leaf surface extraction method was used on S. divinorum, affording a chlorophyll-free extract containing predominantly neoclerodane diterpenoids, including the new salvinorins J (2) and 14 known analogues. Salvinorins J (2) represent an example of a neoclerodane hemiacetal (lactol) susceptible to mutarotation with the formation of an equilibrium mixture of C-17 epimers.

synthetic

J Nat Prod. 2009 Jul;72(7):1361-3. doi: 10.1021/np900181q.

Salvinorins J from Salvia divinorum: mutarotation in the neoclerodane system

Lukasz M Kutzreba [1], Daneel Ferreira, Jordan K Zjawiony

Affiliations + expand

Abstract

The neoclerodane system of Salvia divinorum has been found to undergo mutarotation, resulting in the formation of c-17 epimers. This was confirmed through the isolation and identification of several compounds from the plant using a leaf surface extraction method. The compounds isolated include Salvinorin A (1), Salvinorin B (2), Salvinorins F-H (3-5), (-)-Hardwickiic acid (6) and divinatorins A-C (7-9). These findings provide new insights into the chemical diversity of Salvia divinorum and highlight the importance of considering mutarotation in the study of neoclerodane systems.

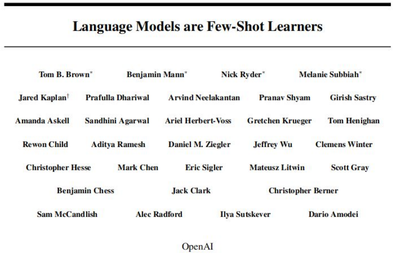
Re-training on synthetic data

Supervised

LOTUS

text ≠ output labels

(Weakly) supervised



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

	precision	recall	f1
Seq2rel	47.3	5.8	10.4
GPT-2	44.8	21.7	29.3
BioGPT	42.2	26.5	32.5

LLM	precision	recall	f1
Llama-7B	27.0	9.04	13.55
Llama-13B	35.64	23.64	28.49
Llama-30B	38.51	23.24	28.99
Llama-65B	40.16	22.97	29.23
Alpaca-7B	15.14	2.21	5.86
Vicuna-13B	38.4	20.43	26.48

Re-training on synthetic data

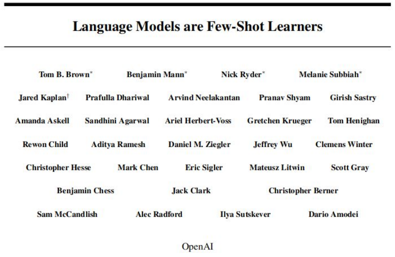
Supervised

LOTUS

text ≠ output labels

	precision	recall	f1
Seq2rel	47.3	5.8	10.4
GPT-2	44.8	21.7	29.3
BioGPT	42.2	26.5	32.5

(Weakly) supervised



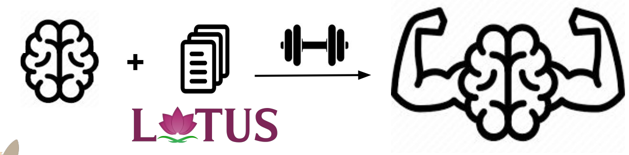
Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

LLM	precision	recall	f1
Llama-7B	27.0	9.04	13.55
Llama-13B	35.64	23.64	28.49
Llama-30B	38.51	23.24	28.99
Llama-65B	40.16	22.97	29.23
Alpaca-7B	15.14	2.21	5.86
Vicuna-13B	38.4	20.43	26.48

Supervised - on synthetic data



LOTUS

text = output labels

model	precision	recall	f1
seq2rel	**65.1** (+17.8)	29.9 (+22.0)	41.0 (+28.9)
GPT2	52.0 (+7.2)	44.6 (+22.9)	48.0 (+18.7)
BioGPT	63.7 (+21.5)	**46.5** (+20.0)	**53.8** (+21.3)

Re-training on synthetic data

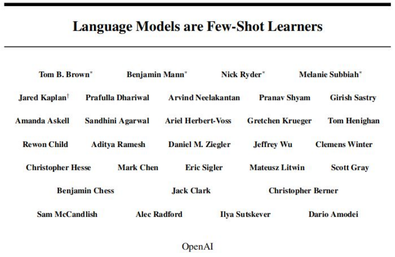
Supervised

LOTUS

text ≠ output labels

	precision	recall	f1
Seq2rel	47.3	5.8	10.4
GPT-2	44.8	21.7	29.3
BioGPT	42.2	26.5	32.5

(Weakly) supervised



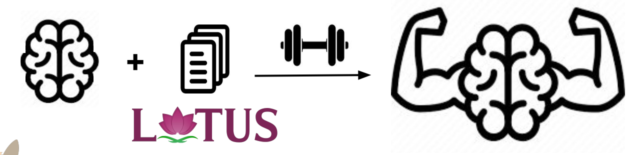
Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray
Benjamin Chess Jack Clark Christopher Berner
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

LLM	precision	recall	f1
Llama-7B	27.0	9.04	13.55
Llama-13B	35.64	23.64	28.49
Llama-30B	38.51	23.24	28.99
Llama-65B	40.16	22.97	29.23
Alpaca-7B	15.14	2.21	5.86
Vicuna-13B	38.4	20.43	26.48

Supervised - on synthetic data



LOTUS

text = output labels

model	precision	recall	f1
seq2rel	**65.1** (+17.8)	29.9 (+22.0)	41.0 (+28.9)
GPT2	52.0 (+7.2)	44.6 (+22.9)	48.0 (+18.7)
BioGPT	63.7 (+21.5)	**46.5** (+20.0)	**53.8** (+21.3)

model	precision	recall	f1
BioGPT-Large	69	51.6	59.0

Conclusion: LLM are versatiles

- Fine-tuning: *garbage in, garbage out !*

Conclusion: LLM are versatiles

- Fine-tuning: *garbage in, garbage out !*

- Impressive few-shot learners, but **language models** above all

Conclusion: LLM are versatiles

- Fine-tuning: *garbage in, garbage out !*
- Impressive few-shot learners, but **language models** above all
- Better Synthetic Data Generator (Knowledge distillation)

Conclusion: LLM are versatiles

- Fine-tuning: *garbage in, garbage out !*
- Impressive few-shot learners, but **language models** above all
- Better Synthetic Data Generator (Knowledge distillation)
- Multi-lingual opportunities

Conclusion: LLM are versatiles

- Fine-tuning: *garbage in, garbage out !*
- Impressive few-shot learners, but **language models** above all
- Better Synthetic Data Generator (Knowledge distillation)
- Multi-lingual opportunities

But,

- Narrow range of “styles” compared to human-written abstracts
- Hard to control the quality
- LLM evolve fast (15 Feb. 2024 - *BioMistral*)

Thanks for your attention

André
Freitas

Magdalena
Wysocka

RELATION EXTRACTION IN UNDEREXPLORED BIOMEDICAL DOMAINS: A DIVERSITY-OPTIMISED SAMPLING AND SYNTHETIC DATA GENERATION APPROACH

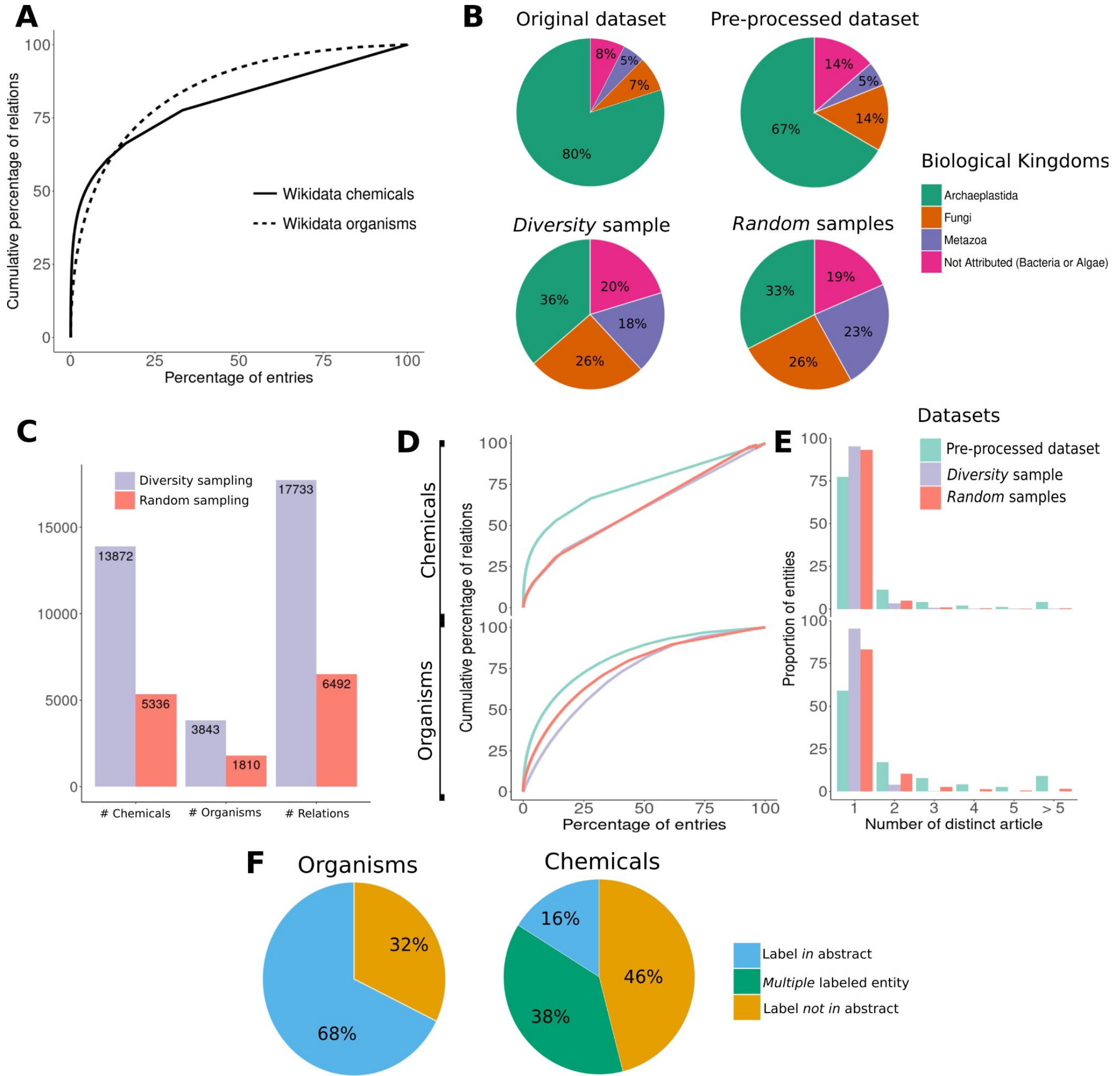
A PREPRINT

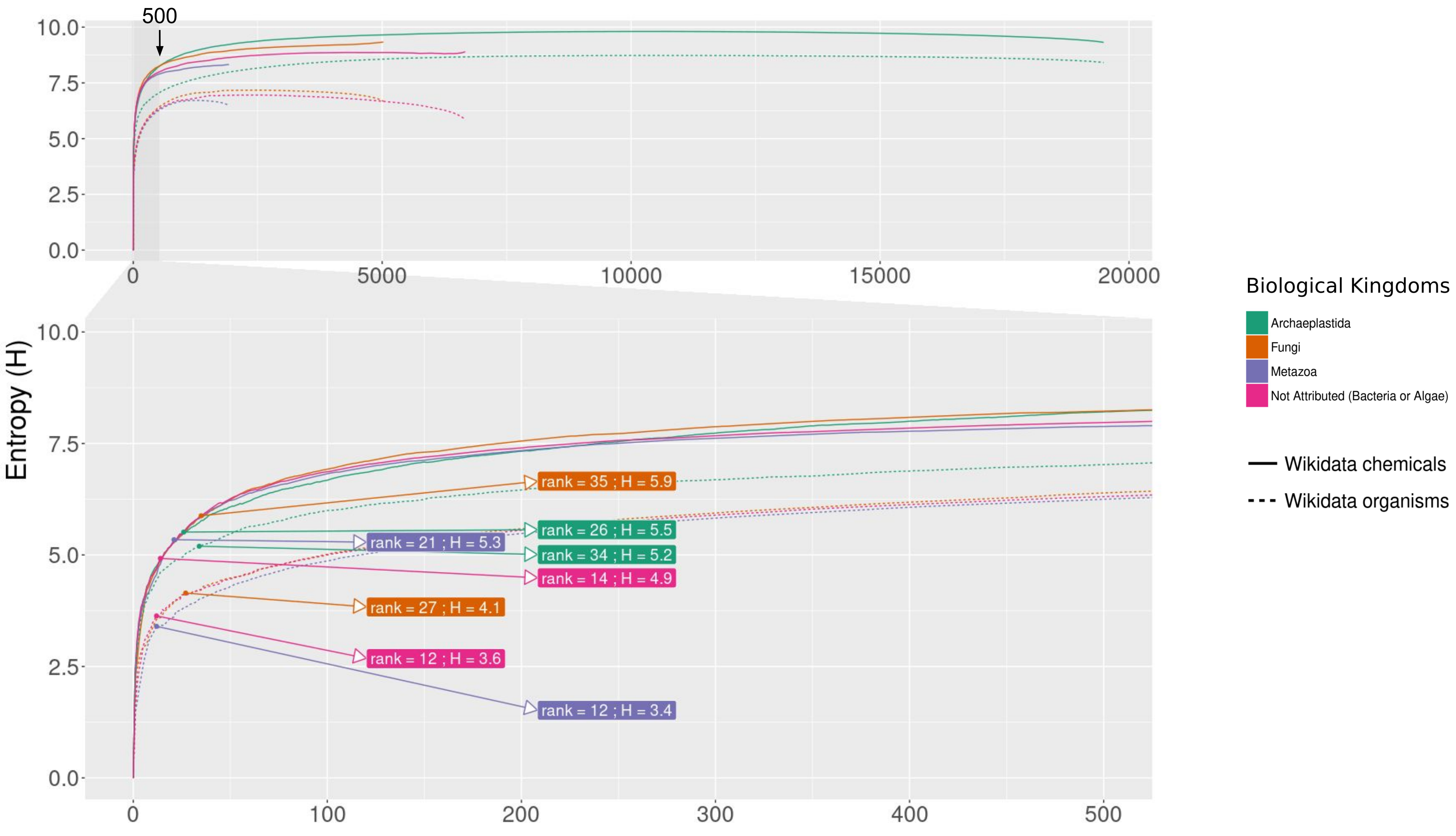
Maxime Delmas*[1], Magdalena Wysocka[2], and André Freitas[1,2,3]

[1]*Idiap Research Institute, Switzerland*
[2]*Digital Experimental Cancer Medicine Team, Cancer Biomarker Centre, CRUK Manchester Institute*
[3]*Department of Computer Science, University of Manchester*

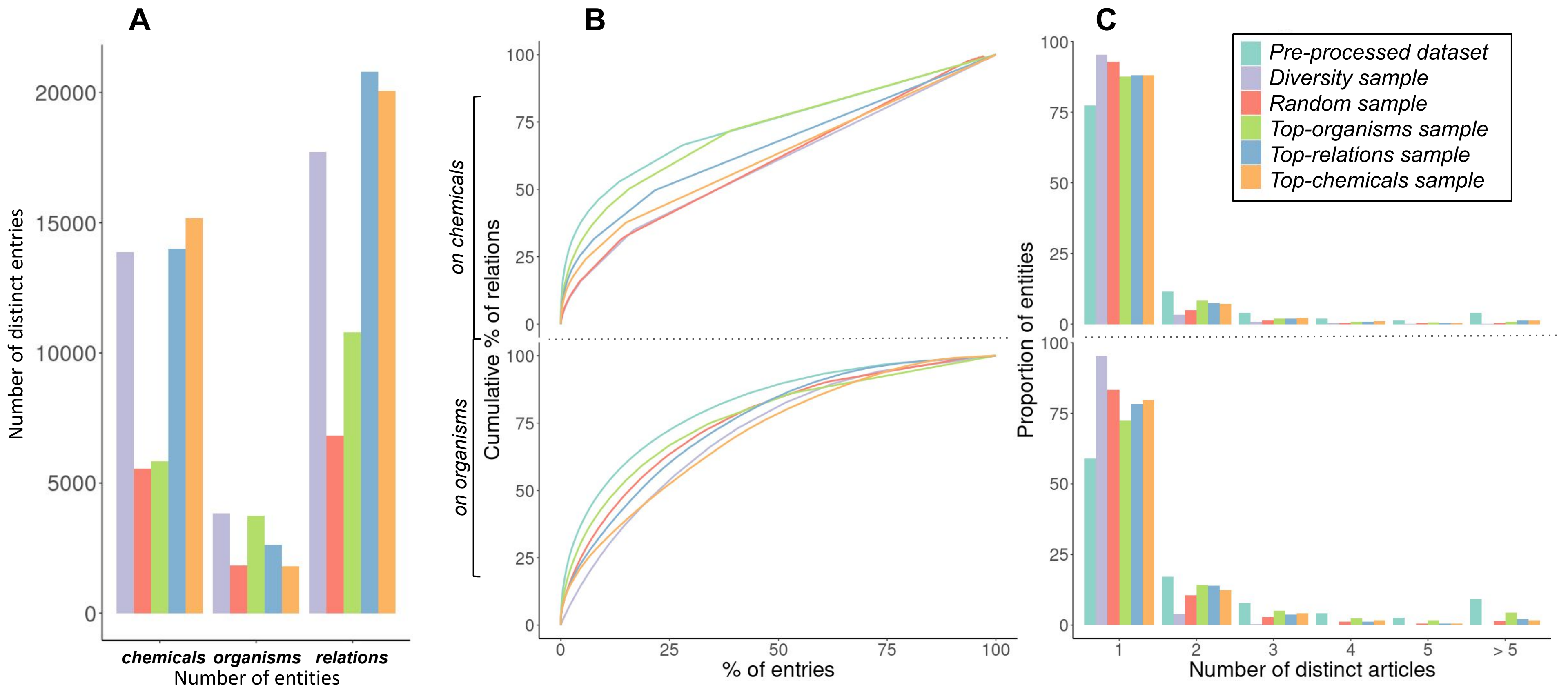
arXiv https://arxiv.org/pdf/2311.06364.pdf

https://github.com/idiap/abroad-re
https://github.com/idiap/gme-sampler

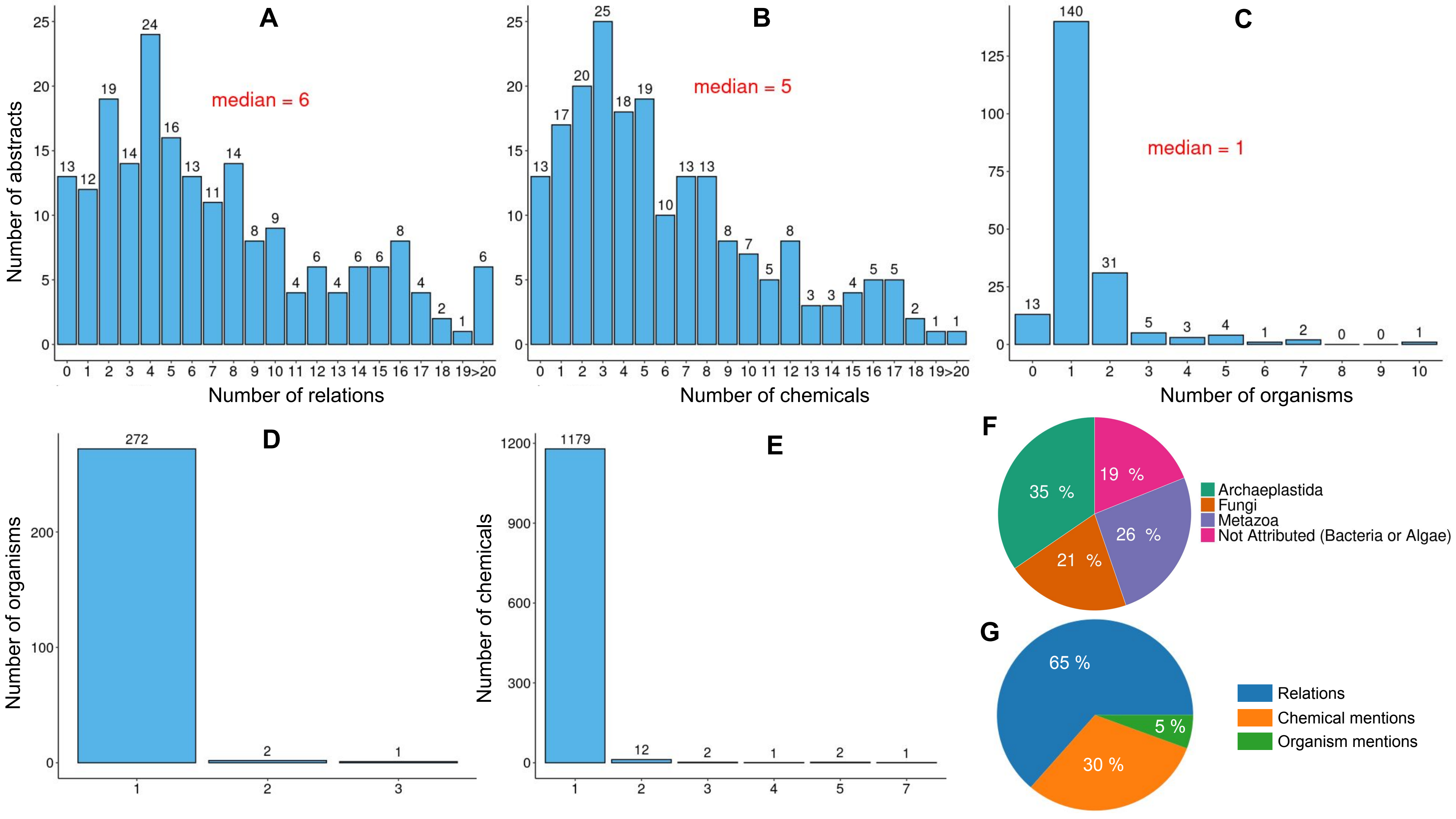

A
Cumulative percentage of relations
Percentage of entries
Wikidata chemicals
Wikidata organisms
B
Original dataset
Pre-processed dataset
Diversity sample
Random samples
Biological Kingdoms
Archaeplastida
Fungi
Metazoa
Not Attributed (Bacteria or Algae)
C
Diversity sampling
Random sampling
13872
5336
3843
1810
17733
6492
Chemicals
Organisms
Relations
D
Chemicals
Organisms
Cumulative percentage of relations
Percentage of entries
E
Datasets
Pre-processed dataset
Diversity sample
Random samples
Proportion of entities
Number of distinct article
F
Organisms
Chemicals
Label in abstract
Multiple labeled entity
Label not in abstract



500
Entropy (H)
rank = 35 ; H = 5.9
rank = 26 ; H = 5.5
rank = 21 ; H = 5.3
rank = 34 ; H = 5.2
rank = 14 ; H = 4.9
rank = 27 ; H = 4.1
rank = 12 ; H = 3.6
rank = 12 ; H = 3.4
Biological Kingdoms
Archaeplastida
Fungi
Metazoa
Not Attributed (Bacteria or Algae)
Wikidata chemicals
Wikidata organisms



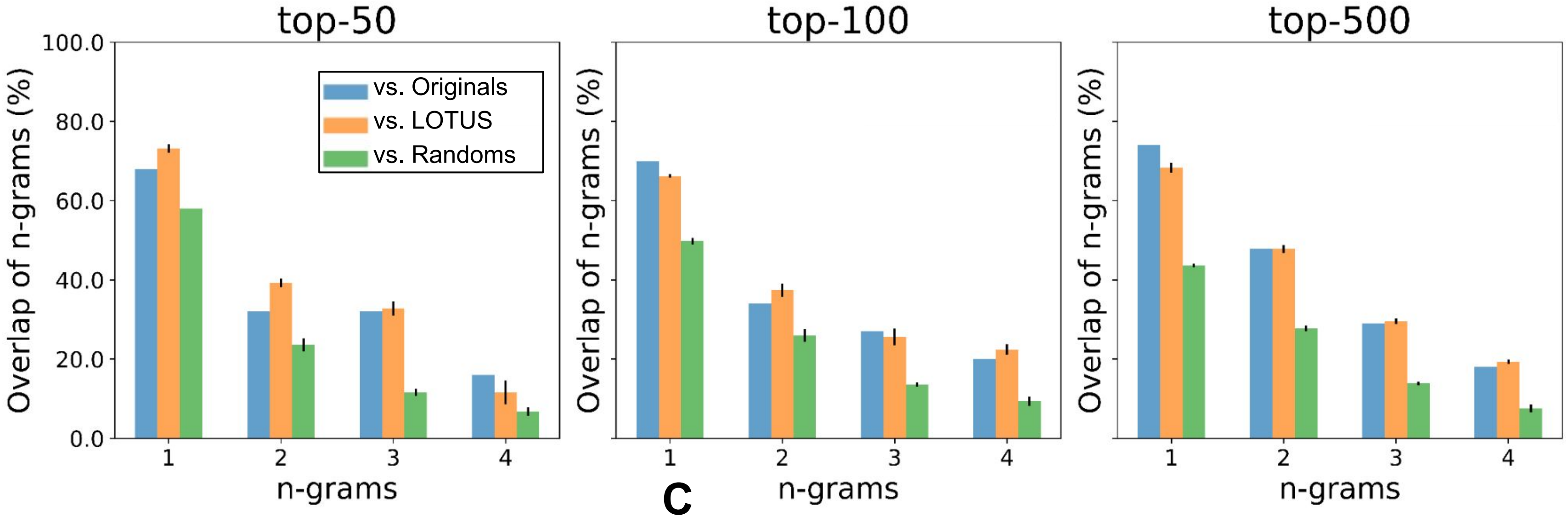
A
B
C
Number of distinct entries
20000
15000
10000
5000
0
chemicals organisms relations
Number of entities
on chemicals
on organisms
Cumulative % of relations
100
75
50
25
0
% of entries
Proportion of entities
Number of distinct articles
1
2
3
4
5
> 5
Pre-processed dataset
Diversity sample
Random sample
Top-organisms sample
Top-relations sample
Top-chemicals sample


A
B
C
D
E
F
G
median = 6
median = 5
median = 1
Number of abstracts
Number of relations
Number of chemicals
Number of organisms
Number of organisms
Number of chemicals
Archaeplastida
Fungi
Metazoa
Not Attributed (Bacteria or Algae)
35 %
19 %
26 %
21 %
65 %
30 %
5 %
Relations
Chemical mentions
Organism mentions

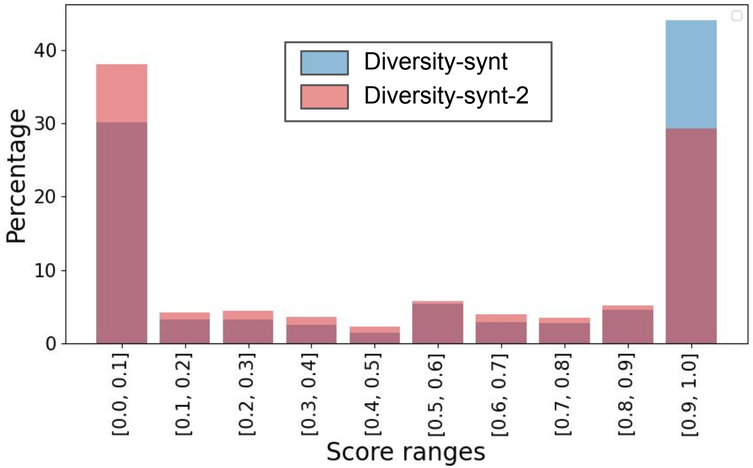
model	Training	precision	recall	f1
Llama-7B	*Few-shots learning (5-shots)*	27.0	9.04	13.55
Llama-13B		35.64	23.64	28.49
Llama-30B		38.51	23.24	28.99
Llama-65B		40.16	22.97	29.23
Alpaca-7B		15.14	2.21	5.86
Vicuna-13B		38.4	20.43	26.48
Seq2rel	*Random-raw*	43.2 +/- (6.67)	4.8 +/- (1.16)	8.6 +/- (2.00)
	Diversty-raw	39.6	5.4	9.5
	Extended-raw	47.3	5.8	10.4
	Full	45.6	7.1	12.2
GPT2-QLoRA	*Random-raw*	32.5 +/- (4.83)	11.8 +/- (5.25)	15.0 +/- (2.54)
	Diversty-raw	22.3	19.2	20.6
	Extended-raw	44.8	21.7	29.3
	Full	**47.5**	22.5	30.5
BioGPT-QLoRa	*Random-raw*	47.2 +/- (4.01)	19.8 +/- (2.71)	27.6 +/- (2.48)
	Diversty-raw	37.1	**28.4**	32.2
	Extended-raw	42.2	26.5	**32.5**
	Full	46.7	21.3	29.3

model	Dataset	precision	recall	f1
	Random-synt.	62.4 +/- (1.03)	26.8 +/- (1.96)	37.5 +/- (1.90)
	Diversty-synt.	61.5	30.7	40.1
Seq2rel	*Extended-synt.*	**65.1**	29.9	41.0
	Random-synt.	42.6 +/- (2.89)	32.7 +/- (2.81)	37.2 +/- (2.80)
	Diversty-synt.	28.5	39.4	33.0
GPT2-QLoRA	*Extended-synt.*	52.0	<u>44.6</u>	<u>48.0</u>
	Random-synt.	56.4 +/- (2.26)	38.8 +/- (1.92)	46.0 +/- 1.08
	Diversty-synt.	53.1	41.6	*46.6*
BioGPT-QLoRa	*Extended-synt.*	<u>63.7</u>	**46.5**	**53.8**

A



B



C

Model	Dataset	Precision	Recall	f1
Seq2rel	*Diversity-synt*	61.5	30.7	40.1
	Diversity-synt-2-selector	**65.2**	28.7	39.9
	Diversity-synt-2-NO-selector	60.3	24.0	34.3
GPT-2	*Diversity-synt*	28.5	39.4	33.0
	Diversity-synt-2-selector	15.3	22.7	18.3
	Diversity-synt-2-NO-selector	15.4	20.1	17.4
BioGPT	*Diversity-synt*	52.5	**41.2**	**46.2**
	Diversity-synt-2-selector	44.5	36.4	40.0
	Diversity-synt-2-NO-selector	45.8	34.3	39.2