PDMR CHAT BOT

PURDUE UNIVERSITY | The Data Mine

PURDUE UNIVERSITY | Turfgrass Science at Purdue
Department of Horticulture & Landscape Architecture

MathWorks

Team Members: Nathan Steurer, Daniel Harradine, Hruthin Muddasani, Yu-Jen Chen, John Graham Chiles, Tejomay Marathe, Vachan Arora, Ken Thai Nguyen, Madhavan Prasanna, Sharan Suthaharan

The Purpose of this Project

- This project aims to create a learned language model (LLM) to optimize the search for turf grass disease treatments using 26 years of published field trials stored in Plant Disease Management Reports (PDMRs).
- Currently, the turfgrass treatment resides in the PDMR's that must be manually searched – a cumbersome process.
- By creating an LLM with the ability to search through PDMR's and create a database of the information, we improve the accessibility of the information for stakeholders.

Our Plan

To increase the ability for an LLM to properly read the PDMR's so it can give accurate responses we split the team into three different groups.

- Parsing – First, for the LLM to more easily read the PDMR's the files needed to be more standardized.
- Database – Next, the PDMR's need to be stored in a data so the LLM could easily search for relevant PDMR's needed for the response.
- LLM – The LLM itself needed to be engineered in a way that it would look for not only the correct treatment but the most recent one.

Research Into The Project

Parsing of the PDMRs

Throughout this semester the parsing team looked at three different ways of parsing the PDMR's:

- Custom Computer Vision (CV) solution, splitting the work into five stages: Fetching, Parsing, Data Cleaning 1, Data Cleaning 2, and Output generation.
- Matlab's Text Analytics Toolbox
- ChatGPT to parse the PDMRs and transform tables into CSV files.

Storing The PDMR's

The best form for storing the CSV files was on a vector database. Which would allow the LLM to more easily find relevant files to any question. This was done through a myriad of tools.

- Docker
- VS Code.
- Milvus

The first step was to code a vector database. Then to encode the data into the database. The final step was to create indicator's allowing for files to be queried faster.

Training the LLM

After much deliberation between open-source or closed-source, we eventually choose an open-source because of the reduced cost.

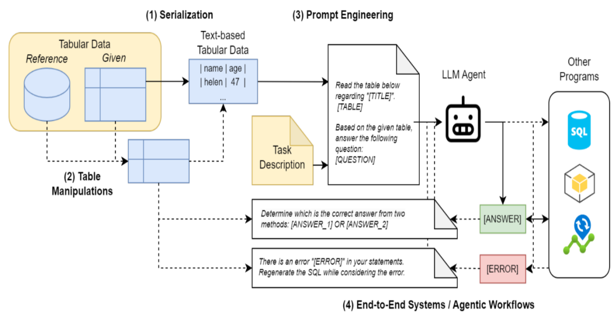
- The LLM that we settled with was lama 3.2. We were able to more easily prompt engineer the LLM by using Ollama to run it without many limits.
- To decrease the amount of time it takes the LLM to respond we had Purdue's Anvil host the Ollama where we could import our desired LLM.

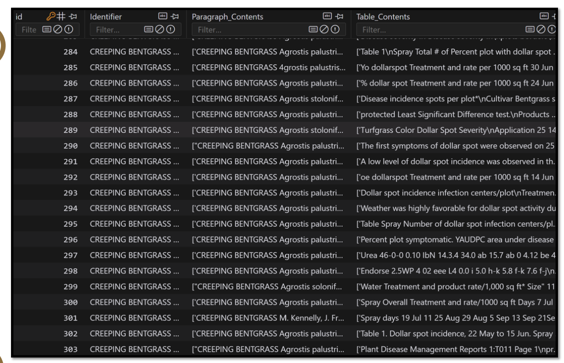
Parsing Team: 1

- After handling text extraction from PDMRs with Computer Vision, we clean it up and separate the paragraphs
- Trained the learning algorithm of ChatGPT to build accuracy
- Used a pipeline into ChatGPT from Python to handle different table formats in PDMRs

Database Team: 2

- To improve LLM performance, we created a vector database for retrieval augmented-generation (RAG)
- Hosted the vector and SQL database of PDMR's on Docker and Milvus as backend
- Training LLM using unsupervised learning on vector database
- Establishing public hosting on Anvil for easy connection/ setup





id	Identifier	Paragraph_Contents	Table_Contents
284	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['Table 1\nSpray Total # of Percent plot with dollar spot .
285	CREEPING BENTGRASS ...	['CREEPING BENTGRASS 4grostis palustris...	['Yo dollarspot Treatment and rate per 1000 sq ft 30 Jun
286	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['% dollar spot Treatment and rate per 1000 sq ft 24 Jun
287	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis stolonif...	['Disease incidence spots per plot*\nCultivar Bentgrass s
288	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['protected Least Significant Difference test.\nProducts ..
289	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis stolonif...	['Turfgrass Color Dollar Spot Severity\nApplication 25 14
290	CREEPING BENTGRASS ...	["CREEPING BENTGRASS Agrostis palustri...	['The first symptoms of dollar spot were observed on 25
291	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['A low level of dollar spot incidence was observed in th.
292	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['oe dollarspot Treatment and rate per 1000 sq ft 14 Jun
293	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['Dollar spot incidence infection centers/plot\nTreatmen.
294	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['Weather was highly favorable for dollar spot activity du
295	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['Table Spray Number of dollar spot infection centers/pl.
296	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['Percent plot symptomatic. YAUDPC area under disease
297	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['Urea 46-0-0 0.10 IbN 14.3.4 34.0 ab 15.7 ab 0 4.12 be 4
298	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['Endorse 2.5WP 4 02 eee L4 0.0 i 5.0 h-k 5.8 f-k 7.6 f-j\n.
299	CREEPING BENTGRASS ...	["CREEPING BENTGRASS Agrostis solonif...	['Water Treatment and product rate/1,000 sq ft* Size" 11
300	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['Spray Overall Treatment and rate/1000 sq ft Days 7 Jul
301	CREEPING BENTGRASS ...	['CREEPING BENTGRASS M. Kennelly, J. Fr...	['Spray days 19 Jul 11 25 Aug 29 Aug 5 Sep 13 Sep 21Se
302	CREEPING BENTGRASS ...	['CREEPING BENTGRASS Agrostis palustri...	['Table 1. Dollar spot incidence, 22 May to 15 Jun. Spray
303	CREEPING BENTGRASS ...	["CREEPING BENTGRASS Agrostis palustri...	['Plant Disease Management Reports 1:T011 Page 1\npr.

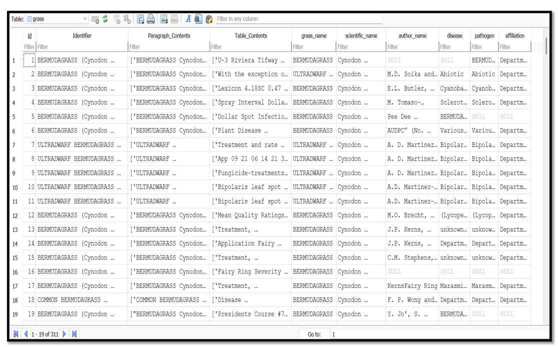


Table: grass

	id	Identifier	Paragraph_Contents	Table_Contents	grass_name	scientific_name	author_name	disease	pathogen	affiliation
1	1	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['U-3 Riviera Tifway ...	BERMUDAGRASS	Cynodon ...	NULL	NULL	BERMUD...	Departm...
2	2	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['With the exception o...	ULTRADWARF ...	Cynodon ...	M.D. Soika and...	Abiotic	Abiotic	Departm...
3	3	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Lexicon 4.18SC 0.47 ...	BERMUDAGRASS	Cynodon ...	E.L. Butler, ...	Cyanoba...	Cyanob...	Departm...
4	4	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Spray Interval Dolla...	BERMUDAGRASS	Cynodon ...	M. Tomaso-...	Sclerot...	Sclero...	Departm...
5	5	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Dollar Spot Infectio...	BERMUDAGRASS	Cynodon ...	Pee Dee ...	BERMUDA...	NULL	NULL
6	6	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Plant Disease ...	BERMUDAGRASS	Cynodon ...	AUDPC" (No. ...	Various...	Variou...	Departm...
7	7	ULTRADWARF BERMUDAGRASS ...	['ULTRADWARF ...	['Treatment and rate ...	ULTRADWARF ...	Cynodon ...	A. D. Martinez...	Bipolar...	Bipola...	Departm...
8	8	ULTRADWARF BERMUDAGRASS ...	['ULTRADWARF ...	['App 09 21 06 14 21 3...	ULTRADWARF ...	Cynodon ...	A. D. Martinez...	Bipolar...	Bipola...	Departm...
9	9	ULTRADWARF BERMUDAGRASS ...	['ULTRADWARF ...	['Fungicide-treatments...	ULTRADWARF ...	Cynodon ...	A. D. Martinez...	Bipolar...	Bipola...	Departm...
10	10	ULTRADWARF BERMUDAGRASS ...	['ULTRADWARF ...	['Bipolaris leaf spot ...	ULTRADWARF ...	Cynodon ...	A.D. Martinez-...	Bipolar...	Bipola...	Departm...
11	11	ULTRADWARF BERMUDAGRASS ...	['ULTRADWARF ...	['Bipolaris leaf spot ...	ULTRADWARF ...	Cynodon ...	A.D. Martinez-...	Bipolar...	Bipola...	Departm...
12	12	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Mean Quality Ratings...	BERMUDAGRASS	Cynodon ...	M.O. Brecht, ...	(Lycope...	(Lycop...	Departm...
13	13	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Treatment, ...	BERMUDAGRASS	Cynodon ...	J.P. Kerns, ...	unknown...	unknow...	Departm...
14	14	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Application Fairy ...	BERMUDAGRASS	Cynodon ...	J.P. Kerns, ...	Departm...	Depart...	Departm...
15	15	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Treatment, ...	BERMUDAGRASS	Cynodon ...	C.M. Stephens,...	unknown...	unknow...	Departm...
16	16	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Fairy Ring Severity ...	BERMUDAGRASS	Cynodon ...	NULL	NULL	NULL	NULL
17	17	BERMUDAGRASS (Cynodon ...	['BERMUDAGRASS Cynodon...	['Treatment, ...	BERMUDAGRASS	Cynodon ...	KernsFairy Ring	Marasmi...	Marasm...	Departm...
18	18	COMMON BERMUDAGRASS ...	['COMMON BERMUDAGRASS ...	['Disease ...	BERMUDAGRASS	Cynodon ...	F. P. Wong and...	Departm...	Depart...	Departm...
19	19	BERMUDAGRASS (Cynodon ...	["BERMUDAGRASS Cynodon...	['Presidents Course #7...	BERMUDAGRASS	Cynodon ...	Y. Jo', S. ...	BERMUDA...	NULL	NULL

1 - 19 of 311 Go to: 1

User Interface

- Created user interface (UI) using pre-existing tooling in order to achieve professional results akin to the DeepSeek or ChatGPT interface.
- Utilized tools such as AI-SDK along with Vercel in order to simplify and speed up deployment process, while also enabling easy switching of LLMs if necessary.

LLM Team: 3

- Used database to query the LLM using RAG.
- Fine-tuning and prompt engineering to ensure correct output.
- Researched state of the art methods to handle tabular data as well as textual data.
- Conducted exhaustive testing with a set question bank in order to verify outputs

Conclusion

- We were able to increase much in the two semesters that we worked on this project. Finding the many different challenges working with LLM's.
- There are many different solutions to the problems that we faced. Leading multiple branched methods of parsing, storing, and reading through the LLM.
- However, in the end we were able to successfully parse the PDMR's into readable CSV files. Also being able to import them into a vector datable using Milvus. With the final out from the LLM at least being more capable compared to the previous version.

Future Goals

- The most pressing goal would be continuing to improve the way the PDMR's can be parsed in a more efficient method, handling edge cases such as tables spanning multiple pages, or non-standardized formats.
- Prompt engineer the LLM to give better responses and reduce the chance that the LLM gives a false response or hallucinates.
- Generalize the solution to account for other types of PDMRs, not just turfgrass.
- Fine-tune response times by resolving bottlenecks in processing and inference.
- Enable LLM to help perform data analysis and other tasks on the PDMRs as directed by the user.

Treatment and rate/1000 ft^2	'U-3'		'Riviera'		'Tifway 419'	
	Phytotoxicity[Z]	Quality[Y]	Phytotoxicity	Quality	Phytotoxicity	Quality
Non-fungicide Treated Control[X]	0.88 D	8.57 A	1.37	8.13	1.50	8.25
Trinity 1.67 SC 0.5 fl oz......................	1.13 CD	8.44 AB	1.63	8.13	1.50	8.25
Trinity 1.67 SC 0.75 fl oz....................	1.13 CD	8.44 AB	1.63	8.13	1.75	8.13
Trinity 1.67 SC 1.0 fl oz......................	1.00 D	8.50 A	1.63	8.00	1.75	8.13
Trinity 1.67 SC 1.5 fl oz......................	1.38 BC	8.31 BC	1.50	8.31	2.00	8.00
Trinity 1.67 SC 2.0 fl oz......................	1.13 CD	8.44 AB	1.63	8.13	1.75	8.13
Banner MAXX 1.24 MEC 2.0 fl oz......	2.13 A	7.81 D	1.63	8.13	2.00	8.00
Bayleton 50 WG 2.0 oz	2.13 A	7.88 D	1.63	8.13	2.00	8.00
Eagle 40 WP 1.2 oz	1.63 B	8.19 C	1.50	8.13	2.00	8.00

[Z] Phytotoxicity was assessed on 7-Aug and based on a scale of 0 - 10 where 0 = no change, 1 = slight change in turfgrass appearance, 5 = distinct change in turfgrass appearance, 10 = very abnormal turfgrass. Means followed by the same letter are not significantly different according to Fisher's test of protected least significant difference where; LSD=0.51; *P*-value=0.02.
[Y] Turfgrass quality was assessed on 7-Aug and based on a scale of 1 - 9 where 1 = no turf present, 5 = unacceptable turfgrass, 7 = acceptable turf, 9 = dense, dark color, thick stand of turfgrass. Means followed by the same letter are not significantly different according to Fisher's test of protected least significant difference where; LSD=0.23; *P*-value=0.01.
[X] Fungicide application was the sub-plot effect of the split-block experimental design.

Acknowledgements

We would like to thank Dr. Miller, Gen Sasaki, Carmen Lee, and Sathvik Hegde for all the help on this project. We would also like to thank the staff at Data Mine team with all the support they have gave us.

References

Ollama. (n.d.). *Ollama Documentation.* Retrieved from https://github.com/ollama/ollama/blob/main/docs/README.md
Milvus. (n.d.). *Milvus Documentation.* Retrieved from https://milvus.io/docs
All the various turfgrass related PDMRs.
Sui, Y., Zhou, M., Zhou, M., Han, S., & Zhang, D. (2024, March). Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. *The 17th ACM International Conference on Web Search and Data Mining (WSDM '24).*