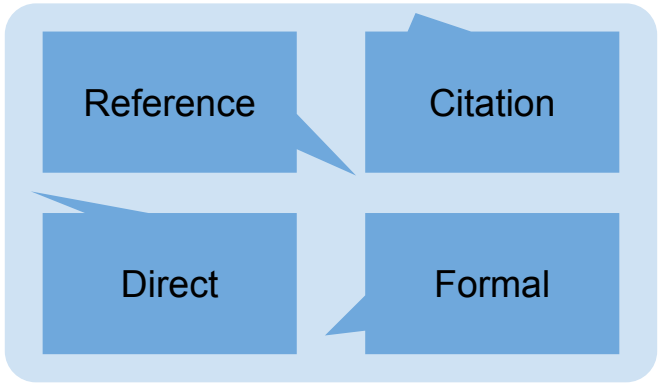
JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

Data citation function classification with large language models (LLMs)

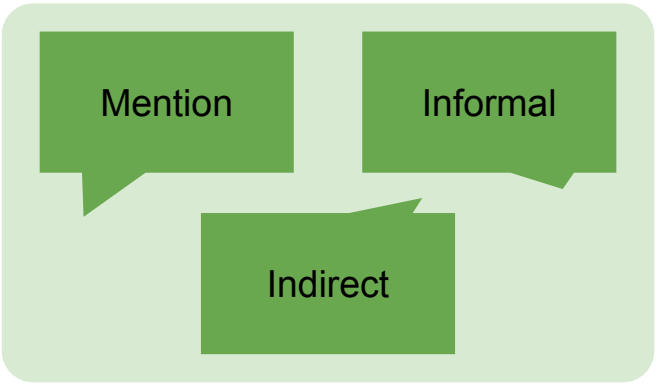
Neil Byers

BRIC 2026
June 4, 2026

Introduction: Data Citations



Structured



Unstructured

Crossref

MAKE DATA COUNT.

NCBI

DataCite
FIND, ACCESS, AND REUSE DATA

ICPSR INSTITUTE FOR SOCIAL RESEARCH UNIVERSITY OF MICHIGAN

Introduction: Data Citations

Entity Representation

dataset	publication
rs2431697	https://doi.org/10.109
H0YGJ7	https://doi.org/10.339
https://doi.org/10.10	https://doi.org/10.109
X67681.1	https://doi.org/10.118

Datasets ↔ Publications

Actual Entity

The BLAST average nucleotide identity (ANIb) values between *Gordonia* sp. JH63 and the most closely related genomes on the databases were calculated using the JSpeciesWS server (10). JH63 shared the highest ANIb with *Gordonia* sp. strain UCD-TK1 (GenBank accession number LZMP00000000) (11), *Gordonia* sp. strain SGD-V-85 (LNQO01000000), and *Gordonia* sp. strain IITR100 (MVKV01000000) (12) at values of 97.31%, 97.12%, and 97.00%, respectively. These values were derived from 89.67, 88.99, and 88.84 aligned percentages, respectively. This result indicated that the ANIb values are above the suggested cutoff (95%) for the delineation of new bacterial species (13), and these strains may form a new species clade of *Gordonia*. Of them, *Gordonia* sp. JH63 represents the first complete

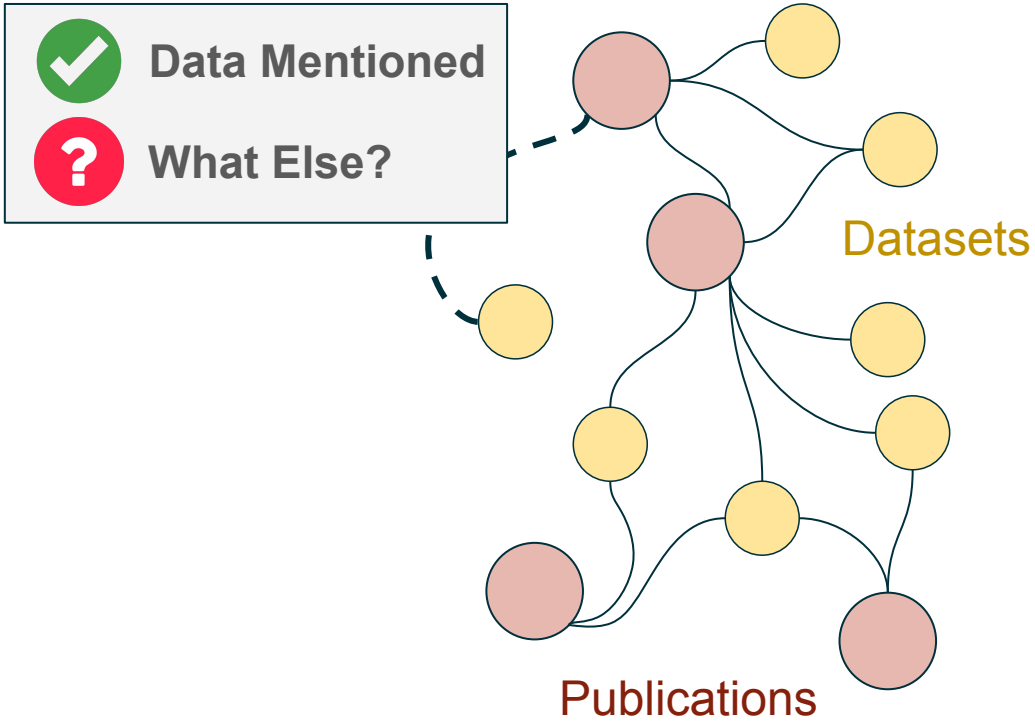
Findings

Datasets

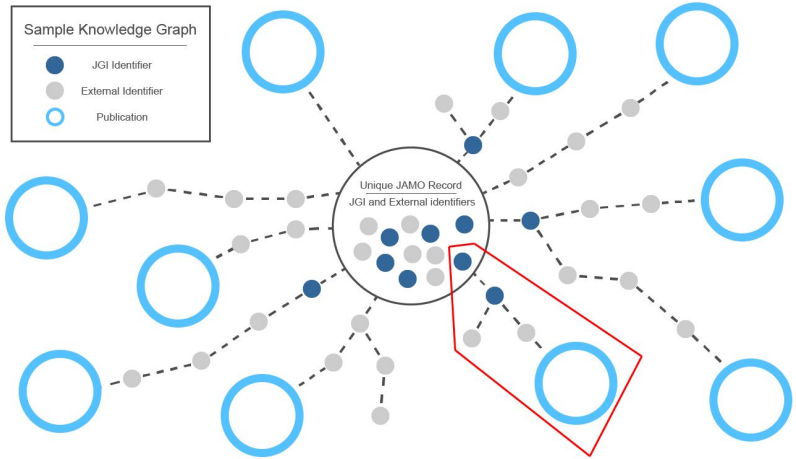
Algorithms, tools, and statistics

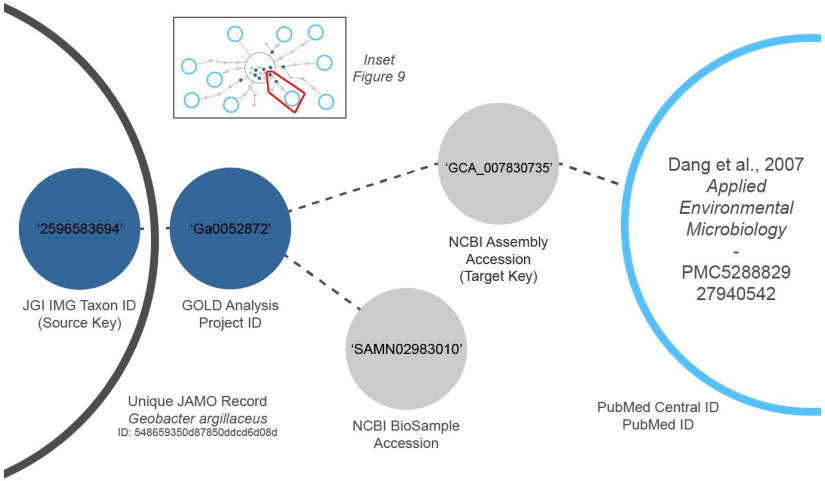
Question: How is data used?

- **In Scope**
  - ***Is*** the data used?
  - ***How*** is it being used?
- **Out of Scope**
  - ***Why*** is the data being used?



Corpus: JGI's Data Citation Explorer





~40,000 publications

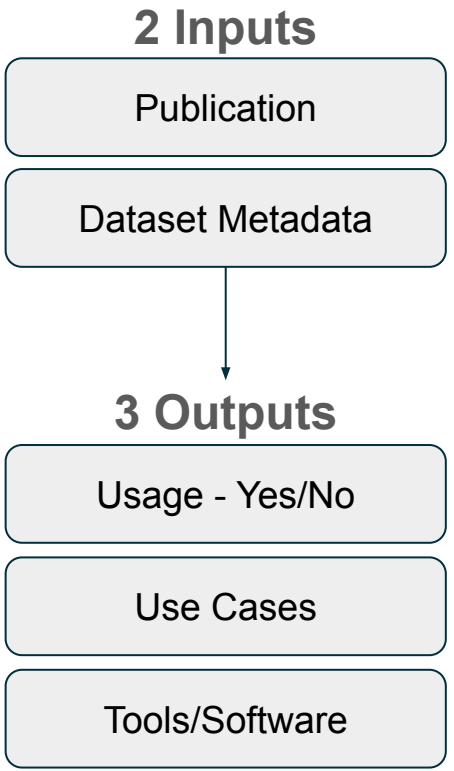
linked to

~50,000 data identifiers

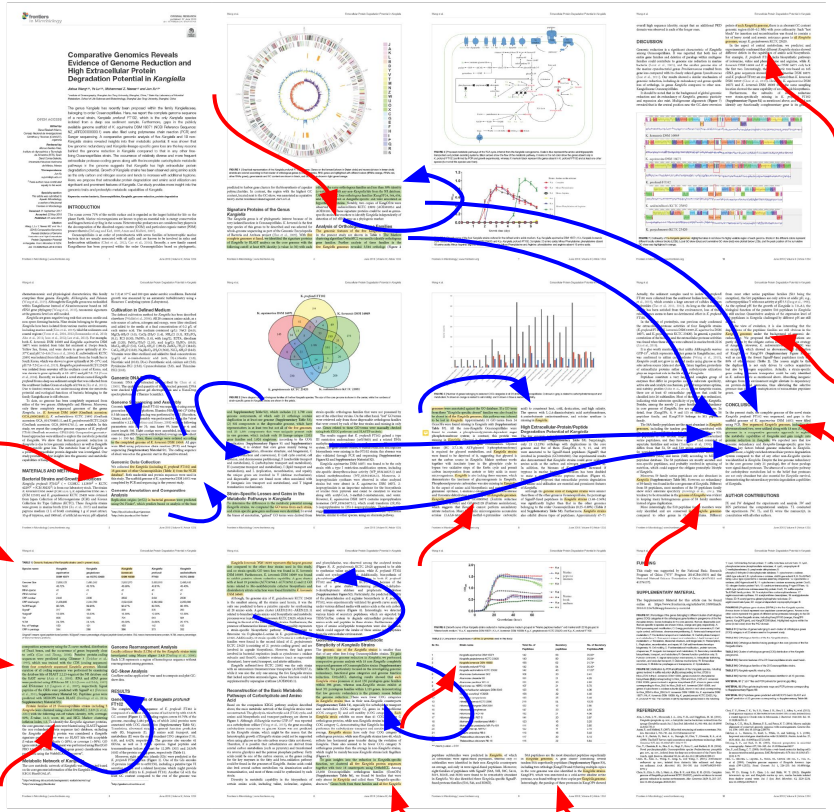
- Full text from PMC XML
- Plain text extracted and transformed using a Java client library
- Front/back matter removed

The Problem: Why LLMs?

- **What we want to know:**
  - ***Is*** the data used?
    - Not all citations indicate use
  - ***How*** is it being used?
    - Nearly infinite possible answers
  - ***Which tools/software*** were involved in that use?
    - In bioinformatics, can tell a lot about usage on its own



The Problem: Why LLMs?



Publication

Wang J, et al. (2018) *Front. Microbiol. 9:1224*. **10.3389/fmicb.2018.01224**

Dataset

Kangiella koreensis DSM 16069 Genome assembly ASM2408v1 - **GCA_000024085.1**

Mentions of dataset

Descriptions of usage

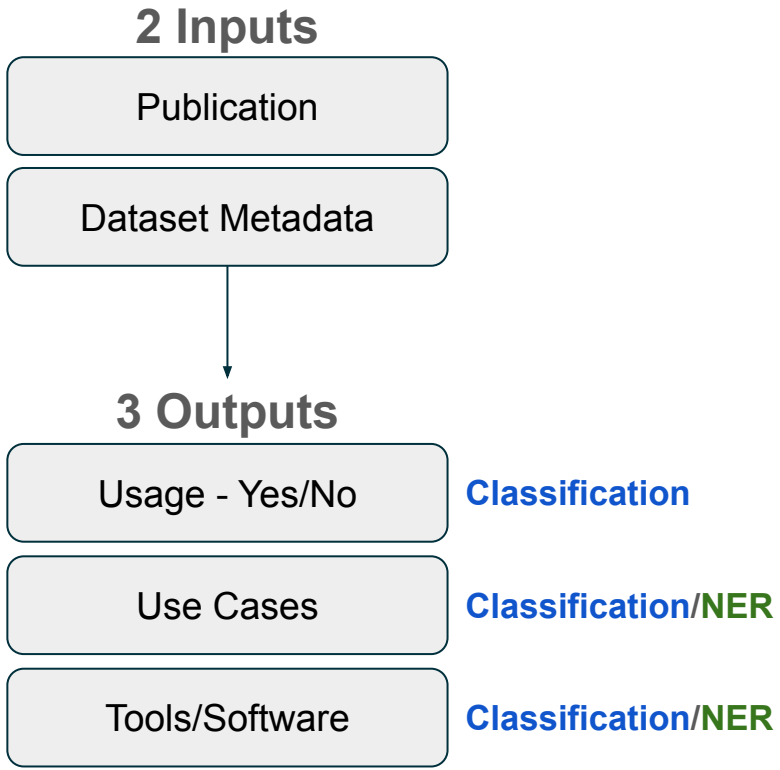
The Problem: Why LLMs?

- **Data Use (Yes/No):**
  - ***Yes***
- **Use Cases**
  - The assembled genome of K. koreensis DSM 16069 was **used as a reference in the annotation** of the newly sequenced K. profundi FT102 genome.
  - The assembled genome of K. koreensis DSM 16069 was **used in wide-ranging comparative genomic analysis** between 5 Kangiella and 18 non-Kangiella Oceanospirillales strains.

- **Tools/Software:**
  - ***Ori-Finder***
  - ***Glimmer 3***
  - ***BLAST 2.2.25+***
  - ***BLASTP***
  - ***tRNAscan-SE 1.3.1***
  - ***RNAmmer 1.2***
  - ***SignalP 4.1***
  - ***OrthoMCL 2.0.9***
  - ***Blast2GO PRO 3.0***
  - ***WebMGA online server***
  - ***Mauve aligner 2.4.0***
  - ***GenSkew***

The Problem: Why LLMs?

- **The Nail**
  - Extracting things from a large text
  - Matching them against pre-labeled examples
  - Binary classification on the whole article
- **The Right Hammer**
  - Pre-training essential. Training set impractical
  - No predefined classes
  - Multiple class assignments per document
  - Heuristics impractical given variation in source data
  - Large semantic context critical



The Problem: Why LLMs?

A difficult nail requires the right hammer

	GPT (LLMs)	BERT / "Long-BERT"	TextCNN	Traditional Classifiers (Bayes, Regression, SVMs, etc.)
Context Size	Large	Small-Medium	Very Small	NA
Pretrained	Yes	Yes	Yes	No
Classification	Yes	Yes	Yes	Yes
Entity Extraction	Yes	Yes	No	No
Requires Heuristics	No-ish	Maybe	Maybe	Maybe
External Info Accepted	Yes	Maybe	No	NA

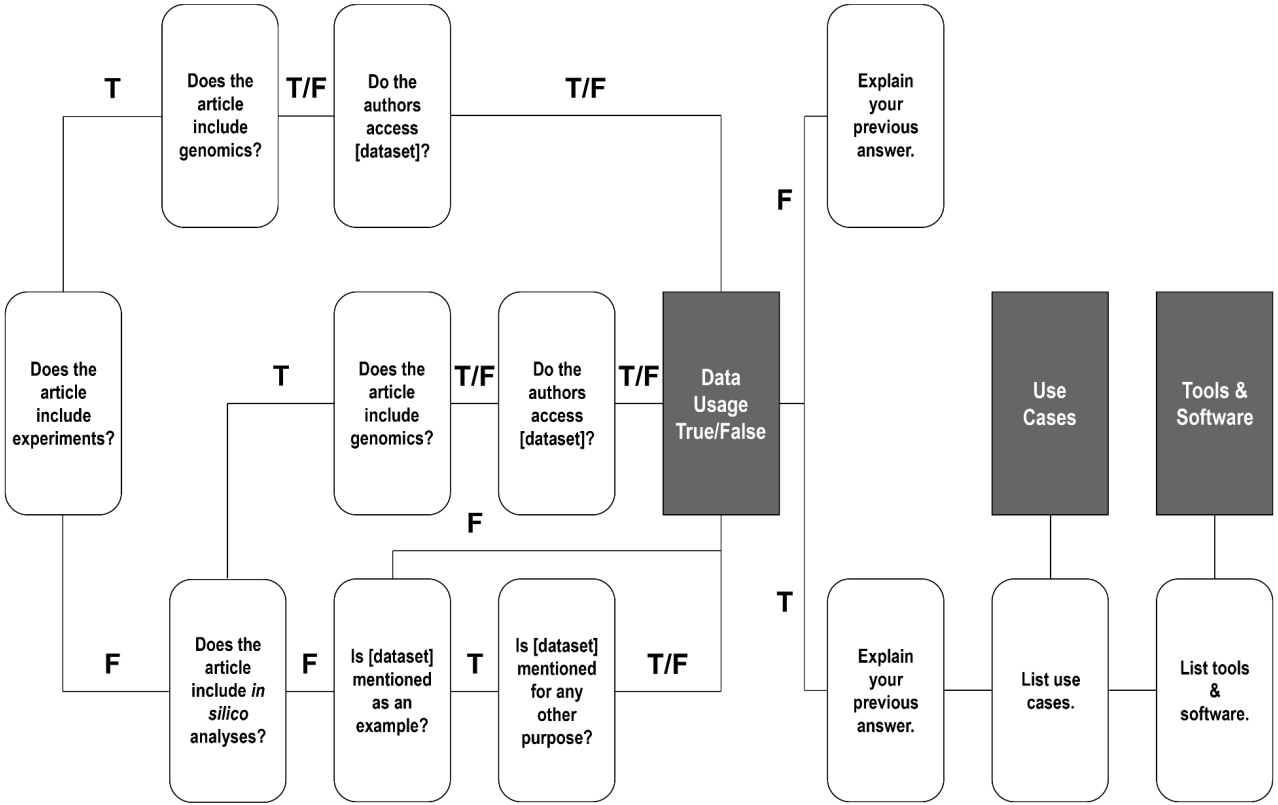
Hardware, Software, & Parameters

JGI JOINT GENOME INSTITUTE

Choices, choices choices…

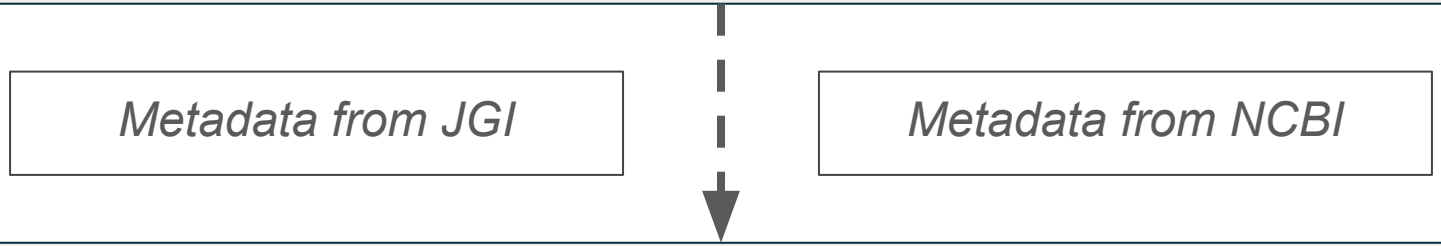
Which Model?	Open Models *or* Commercial Models	Which software?	*Python* Hugging Face OpenAI Pytorch
Where is it?	Remotely-hosted *or* Locally-hosted	Which settings?	Sampling disabled = Deterministic

Prompting & RAG



Prompting & RAG

The accession "**<ACCESSION>**" refers to a **<RECORD TYPE>** record from the **<ORGANIZATION> <DATABASE>**. The record contains **<DATA TYPE>** from **<SPECIES-STRAIN NAME>** (a **<ORGANISM TYPE>** from the species **<SPECIES>**). The sequenced strain is referred to as "**<STRAIN IDS>**".



The accession "**CP001672.1**" refers to a **Nucleotide Sequence** record from the **National Center for Biotechnical Information (NCBI) GenBank Nucleotide Database.** The record contains **nucleotide sequence data** from **Methylotenera mobilis JLW8** (a **Prokaryote** from the species **Methylotenera mobilis**). The sequenced strain is referred to as "**JLW8**".

Publication Sampling

"Initial Set"		"Evaluation Set"
22	*Publications*	20
56	*Identifiers*	34
Low	*Selection Rigor*	High
• Publisher • Age • Length • Cited ID classes	*Criteria*	• Model training date • Publisher • Age • Length • Cited ID classes

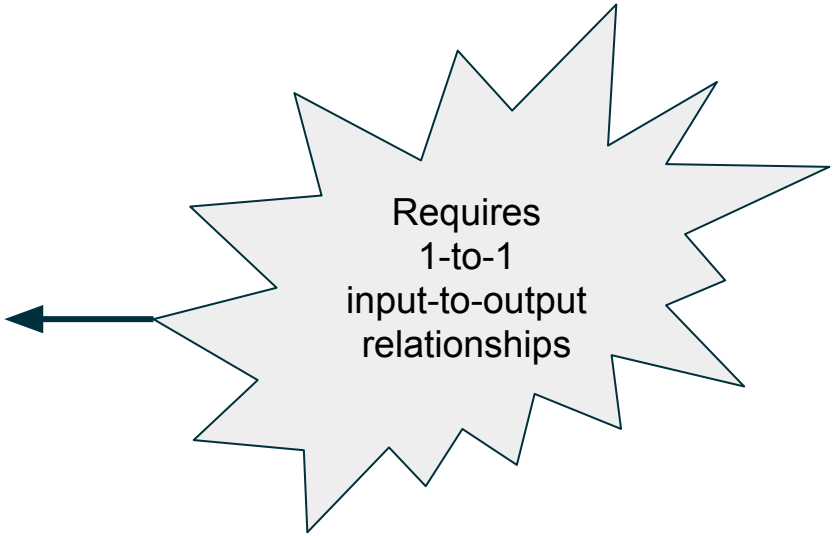
Evaluation: Manual Annotation

	Initial Set		Evaluation Set	
	Human	Machine	Human	Machine
Data Use	56	56	34	34
Use Cases	73	~200-600	49	~60-300
Tools/ Software	229	~200-600	145	~90-300
Total	358	~400-1300	228	~200-700

- Data Use: *Yes*
- Use Case: *The assembled genome of* K. koreensis *DSM 16069 was used in a comparative genomic analysis between 5* Kangiella *and 18* non-Kangiella *Oceanospirillales strains.*
- Use Case: *The assembled genome of* K. koreensis *DSM 16069 was used as a reference in the annotation of the newly sequenced* K. profundi *FT102 genome.*
- Tool/Software: *Ori-Finder*
- Tool/Software: *Glimmer 3*
- Tool/Software: *BLAST 2.2.25+*
- Tool/Software: *BLASTP*
- Tool/Software: *tRNAscan-SE 1.3.1*
- Tool/Software: *RNAmmer 1.2*
- Tool/Software: *SignalP 4.1*
- Tool/Software: *OrthoMCL 2.0.9*
- Tool/Software: *Blast2GO PRO 3.0*
- Tool/Software: *WebMGA online server*
- Tool/Software: *Mauve aligner 2.4.0*
- Tool/Software: *GenSkew*

Evaluation: Scoring

- **Precision**
  - ***How many machine outputs were 'good'?***
- **Recall**
  - ***How many of the human-labeled 'good' outputs were ID'd by the machine?***
- **F1**
  - ***How 'good' was the machine on average?***



Evaluation: Scoring

Precision, Recall and F1 Calculations for LLM outputs

Category	Human Value	Machine Value	Assessment
Data Use	TRUE	TRUE	TP
Tools & Software	BLAST		FN
		NCBI Database search function	FP
Use Cases	comparative analysis of dehydrogenase and alcohol dehydrogenase genes	comparative analysis of alcohol dehydrogenase gene counts across bacterial species	TP

	Retrieved	Not Retrieved
Relevant	2	1
Irrelevant	1	NA

- ***Precision:*** 0.66
- ***Recall:*** 0.66
- ***F1:*** 0.66

Publication
Otun SO, Ntushelo K. (2023) *Data Brief* 52:109918. doi: **10.1016/j.dib.2023.109918**.
Dataset
Methylotenera mobilis JLW8, complete genome - **CP001672**
Model
Claude Opus 4.1

Evaluation: SARGO

Selective Aggregation of Redundant and Granular Outputs

Category	Human Value	Machine Value	Assessment
Data Use	TRUE	TRUE	TP
Tools/Software	BEAST 2.4.5		FN
	CloanalFrameML		FN
	gingr		FN
	ITL		FN
	Parsnp		FN
	PhyML	PhyML	TP
	R		FN
	RAxML 8.2.11	RAxML	TP
	VariScan		FN
Use Cases	Outgroup in molecular dating analysis		FN
	Outgroup in phylogenetic analysis	Outgroup selection in phylogenetic analysis	TP - AGGREGATE
		Rooting of the phylogenetic tree	

	Retrieved	Not Retrieved
Relevant	4	8
Irrelevant	0	NA

- ***Precision:*** 1.0
- ***Recall:*** 0.33
- ***F1:*** 0.5

Publication
Frisch, M, et al. (2018) *mSphere. (3):e00571-17. doi:* ***10.1128/mSphere.00571-17***
Dataset
Staphylococcus aureus subsp. aureus COL, complete genome **CP000046.1**
Model
Llama 3.1-405B

Evaluation: SARGO

Selective Aggregation of Redundant and Granular Outputs

Category	Human Value	Machine Value	Assessment
Data Use	TRUE	TRUE	TP
Tools/Software	COG		FN
	BLAST		FN
	MUMmer		FN
		GLIMMER	FP
Use Cases	Used in a comparative study of Brucella genomes to investigate veterinary pathogenicity	comparative analysis of genomic features	TP - AGGREGATE
		comparative bestmatch blastp searches	
		suffix tree analysis using MUMmer	

	Retrieved	Not Retrieved
Relevant	2	3
Irrelevant	1	NA

- ***Precision:*** 0.66
- ***Recall:*** 0.4
- ***F1:*** 0.5

Publication
Tsolis r, et al. (2009) *PLoS One. 2009;4(5):e5519. doi:* ***10.1371/journal.pone.0005519***
Dataset
Brucella melitensis bv. 1 str. 16M chromosome I, complete sequence **NC_003317.1**
Model
Llama 3.1-405B

Results: Initial vs. Evaluation Sets

Response Category	Data Used Y/N		Use Cases		Tools & Software		Overall	
Publication Set	Initial	Evaluation	Initial	Evaluation	Initial	Evaluation	Initial	Evaluation
Recall	1	1	0.822	0.653	0.686	0.476	0.76	0.579
Precision	1	0.964	0.845	0.8	0.813	0.758	0.851	0.805
F1 Score	1	0.982	0.833	0.719	0.744	0.805	0.805	0.674

Model: Llama 3.1-405B

- Lower recall against evaluation set than against the initial set
- Lower F1 against evaluation set than against the initial set, mostly due to recall disparity
- Comparable scores on data use assessment across both sets
- Slightly lower precision against the evaluation set as compared with the initial set

Results: Initial vs. Evaluation Sets

Response Category	Data Used Y/N		Use Cases		Tools & Software		Overall	
Publication Set	Initial	Evaluation	Initial	Evaluation	Initial	Evaluation	Initial	Evaluation
Recall	1	1	0.822	0.653	0.686	0.476	0.76	0.579
Precision	1	0.964	0.845	0.8	0.813	0.758	0.851	0.805
F1 Score	1	0.982	0.833	0.719	0.744	0.805	0.805	0.674

Model: Llama 3.1-405B

- Lower recall against evaluation set than against the initial set
- Lower F1 against evaluation set than against the initial set, mostly due to recall disparity
- Comparable scores on data use assessment across both sets
- Slightly lower precision against the evaluation set as compared with the initial set

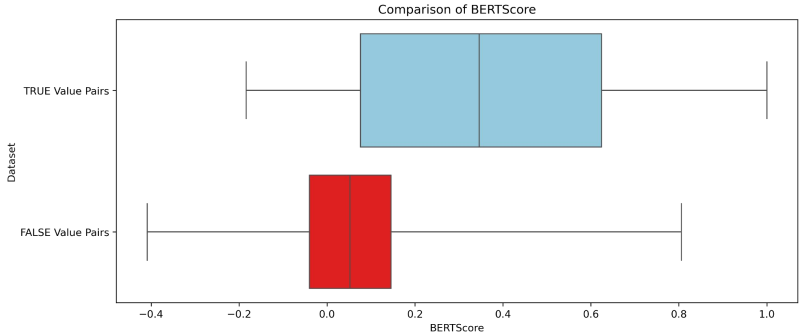
Results: Evaluation Set by Model

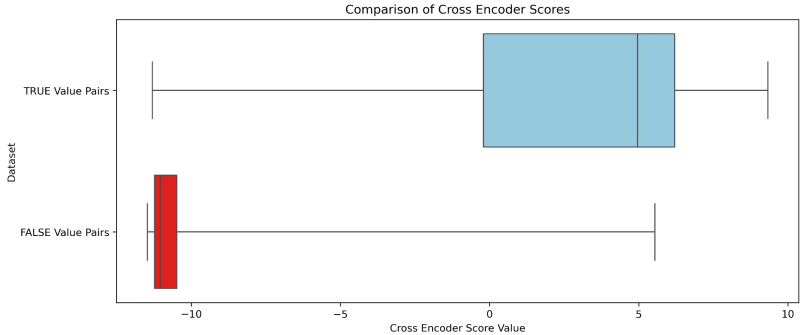
Response Category	Data Used Y/N			Use Cases			Tools & Software			Overall		
Publication Set	Llama 3.1405B	Claude Opus 4.1	Gemini Pro 3.1	Llama 3.1405B	Claude Opus 4.1	Gemini Pro 3.1	Llama 3.1405B	Claude Opus 4.1	Gemini Pro 3.1	Llama 3.1405B	Claude Opus 4.1	Gemini Pro 3.1
Recall	1	0.926	0.889	0.653	0.837	0.918	0.476	0.759	0.959	0.579	0.796	0.941
Precision	0.964	1	0.923	0.8	0.82	0.35	0.758	0.714	0.408	0.805	0.769	0.42
F1 Score	0.982	0.962	0.906	0.719	0.828	0.508	0.805	0.736	0.572	0.674	0.782	0.581

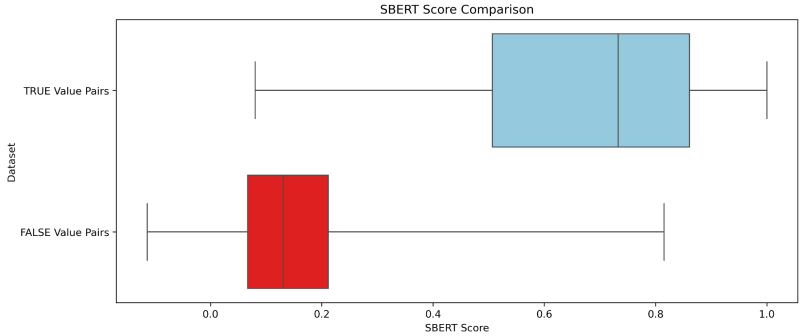
	Llama 3.1-405B	**Claude Opus 4.1**	**Gemini 3.1 Pro**
Recall	Bad	Good/Decent	Great
Precision	Good/Decent	Good/Decent	Bad

Evaluation: Scoring Automation?

Semantic Similarity: Can we standardize the evaluation component?







Use Case Value Comparisons:

- SBERT
- Cross Encoder
- BERTScore

Outputs from Model:
Claude 4.1 Opus

Discussion: Applied Implications

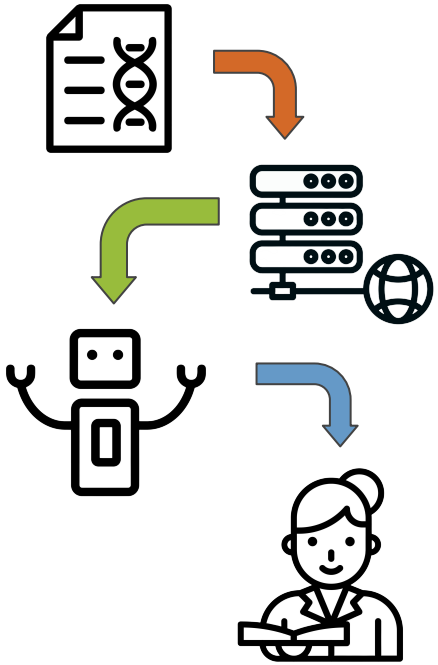
What does all of this mean for a practitioner?

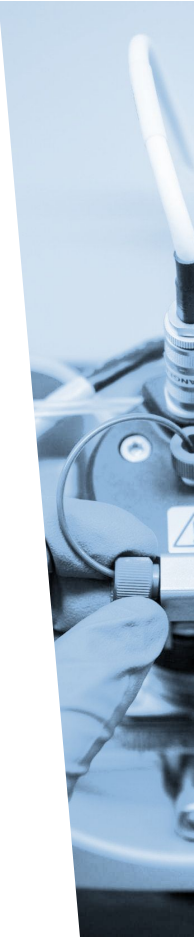
- **Recall Results - missing the things that are good**
  - Best model would still miss ~20-25% of relevant information
- **Precision results - what we did get was not all good**
  - Best model would provide 20% hallucinated responses.
- **Even the best models still need work**
  - Claude's ~.782 F1 score still falls short
- **What performance can be 'robust' changes in context**
  - 95%?
  - 75%
  - 99.9%?

Discussion: Potential Improvements

What can/can't we do to make this better?

- **Post-training / Fine-tuning**
  - Eliminates zero shot motivation
- **Chunking / lowering context window**
  - Eliminates needed context clues from across publication
- **Better RAG/Prompting**
  - Provide more examples?
  - Our workflow may not be optimal
- **Bigger/newer models?**
  - Not always better
  - Each must be prompt-engineered-to
  - Expense component
- **Agentic framework?**
  - Extra validation layers
  - Even more stochasticity



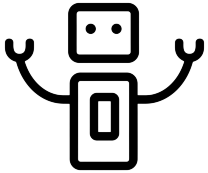


Discussion: Requirements & Cost

How much money do we actually save?

- **HPC reqs for local hosting**
  - GPUs for the largest models
- **API/token cost**
  - Anywhere from $20k to $100k to run the entire DCE corpus
  - Increases with reasoning models and agents
- **Datasets are not free / easily attainable**
  - Data citation corpora aren't available for everyone
  - Full-text access to publications is limited
- **Eval is expensive and crucial**
  - How to eval without lots of FTE hours?

Takeaways



LLMs are good
…
not perfect.

This is still expensive.

There is no silver bullet.

Evaluation is everything.

Thank you!

JGI/DCE Team:

Valerie Skye

Chris Beecroft

Ali Zaidi

Kiersten Fagnan

Further Reading:

Preprint for this work:
https://doi.org/10.48550/arXiv.2511.02936

Repo for this work:
https://github.com/npbyers0401/citation-function-analysis

DCE publication:
https://doi.org/10.1038/s41597-024-04049-7