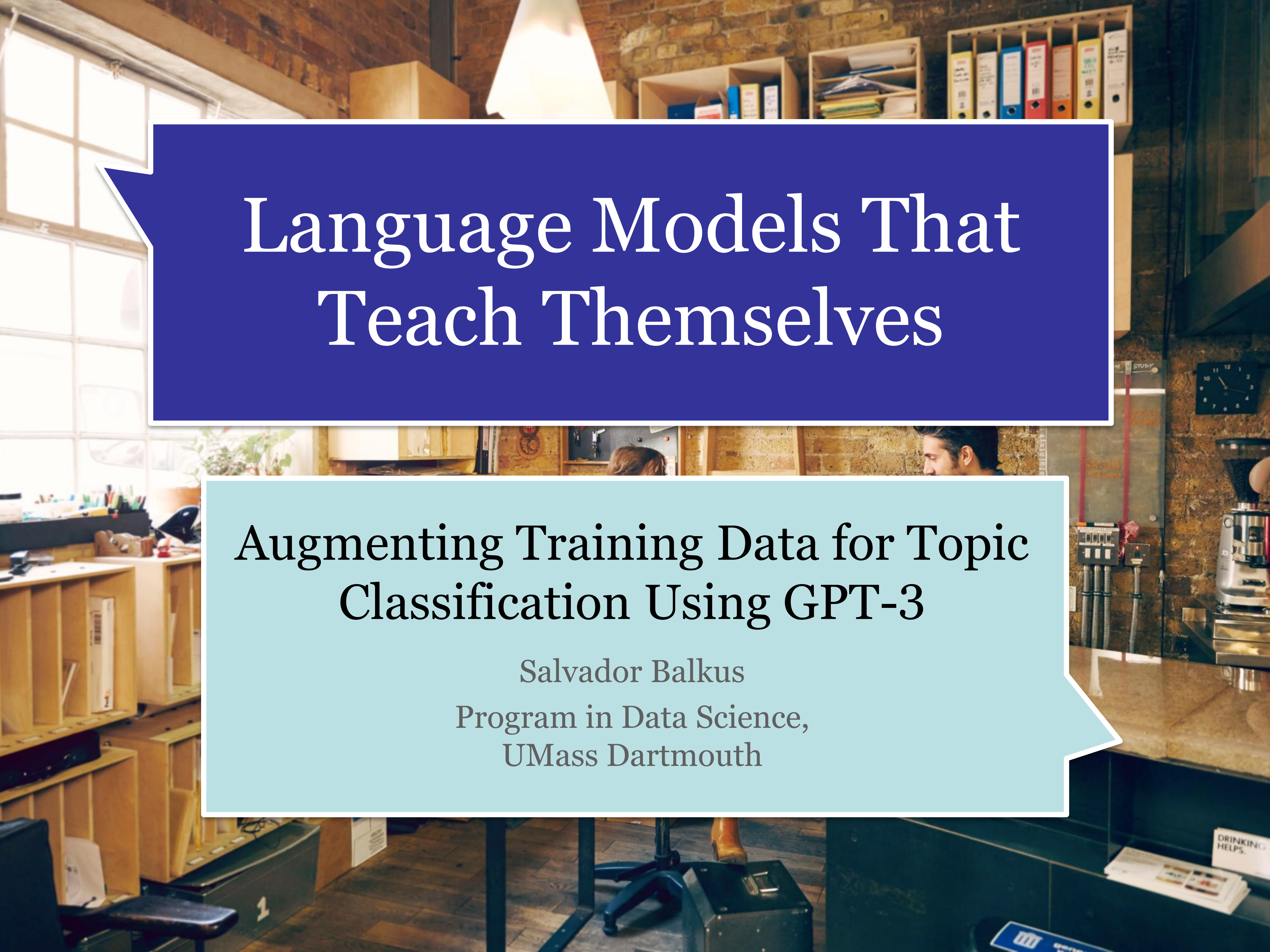


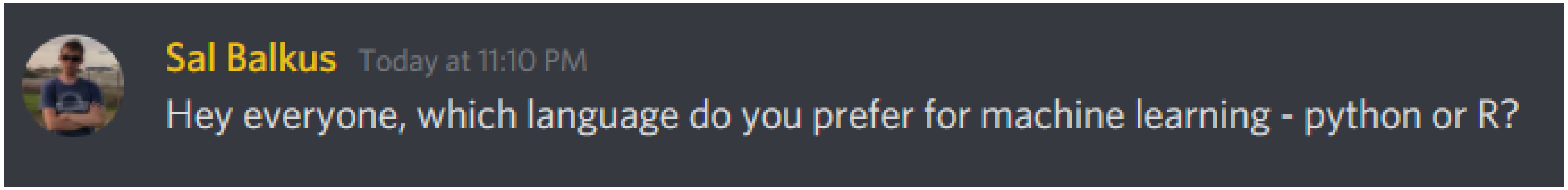
Language Models That Teach Themselves

Augmenting Training Data for Topic Classification Using GPT-3

Salvador Balkus

Program in Data Science,
UMass Dartmouth

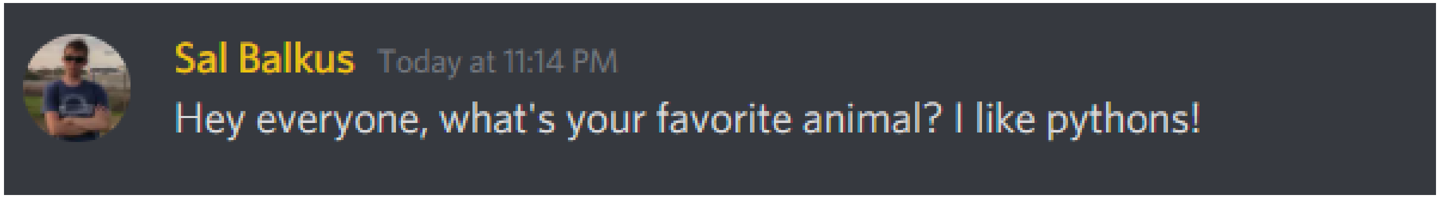
Short-Text Topic Classification



Sal Balkus Today at 11:10 PM

Hey everyone, which language do you prefer for machine learning - python or R?

Topic: Data Science



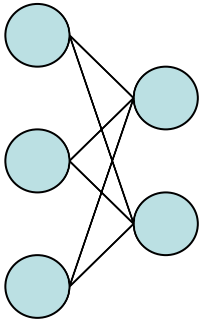
Sal Balkus Today at 11:14 PM

Hey everyone, what's your favorite animal? I like pythons!

Topic: Other

- Gathered 72 questions from UMass Dartmouth Big Data Club Discord Server

GPT-3: A Transfer Learning NLP Model [1]



Large Transformer Model (~175 billion parameters)

WIKIPEDIA
The Free Encyclopedia

Train on Common Crawl Dataset (internet text)

```
It’s a bird!
No, it’s a
plane! No,
it’s ____
```

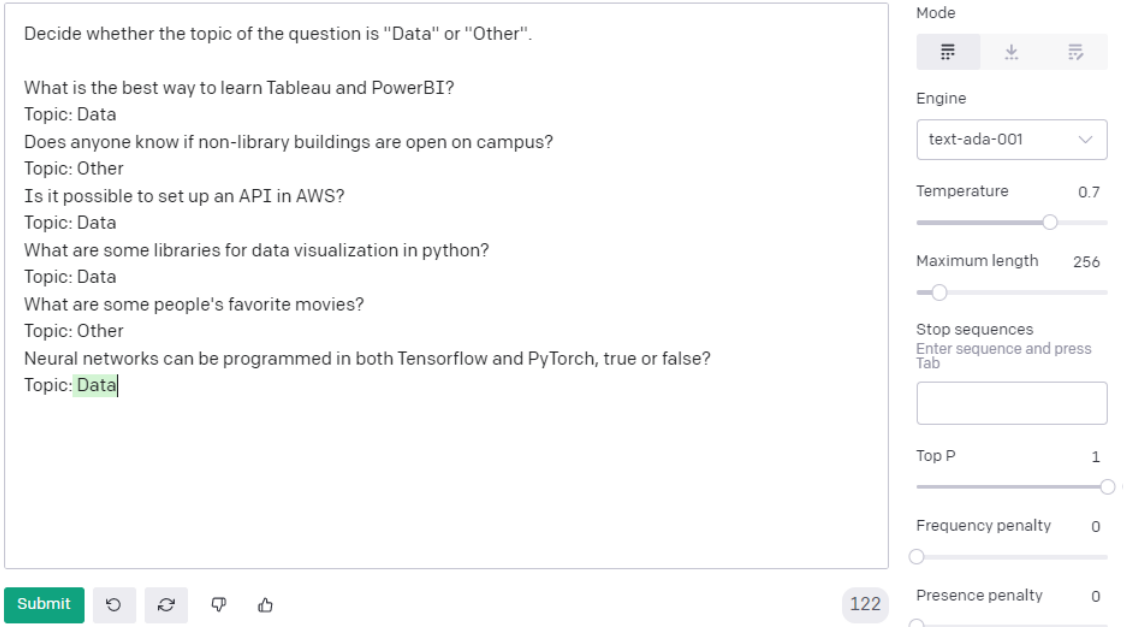
Train to predict next token in text string

Transfer to other tasks (classification, etc.)

OpenAI

`Topic: Data Science`

Completion Endpoint



Decide whether the topic of the question is "Data" or "Other".

What is the best way to learn Tableau and PowerBI?
Topic: Data
Does anyone know if non-library buildings are open on campus?
Topic: Other
Is it possible to set up an API in AWS?
Topic: Data
What are some libraries for data visualization in python?
Topic: Data
What are some people's favorite movies?
Topic: Other
Neural networks can be programmed in both Tensorflow and PyTorch, true or false?
Topic: Data

Mode
Engine: text-ada-001
Temperature: 0.7
Maximum length: 256
Stop sequences: Enter sequence and press Tab
Top P: 1
Frequency penalty: 0
Presence penalty: 0

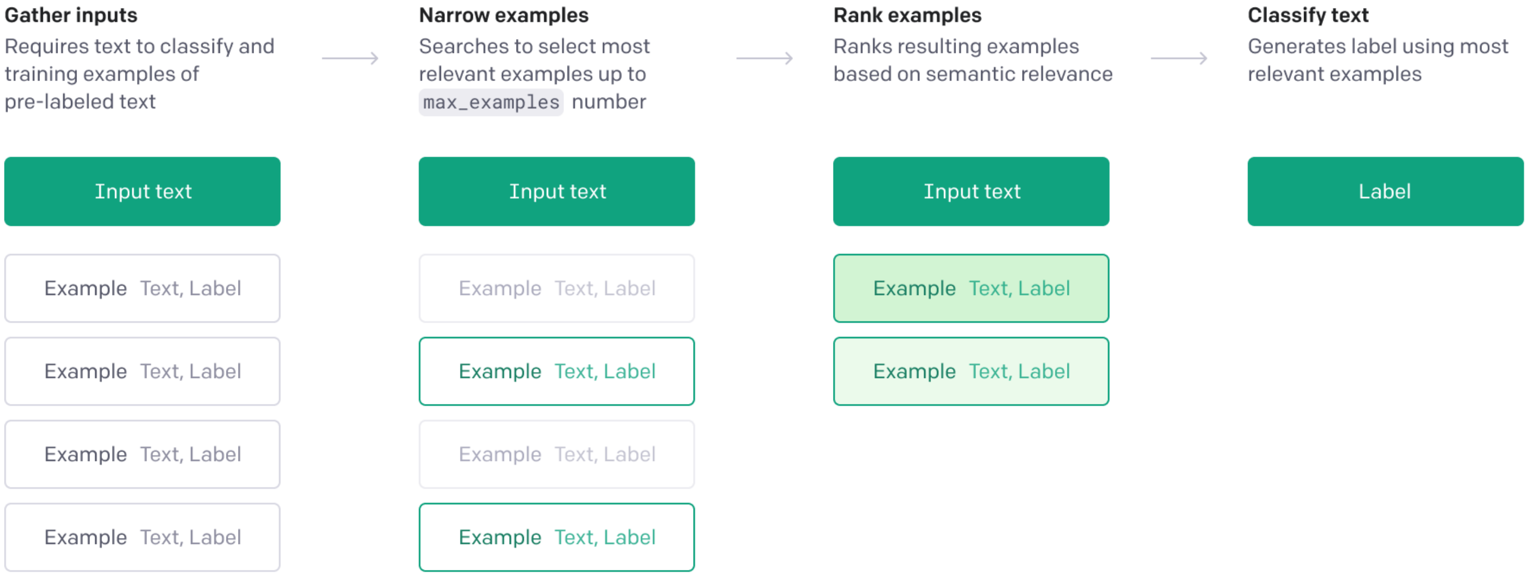
+ Only requires minimal training examples
- Sensitive to example selection

How to choose correct examples?

Classification Endpoint

+ Specifically designed for task
- Good performance requires more examples

How to use with limited data?



Gather inputs
Requires text to classify and training examples of pre-labeled text

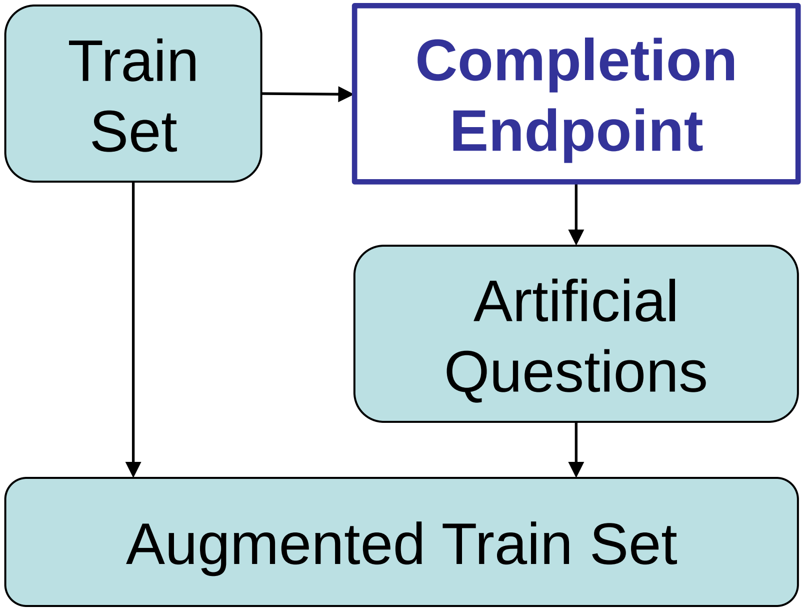
Narrow examples
Searches to select most relevant examples up to `max_examples` number

Rank examples
Ranks resulting examples based on semantic relevance

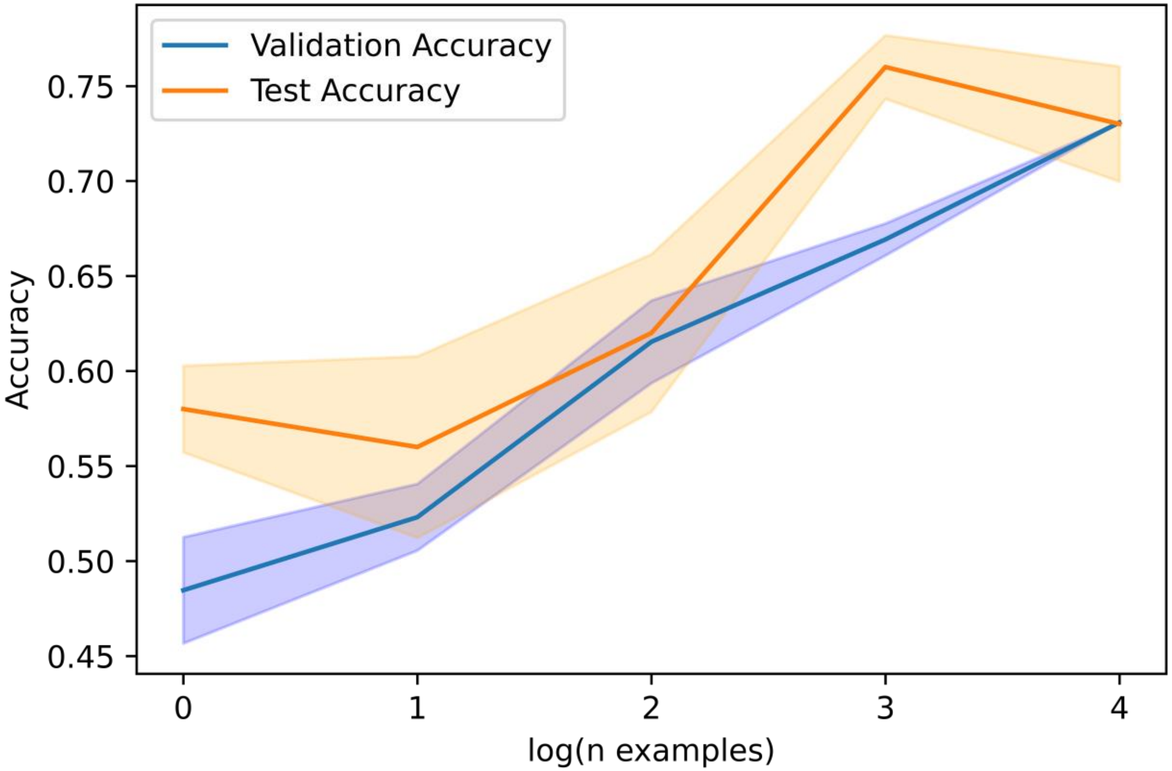
Classify text
Generates label using most relevant examples

Solution:
Augment the training data

- GPT-3 exceptional at *generating* text
- Idea: create new training examples using Completion endpoint

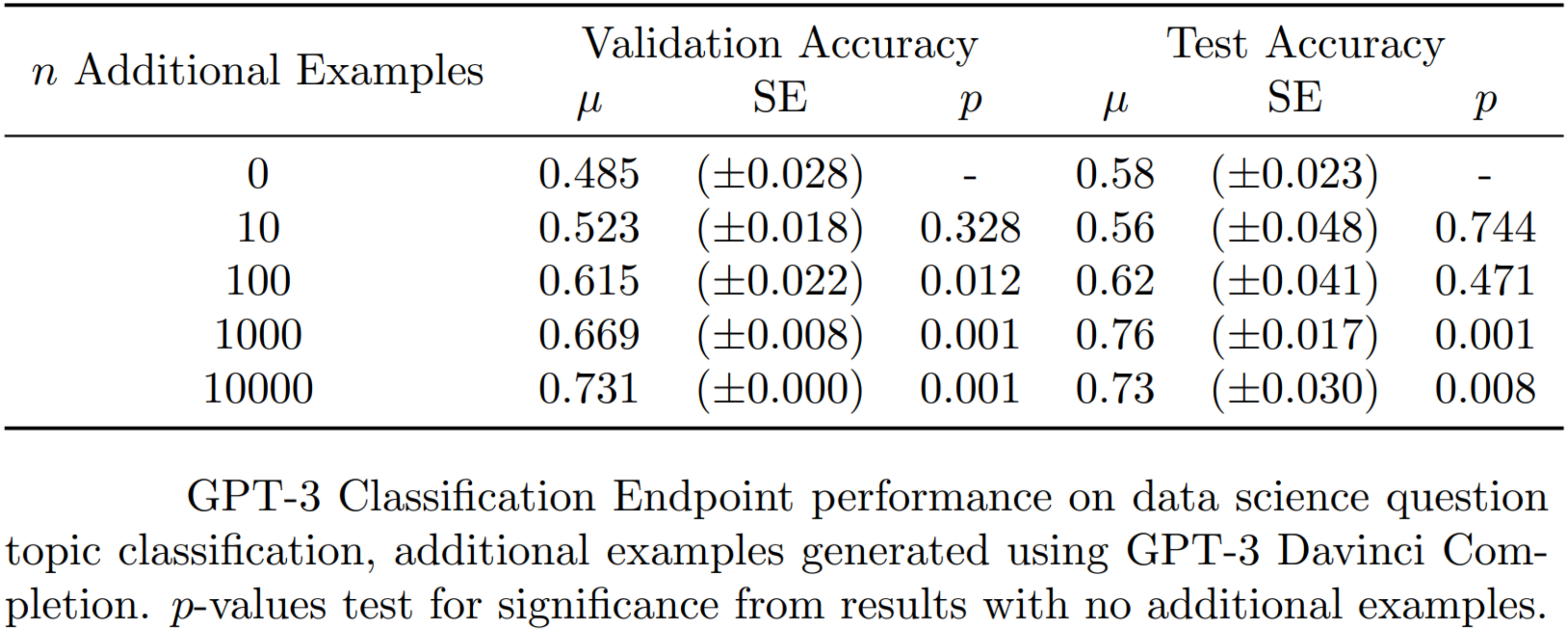


Results – Classification Endpoint



Classification Endpoint accuracy on Validation (n = 26) and Test (n = 20) question sets given different numbers of additional examples added to Train (n = 26) set. All questions posed by Big Data Club Discord Server.

Results – Classification Endpoint



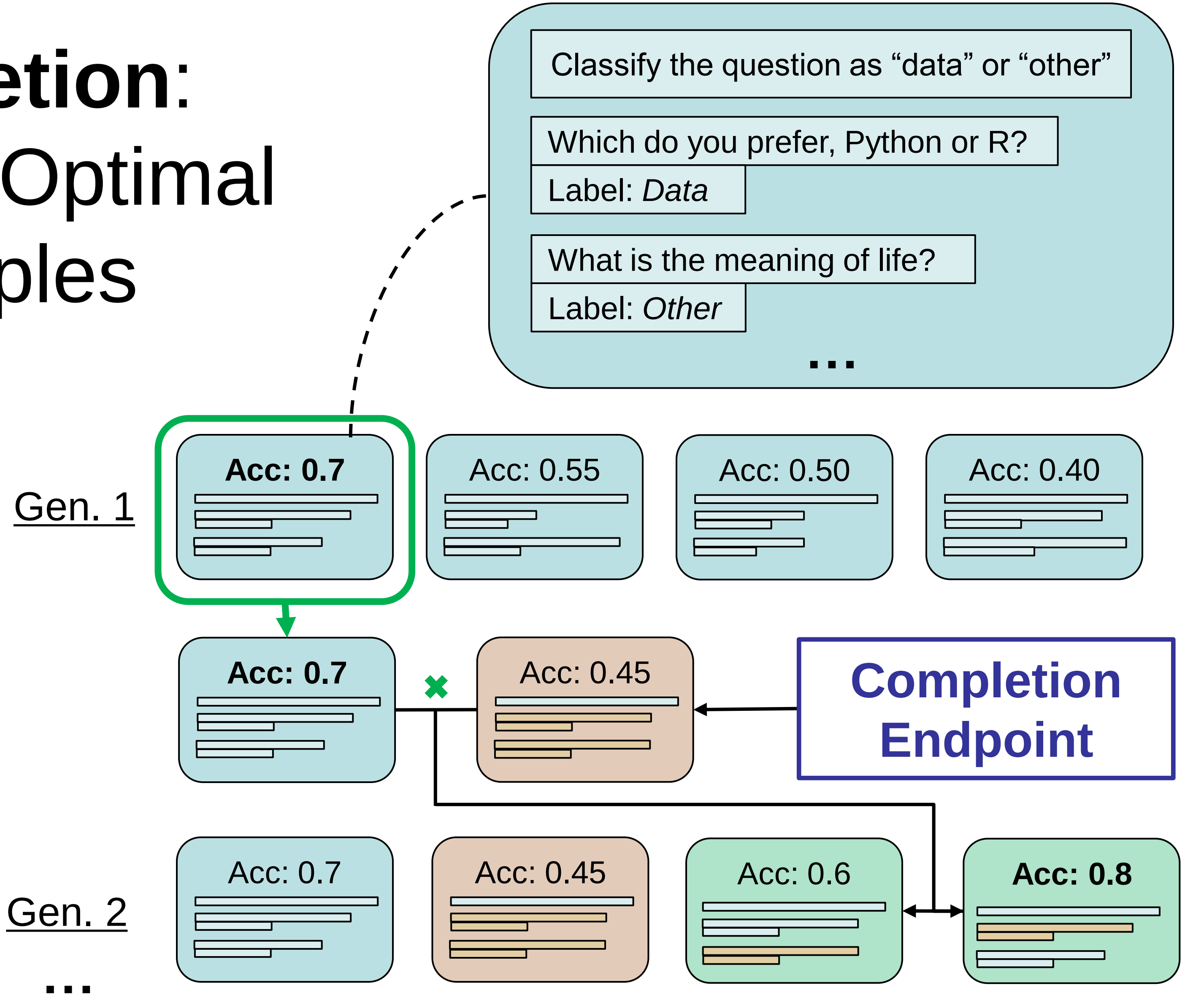
n Additional Examples	Validation Accuracy			Test Accuracy		
	μ	SE	p	μ	SE	p
0	0.485	(±0.028)	-	0.58	(±0.023)	-
10	0.523	(±0.018)	0.328	0.56	(±0.048)	0.744
100	0.615	(±0.022)	0.012	0.62	(±0.041)	0.471
1000	0.669	(±0.008)	0.001	0.76	(±0.017)	0.001
10000	0.731	(±0.000)	0.001	0.73	(±0.030)	0.008

GPT-3 Classification Endpoint performance on data science question topic classification, additional examples generated using GPT-3 Davinci Completion. p-values test for significance from results with no additional examples.

What about the Completion Endpoint?

Completion: Selecting Optimal Examples

- **Genetic algorithm** selects best examples
- GPT-3 creates *new* candidates at each generation



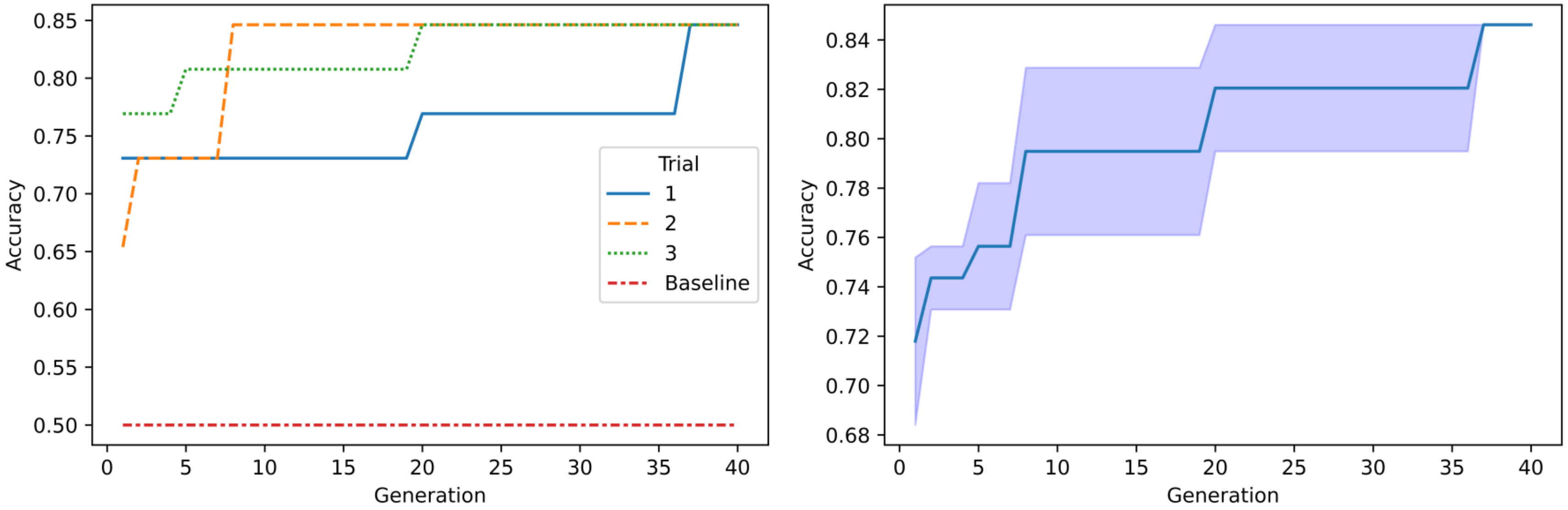
Classify the question as “data” or “other”

Which do you prefer, Python or R?
Label: *Data*

What is the meaning of life?
Label: *Other*

…

Validation Results – Completion Endpoint (Optimizing via Genetic Algorithm)



Completion Endpoint accuracy on Validation (n = 26) question sets over 40 generations using random subsets of 8 examples from Train (n = 26) set as initial population. All questions posed by Big Data Club Discord Server.

Test Results – Completion Endpoint (Optimizing via Genetic Algorithm)

Best Val. Accuracy	0.85 (+/- 0.00)
Test Accuracy	0.67 (+/- 0.02)

Drastic underperformance compared to validation set!

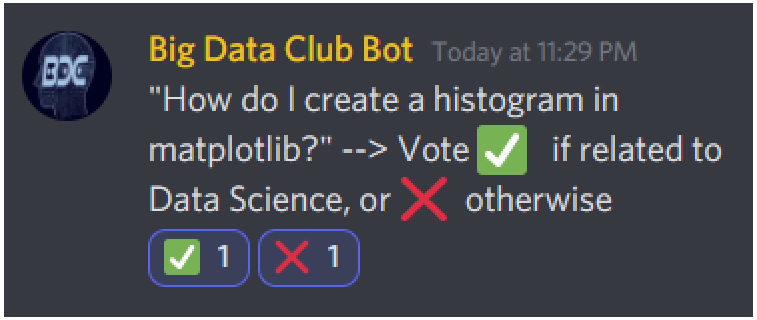
Discussion & Future Work

Subjective human language → imperfections

Common Mistakes:

- “Can someone help me understand how to SSH into the computing cluster on Friday?” [Data]
- “Here's a link on how to make custom Jupyter notebook themes. Big Data Club-themed notebooks, anyone?” [Data]
- “Are you coming to Big Data Club tomorrow?” [Other]

Want data quality control

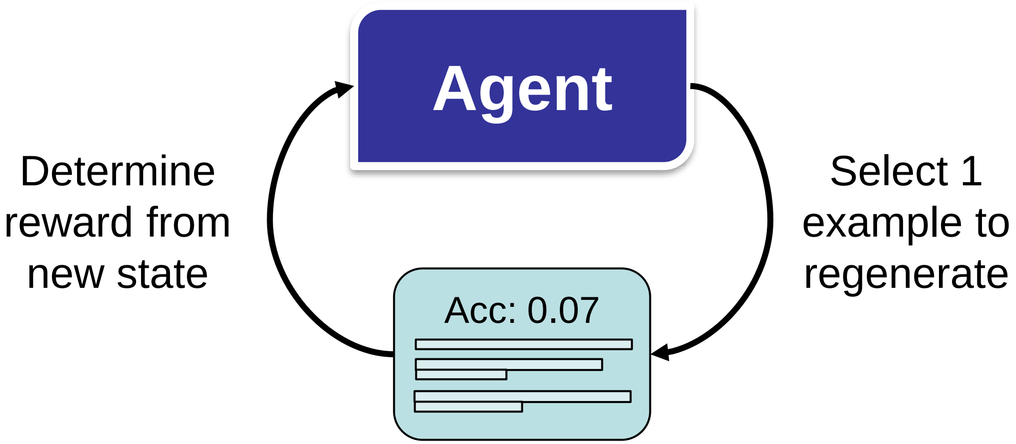


Big Data Club Bot Today at 11:29 PM

"How do I create a histogram in matplotlib?" --> Vote ✅ if related to Data Science, or ❌ otherwise

✅ 1 ❌ 1

Discussion & Future Work


Agent
Determine reward from new state
Select 1 example to regenerate
Acc: 0.07

Completion accuracy loss *validation* → *test* – why?

- Overfitting
- Does not “understand” prompt [2]
- GPT-3 weakest in classification [1, 3]

➢ *Potential Improvement:* **Reinforcement learning**

➢ Test other applications (sentiment, summarization, etc.)

Acknowledgements

Thank you to…

- The UMass Dartmouth **Program in Data Science**, for financial support.
- The UMass Dartmouth **Big Data Club** members for submitting questions used in this study

Thank you!

Questions?

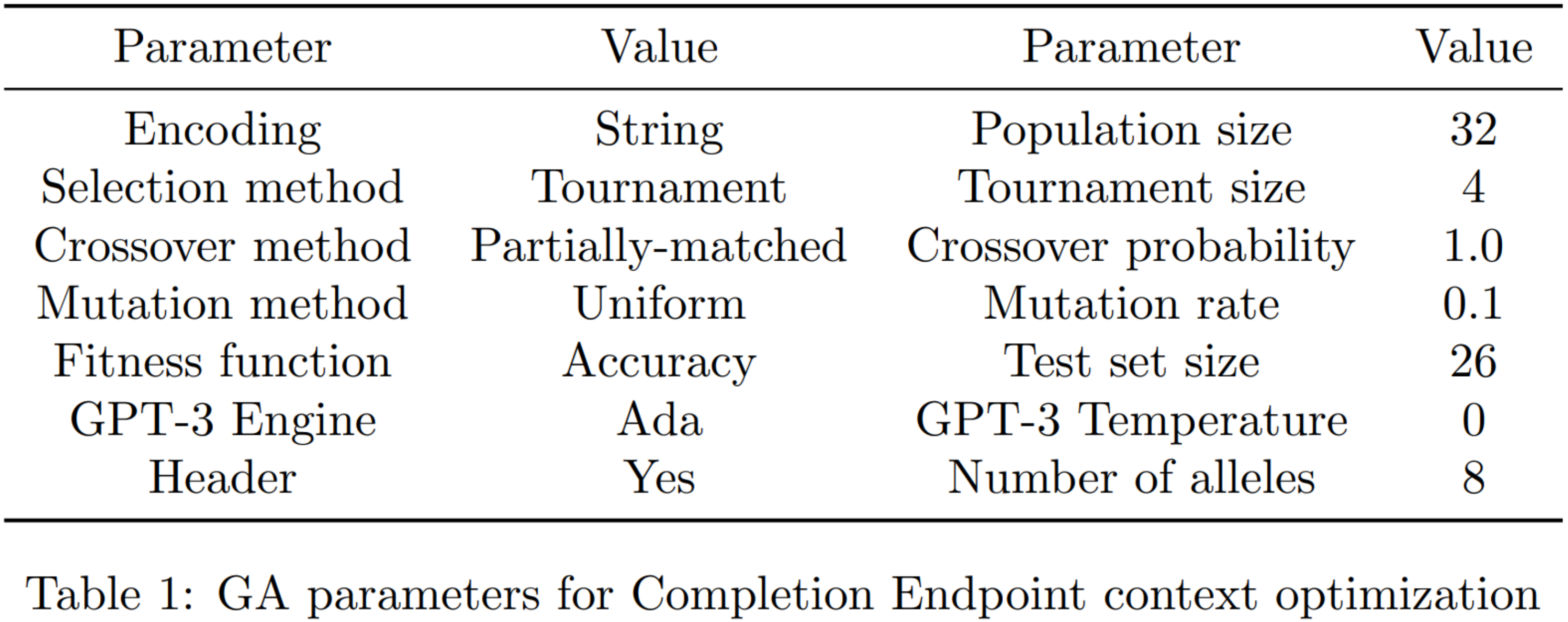
References

[1] T Brown et al. “Language Models are Few-Shot Learners,” 2020. Available: https://arxiv.org/pdf/2005.14165.pdf.

[2] A Webson and E Pavlick. “Do Prompt-Based Models Really Understand the Meaning of their Prompts?” 2021. Available: https://arxiv.org/pdf/2109.01247.pdf

[3] R Habib “How good is GPT-3 in practice?” 2021. Available: https://humanloop.com/blog/how-good-is-gpt-3-in-practice

Appendix: GA Parameters



Parameter	Value	Parameter	Value
Encoding	String	Population size	32
Selection method	Tournament	Tournament size	4
Crossover method	Partially-matched	Crossover probability	1.0
Mutation method	Uniform	Mutation rate	0.1
Fitness function	Accuracy	Test set size	26
GPT-3 Engine	Ada	GPT-3 Temperature	0
Header	Yes	Number of alleles	8

Table 1: GA parameters for Completion Endpoint context optimization