CS8803-MDS

Project Proposal

Lecture 10
09/25/22

Presentation guidelines

Total time: 5min per group

What to cover

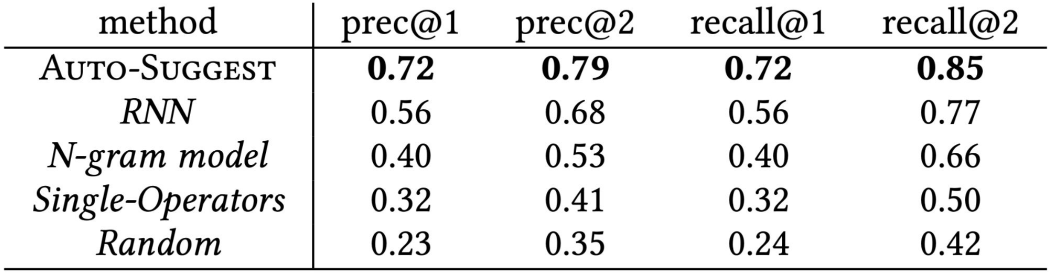
- A short introduction/problem slide
- Your bit-flip slide
- A solution/plan slide

We will have time for ~2min Q&A after each presentation.

Evaluating LLMs on Data Preparation Recommendation Tasks

Introduction

- Data preparation is time-consuming process
- Evaluate the usage of general-purpose LLMs against specialized models
- Auto-Suggest
  - Specialized machine learning model
  - "an end-to-end system that harvests public notebooks on GitHub to recommend data prep steps"
  - Next operator prediction



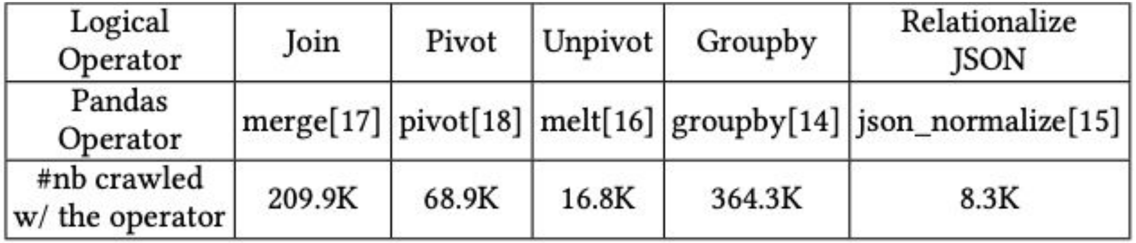
method	prec@1	prec@2	recall@1	recall@2
AUTO-SUGGEST	**0.72**	**0.79**	**0.72**	**0.85**
RNN	0.56	0.68	0.56	0.77
N-gram model	0.40	0.53	0.40	0.66
Single-Operators	0.32	0.41	0.32	0.50
Random	0.23	0.35	0.24	0.42

Bit-Flip

- Current ML models can predict data preparation queries
- Bit!
  - Accuracy of **Auto-Suggest**
- Flip!
  - Compare accuracy of Auto-Suggest to accuracies of …
    - **Chat GPT**
    - **GitHub CoPilot**

Solution and Evaluation

- Develop method for encoding Auto-Suggest dataset inputs as **text prompts**
- Use OpenAI API to test **LLM** accuracy on **single operator prediction** (JOIN, PIVOT, UNPIVOT, GROUPBY/AGGREGATION)
- Manually test **Github Copilot** generation output on a subset of dataset
- Evaluate based on popular **Information Retrieval** metrics
    - Precision@K, NDCG@K, Full Accuracy, etc.
- Compare metrics with results from **Auto-Suggest** paper



Logical Operator	Join	Pivot	Unpivot	Groupby	Relationalize JSON
Pandas Operator	merge[17]	pivot[18]	melt[16]	groupby[14]	json_normalize[15]
#nb crawled w/ the operator	209.9K	68.9K	16.8K	364.3K	8.3K

Project Group 2

Introduction: Automated Financial Reporting Data Extraction and Joining

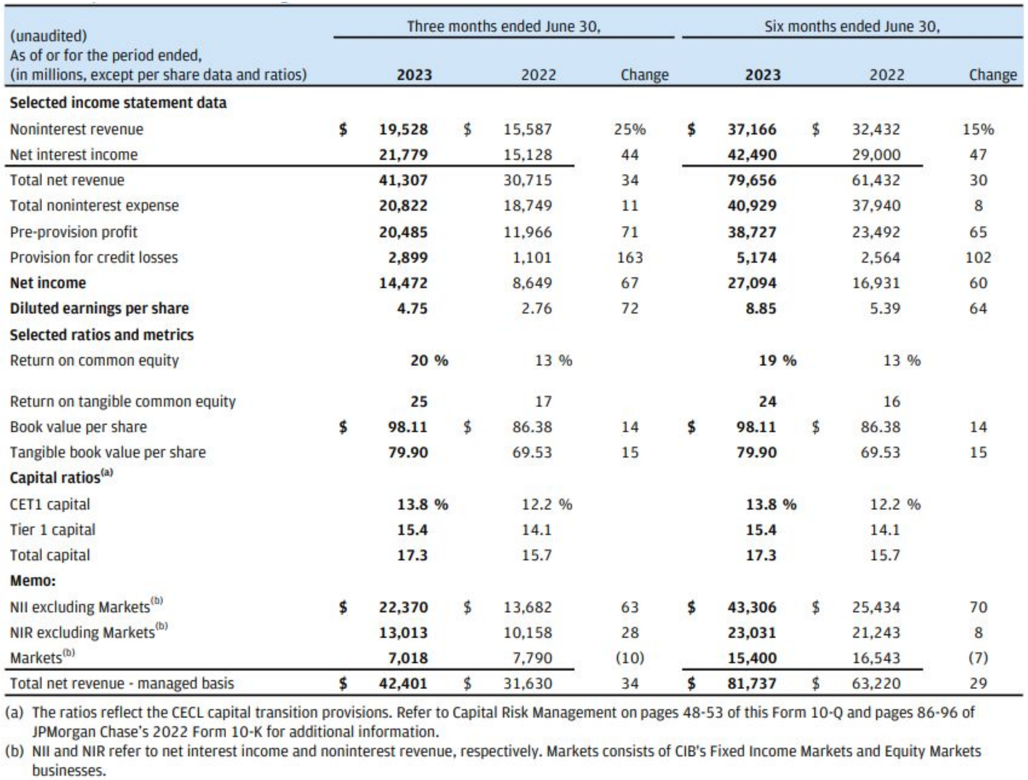
Objective: Create a tailored solution to the problem of joining disparate data from mandatory SEC filings from different companies.

- Although mandatory, financial reports have unique and inconsistent formats even from the same companies due to different scales, business functions, and regulatory requirements.
- Reduces the scale and granularity at which financial analysis can be done.
- Manual methods for joining require extensive domain knowledge.

Bit Flip

- Create a methodology for the standardization of data from an assortment of financial documents with differing content and formatting.
  - Leveraging fuzzy, semantic, and entity recognition to recommend intelligent joins.
  - Ensuring document specific content alignment based on assumed regulatory compliance.
- An adaptable system for future changes in report structures and/or regulatory requirements.
- This system alleviates the requirement of specialized domain knowledge and expert intuition for company specific reporting choices when aggregating data from different filings and companies.

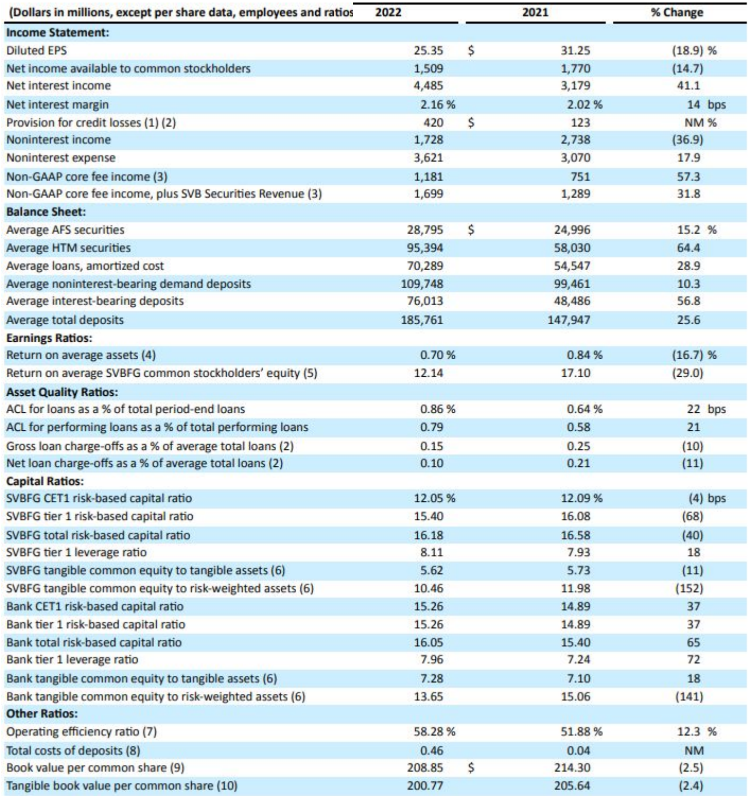
Examples



(unaudited) As of or for the period ended, (in millions, except per share data and ratios)	Three months ended June 30,			Six months ended June 30,		
	2023	2022	Change	**2023**	2022	Change
Selected income statement data						
Noninterest revenue	$ **19,528**	$ 15,587	25%	$ **37,166**	$ 32,432	15%
Net interest income	**21,779**	15,128	44	**42,490**	29,000	47
Total net revenue	**41,307**	30,715	34	**79,656**	61,432	30
Total noninterest expense	**20,822**	18,749	11	**40,929**	37,940	8
Pre-provision profit	**20,485**	11,966	71	**38,727**	23,492	65
Provision for credit losses	**2,899**	1,101	163	**5,174**	2,564	102
Net income	**14,472**	8,649	67	**27,094**	16,931	60
Diluted earnings per share	**4.75**	2.76	72	**8.85**	5.39	64
Selected ratios and metrics						
Return on common equity	**20 %**	13 %		**19 %**	13 %	
Return on tangible common equity	**25**	17		**24**	16	
Book value per share	$ **98.11**	$ 86.38	14	$ **98.11**	$ 86.38	14
Tangible book value per share	**79.90**	69.53	15	**79.90**	69.53	15
Capital ratios[(a)]						
CET1 capital	**13.8 %**	12.2 %		**13.8 %**	12.2 %	
Tier 1 capital	**15.4**	14.1		**15.4**	14.1	
Total capital	**17.3**	15.7		**17.3**	15.7	
Memo:						
NII excluding Markets[(b)]	$ **22,370**	$ 13,682	63	$ **43,306**	$ 25,434	70
NIR excluding Markets[(b)]	**13,013**	10,158	28	**23,031**	21,243	8
Markets[(b)]	**7,018**	7,790	(10)	**15,400**	16,543	(7)
Total net revenue - managed basis	$ **42,401**	$ 31,630	34	$ **81,737**	$ 63,220	29

(a) The ratios reflect the CECL capital transition provisions. Refer to Capital Risk Management on pages 48-53 of this Form 10-Q and pages 86-96 of JPMorgan Chase's 2022 Form 10-K for additional information.

(b) NII and NIR refer to net interest income and noninterest revenue, respectively. Markets consists of CIB's Fixed Income Markets and Equity Markets businesses.



(Dollars in millions, except per share data, employees and ratios	2022		2021	% Change
Income Statement:				
Diluted EPS	25.35	$	31.25	(18.9) %
Net income available to common stockholders	1,509		1,770	(14.7)
Net interest income	4,485		3,179	41.1
Net interest margin	2.16 %		2.02 %	14 bps
Provision for credit losses (1) (2)	420	$	123	NM %
Noninterest income	1,728		2,738	(36.9)
Noninterest expense	3,621		3,070	17.9
Non-GAAP core fee income (3)	1,181		751	57.3
Non-GAAP core fee income, plus SVB Securities Revenue (3)	1,699		1,289	31.8
Balance Sheet:				
Average AFS securities	28,795	$	24,996	15.2 %
Average HTM securities	95,394		58,030	64.4
Average loans, amortized cost	70,289		54,547	28.9
Average noninterest-bearing demand deposits	109,748		99,461	10.3
Average interest-bearing deposits	76,013		48,486	56.8
Average total deposits	185,761		147,947	25.6
Earnings Ratios:				
Return on average assets (4)	0.70 %		0.84 %	(16.7) %
Return on average SVBFG common stockholders' equity (5)	12.14		17.10	(29.0)
Asset Quality Ratios:				
ACL for loans as a % of total period-end loans	0.86 %		0.64 %	22 bps
ACL for performing loans as a % of total performing loans	0.79		0.58	21
Gross loan charge-offs as a % of average total loans (2)	0.15		0.25	(10)
Net loan charge-offs as a % of average total loans (2)	0.10		0.21	(11)
Capital Ratios:				
SVBFG CET1 risk-based capital ratio	12.05 %		12.09 %	(4) bps
SVBFG tier 1 risk-based capital ratio	15.40		16.08	(68)
SVBFG total risk-based capital ratio	16.18		16.58	(40)
SVBFG tier 1 leverage ratio	8.11		7.93	18
SVBFG tangible common equity to tangible assets (6)	5.62		5.73	(11)
SVBFG tangible common equity to risk-weighted assets (6)	10.46		11.98	(152)
Bank CET1 risk-based capital ratio	15.26		14.89	37
Bank tier 1 risk-based capital ratio	15.26		14.89	37
Bank total risk-based capital ratio	16.05		15.40	65
Bank tier 1 leverage ratio	7.96		7.24	72
Bank tangible common equity to tangible assets (6)	7.28		7.10	18
Bank tangible common equity to risk-weighted assets (6)	13.65		15.06	(141)
Other Ratios:				
Operating efficiency ratio (7)	58.28 %		51.88 %	12.3 %
Total costs of deposits (8)	0.46		0.04	NM
Book value per common share (9)	208.85	$	214.30	(2.5)
Tangible book value per common share (10)	200.77		205.64	(2.4)

Solution and Evaluation

- Test the system's efficacy on various mandatory regulatory filings such as 8-K, S-1, 13-D & G, annual reports, call reports, and more
- Retrieved using python-edgar (index of SEC filings since 1993)
- Evaluate with benchmarks of simpler PDF data extraction tools and ground truths from documents
    - Compares accuracy
    - Integration efficacy
    - Semantic alignment
    - Processing speed

Project Group 3

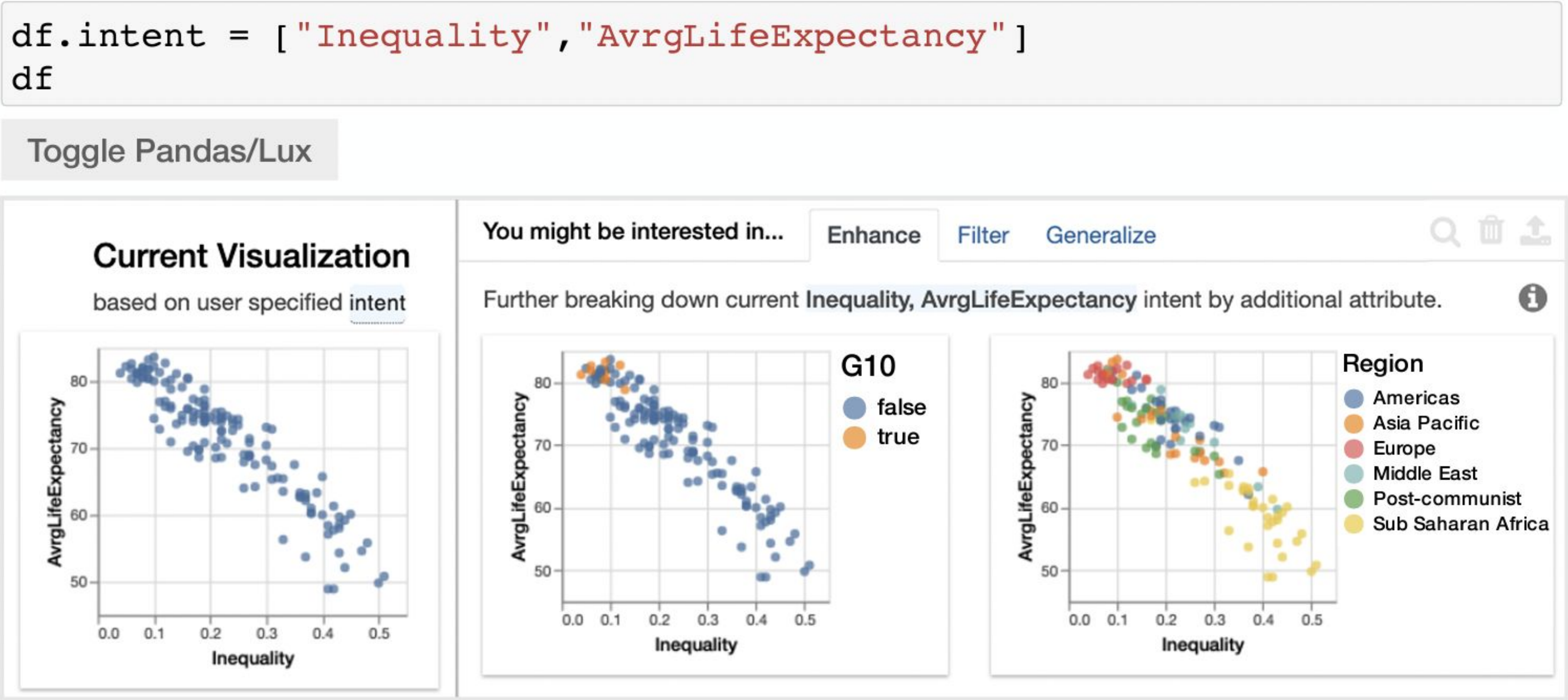
Impact of data cleaning on visualization recommendations

Introduction

→ Lux offers data analysts a unique capability to automate visualizations.

→ Here, We assess whether various data cleaning methods influence the quality and suggestions of visualizations by Lux.

Lux Example



```
df.intent = ["Inequality","AvrgLifeExpectancy"]
df
```

Toggle Pandas/Lux

Current Visualization

based on user specified intent

You might be interested in...

Further breaking down current **Inequality, AvrgLifeExpectancy** intent by additional attribute.

Bit-Flip

⇢ Most visualization tools, including Lux, largely revolve around visualization recommendations with an implicit assumption that data is **pre-processed** and free of discrepancies.

↩ There is inadequate research on how different data cleaning methods may influence the outcome of visual recommendations in systems like Lux.

Solution and Evaluation

1. Utilize common data metrics to determine if data cleaning has an effect on Lux and its ranking preferences
    a. Regression Models: Root mean square error, R-squared

2. Utilize state of the art automatic data cleaning methods, such as HoloClean and AlphaClean, in tandem with Lux

3. If none of the state of the art work quickly and well, create own data cleaning method to which the analysts can use easily and iteratively
    a. For example: a data cleaning method might be preferred over another one and if the analysts selects one, keep using that one or adapt to it

Project Group 4

Effective Pipeline Transfer in Automated Data Cleaning

Introduction

- Automated data cleaning methods like BoostClean [1], ActiveClean [2] save significant amount of time from manual cleaning approaches.

- They also enable cleaning to maximize downstream model performance (“cleaning for ML”).

- On analyzing them, we found that:
    - These methods do not deeply analyze how the cleaning performance is affected by different end models trained.
    - Many of them retrain the end model multiple times, making them computationally expensive for large models.

[1] Sanjay Krishnan, E. (2017). BoostClean: Automated Error Detection and Repair for Machine Learning. ACM.
[2] Sanjay Krishnan, K. (2016). ActiveClean: Interactive Data Cleaning For Statistical Modeling. VLDB.

Bit-Flip

1. For a given model, are there some automated cleaning methods which are more advantageous to use than others?

   i.e. do some methods preferentially improve performance more for some models than others?

2. To avoid iterative retraining of large models, can we transfer datasets cleaned with simpler models (that are faster) to complex end models?

3. Can we transfer data cleaning pipelines trained for one ML task to effectively train a model for another one?

Solution

To achieve this

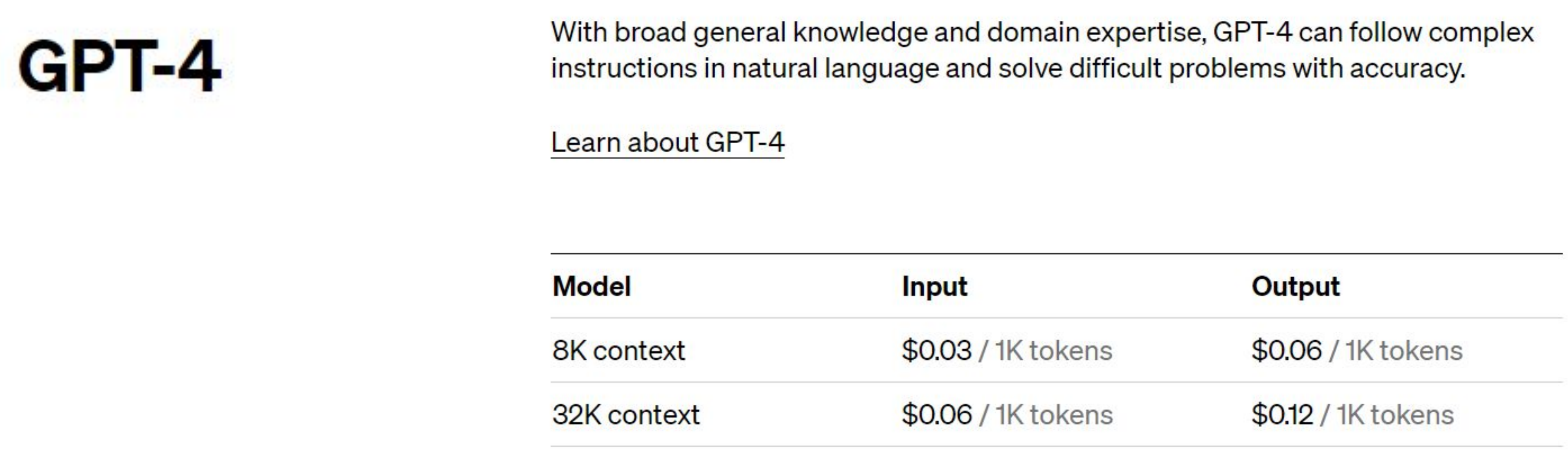
- We train smaller models using these cleaning methods, and then transfer the cleaned dataset to train a larger end model once, saving time by avoiding iteratively retraining complex end models.
- We transfer a dataset cleaned for one task to train another model against a new target variable to avoid recleaning.
- We define transferability based on the runtime and performance difference of these models when cleaned with and without transfer.
- We evaluate transferability using benchmark datasets from CleanML [3] and understand interplay between model and cleaning method selection.
- Through our results, we aim to guide data scientists to save time and choose effective automated cleaning methods for their models.

[3] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, & Ce Zhang. (2021). CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks.

Project Group 5

Panini: Context-aware Waste-minimizing Speculative Decoding for LLMs

26th September, 2023



GPT-4

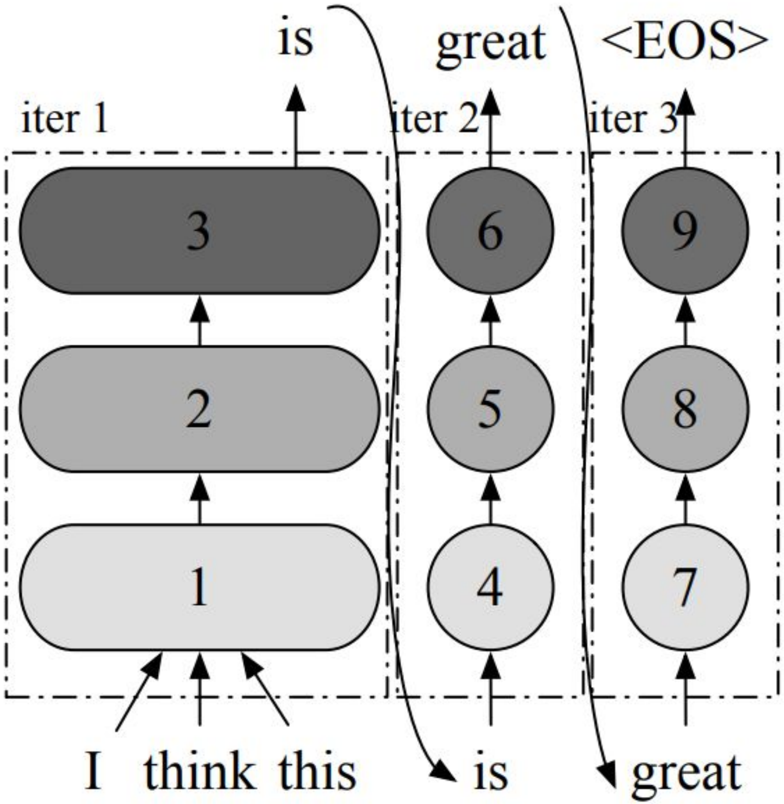
With broad general knowledge and domain expertise, GPT-4 can follow complex instructions in natural language and solve difficult problems with accuracy.

Learn about GPT-4

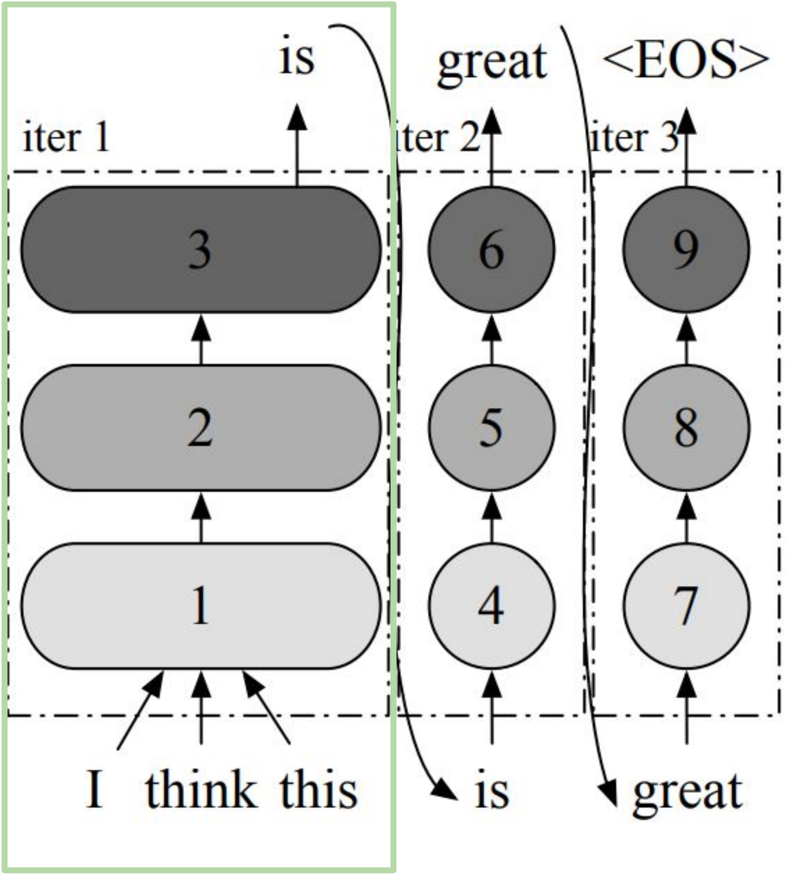
Model	Input	Output
8K context	$0.03 / 1K tokens	$0.06 / 1K tokens
32K context	$0.06 / 1K tokens	$0.12 / 1K tokens

Why OpenAI charges 2x price for output tokens? 🤔

Autoregressive Generation🔁

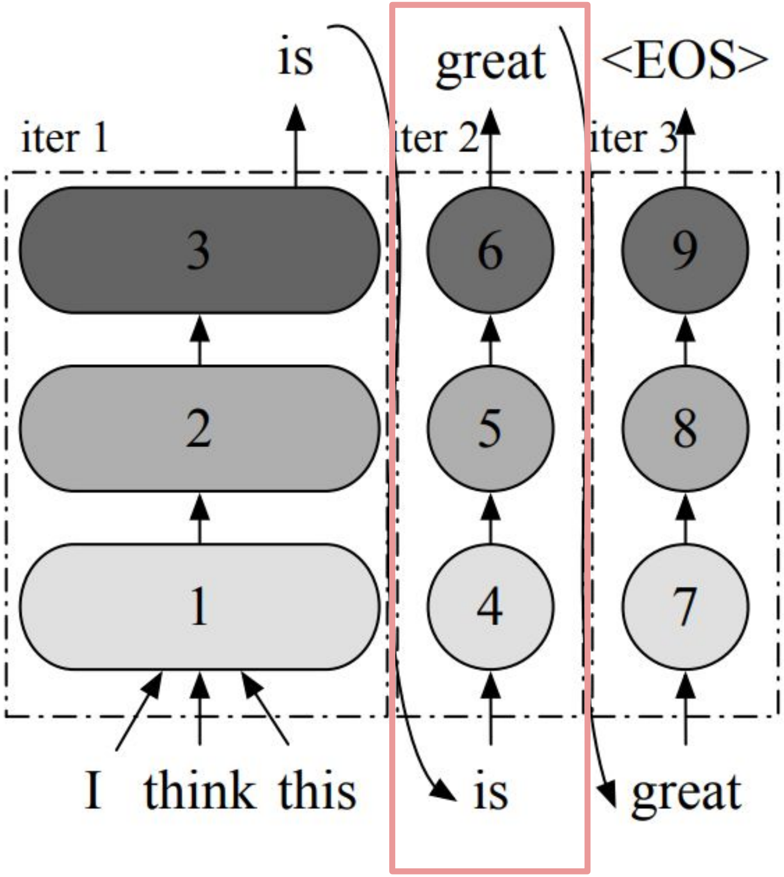


Yu et al. OSDI’22



Prefill

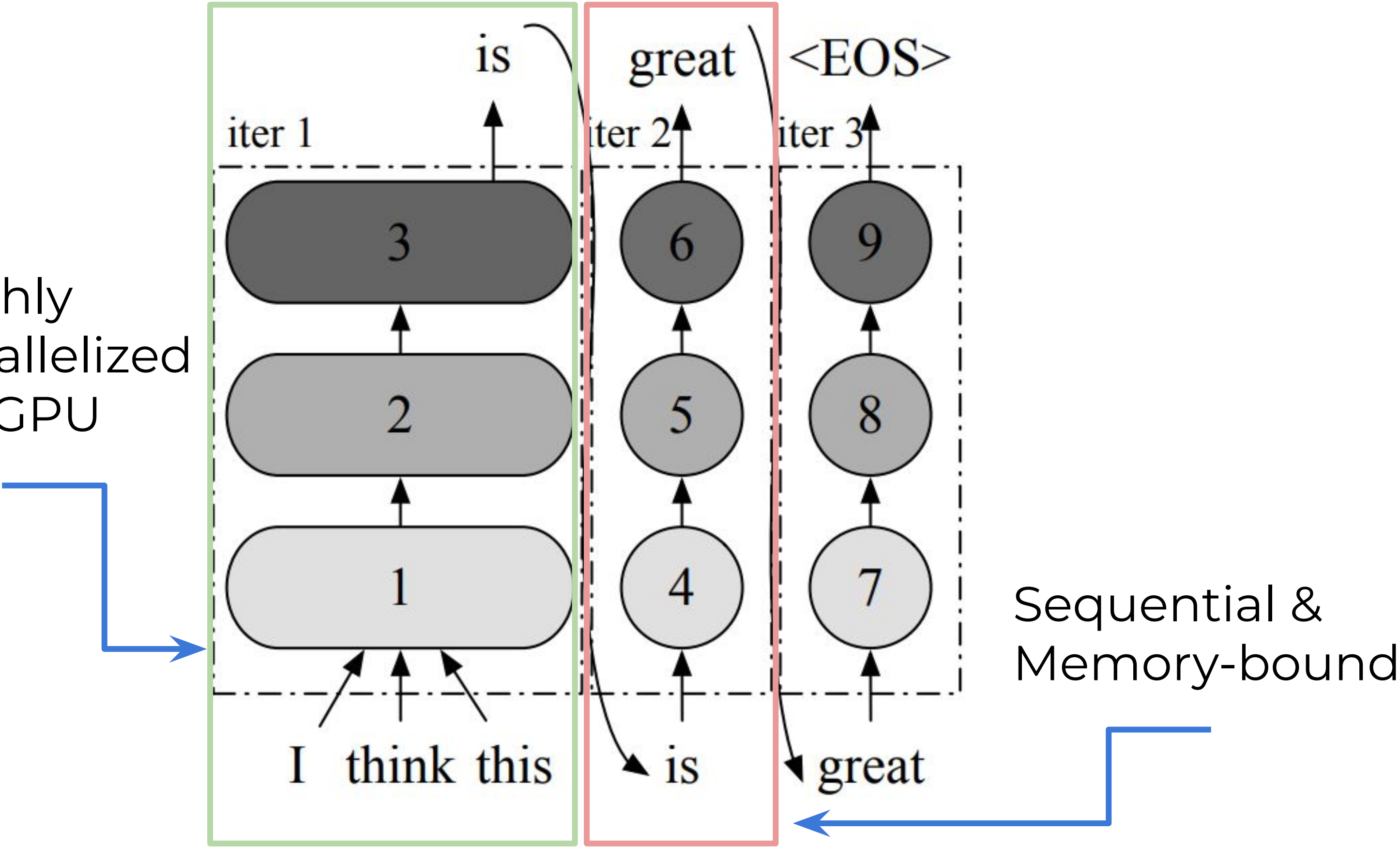
Yu et al. OSDI'22



Decode

Yu et al. OSDI’22

Highly Parallelized on GPU



Sequential & Memory-bound

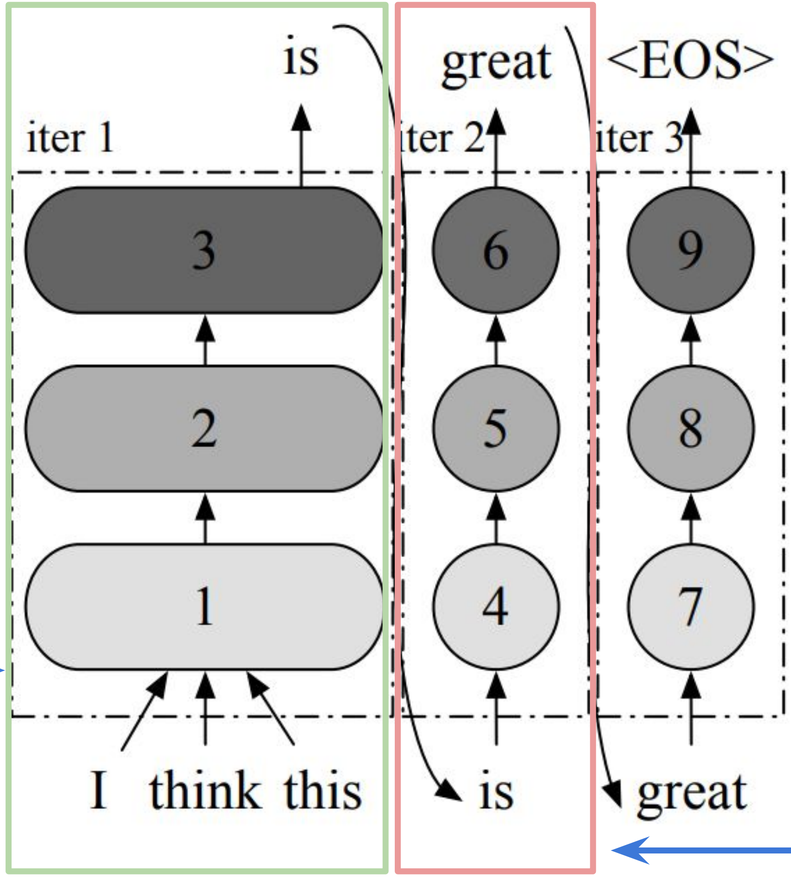
Prefill

Decode

Yu et al. OSDI'22

100x Higher Decode Cost 🥵

Highly Parallelized on GPU



Sequential & Memory-bound

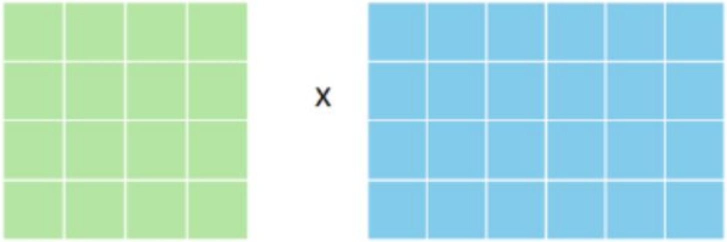
Prefill

Decode

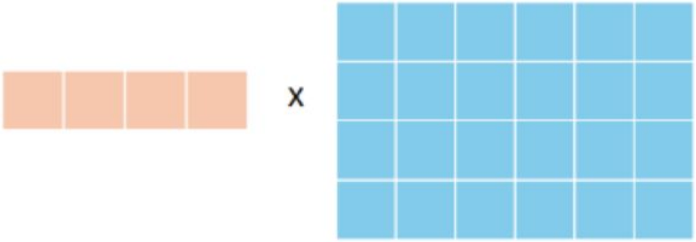
Yu et al. OSDI’22

Understanding Matmul Performance🕒

Prefill



Decode



Arithmetic Intensity 🧮

1️⃣Data movement (HBM -> GPU cores) is much more costly than actual computation.

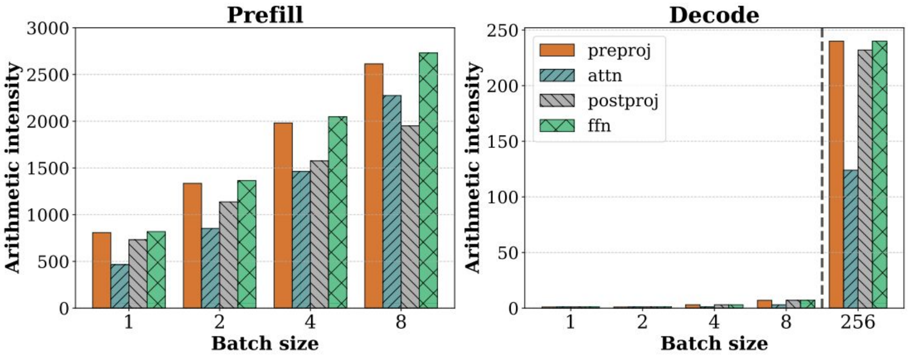
2️⃣Arithmetic intensity measures the amount of compute performed for each byte of data moved.

3️⃣A GPUs compute saturation point can be computed,

I = FP16 FLOPS/Memory Band

4️⃣For matmuls, we can compute intensity as,

$$\frac{M \cdot N \cdot K}{M \cdot K + N \cdot K + M \cdot N}$$



Arithmetic intensity for GPT-like models ~ Number of tokens in batch

Agrawal et al. ‘23

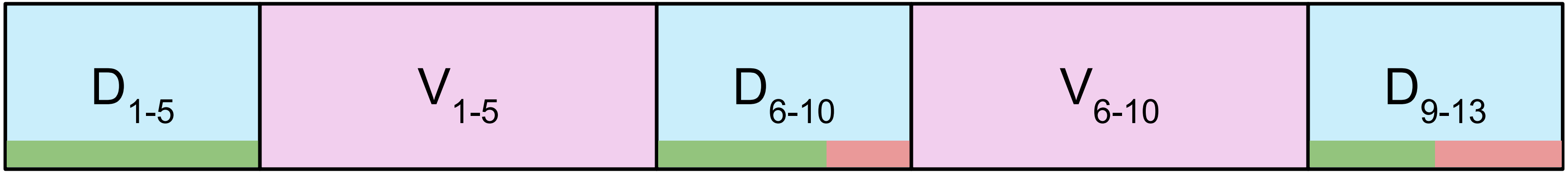
Speculative Decoding 🔮

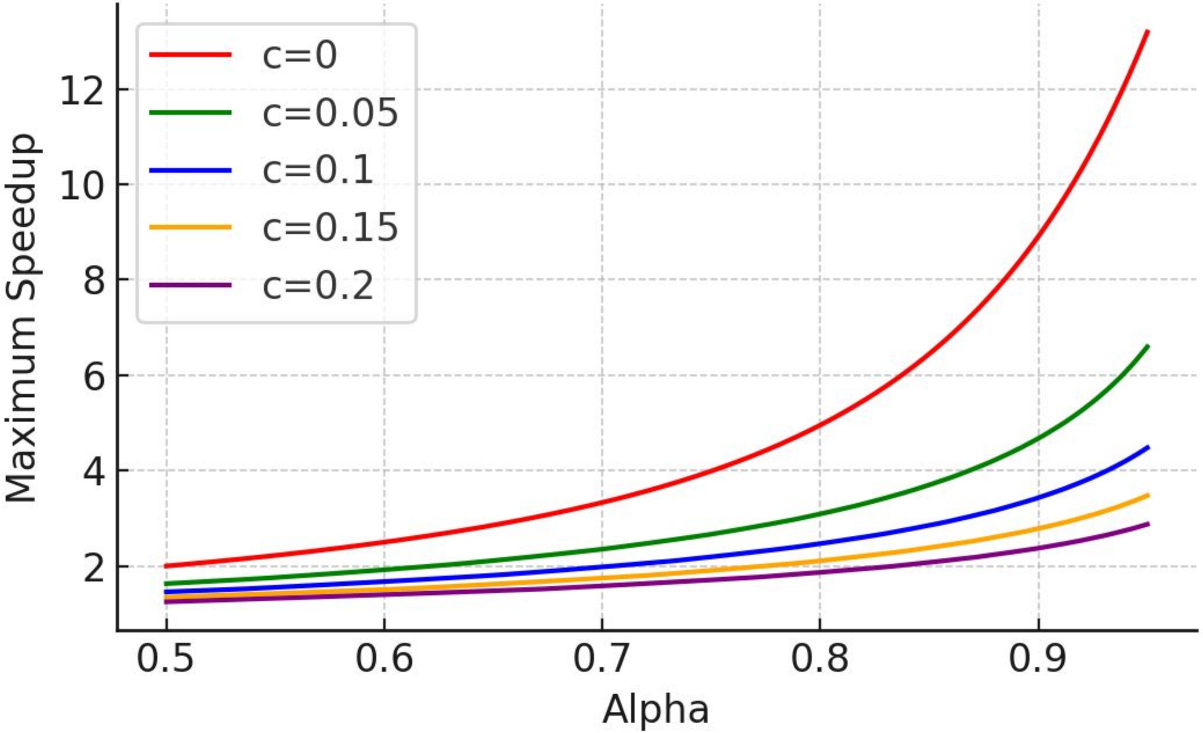
Speculative Decode 🔮

1️⃣ Pick a small “drafter” model (~10x smaller), use that to generate a draft of 5-10 tokens in a auto-regressive manner, and pass that draft to the original model for verification [Leviathan et al. 22’, Chen et al. 23’].

2️⃣ All the tokens can be verified in parallel without any additional cost (remember vector - matrix multiplication).

3️⃣ Rejection sampling assures that we get provably correct output.

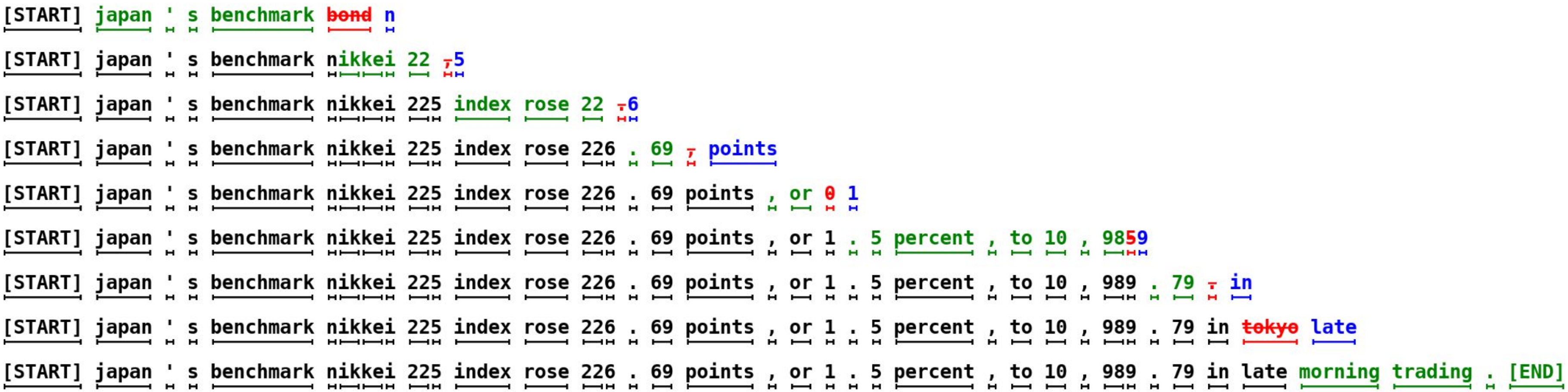




C: Decode runtime ratio between drafter and verification model

Alpha: Acceptance ratio

Maximum decoding speedup with speculative execution at different acceptance rates.



[START] japan ' s benchmark ~~bond~~ n

[START] japan ' s benchmark nikkei 22 ~~,~~5

[START] japan ' s benchmark nikkei 225 index rose 22 ~~.~~6

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 ~~,~~ points

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or ~~0~~ 1

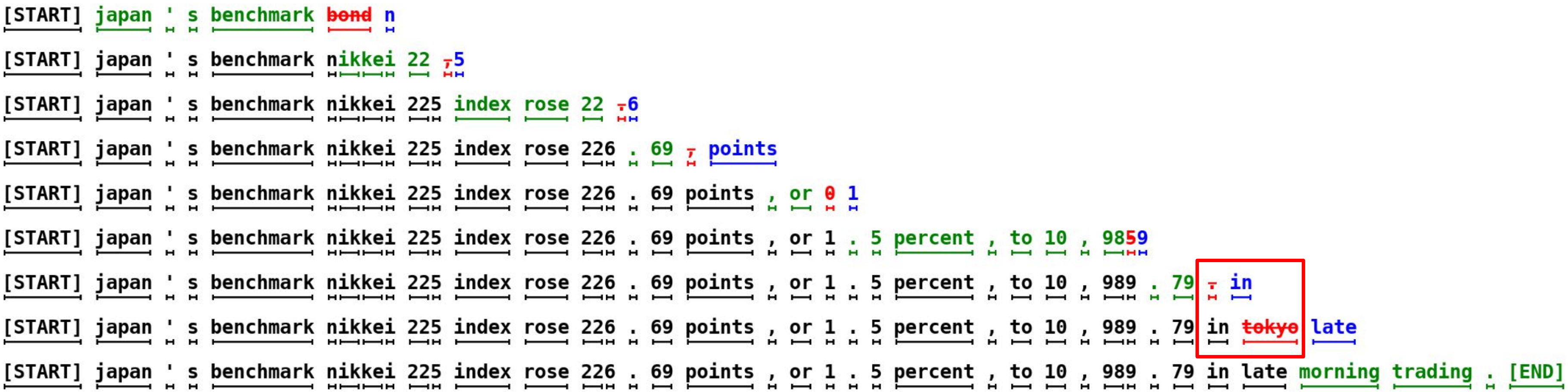
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 98~~5~~9

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 ~~.~~ in

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in ~~tokyo~~ late

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in late morning trading . [END]

Leviathan et al. ‘22



[START] japan ' s benchmark ~~bond~~ n

[START] japan ' s benchmark nikkei 22 ~~,~~5

[START] japan ' s benchmark nikkei 225 index rose 22 ~~.~~6

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 ~~,~~ points

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or ~~0~~ 1

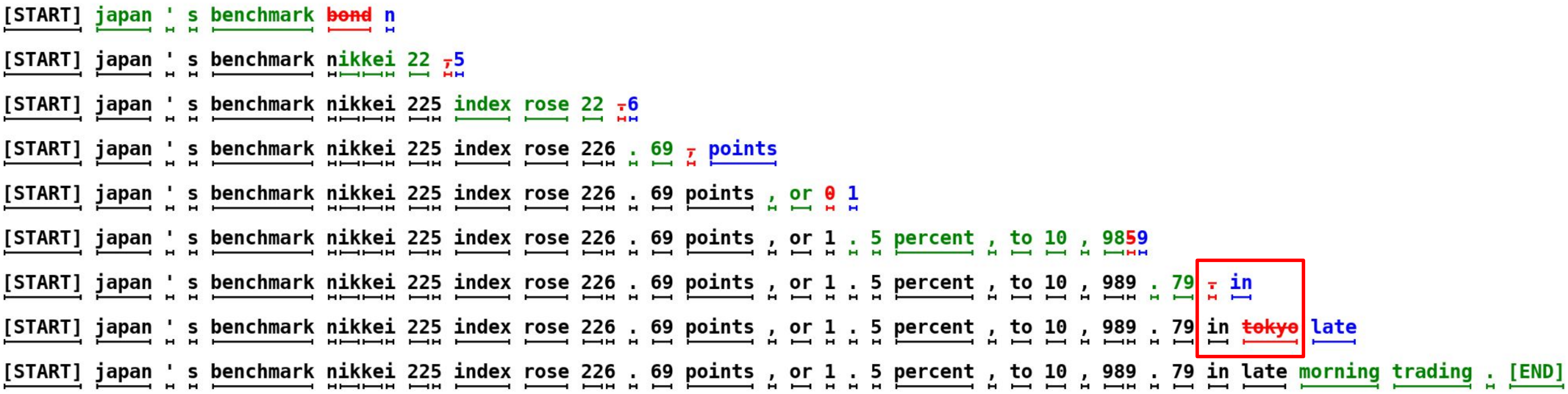
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 98~~5~~9

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 ~~.~~ in

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in ~~tokyo~~ late

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in late morning trading . [END]

Leviathan et al. ‘22



[START] japan ' s benchmark ~~bond~~ n

[START] japan ' s benchmark nikkei 22 ~~.~~5

[START] japan ' s benchmark nikkei 225 index rose 22 ~~.~~6

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 ~~,~~ points

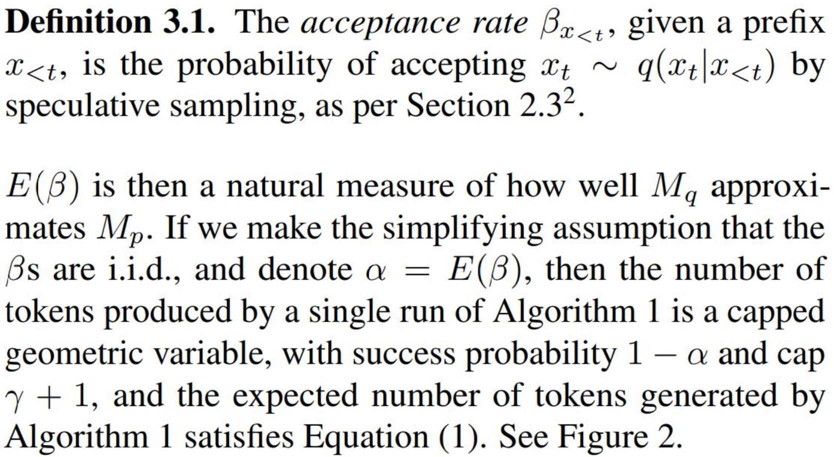
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or ~~0~~ 1

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 98~~5~~9

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 ~~.~~ in

[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in ~~tokyo~~ late

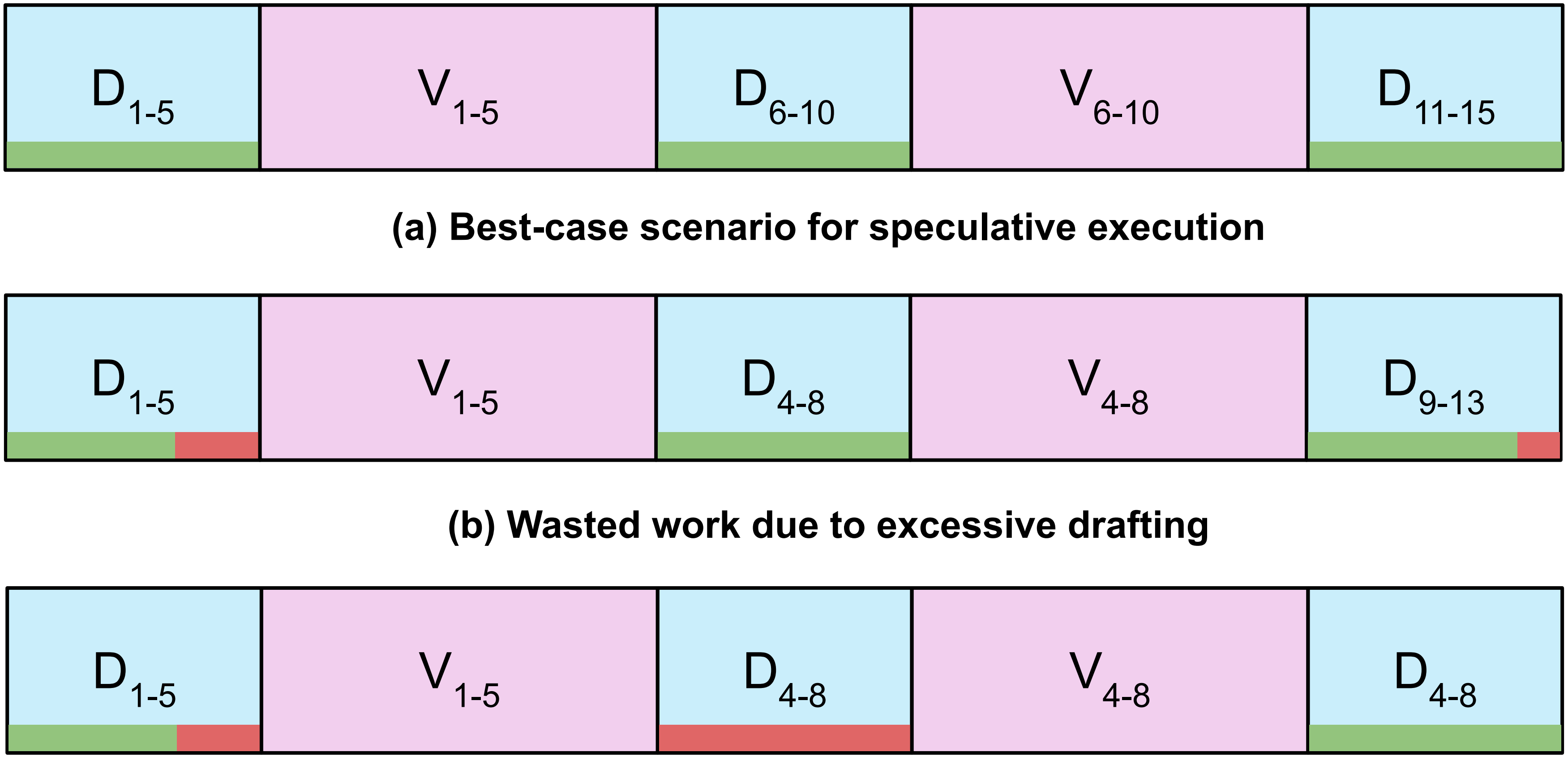
[START] japan ' s benchmark nikkei 225 index rose 226 . 69 points , or 1 . 5 percent , to 10 , 989 . 79 in late morning trading . [END]



Definition 3.1. The *acceptance rate* $\beta_{x_{<t}}$, given a prefix $x_{<t}$, is the probability of accepting $x_t \sim q(x_t|x_{<t})$ by speculative sampling, as per Section 2.3[2].

$E(\beta)$ is then a natural measure of how well M_q approximates M_p. If we make the simplifying assumption that the βs are i.i.d., and denote $\alpha = E(\beta)$, then the number of tokens produced by a single run of Algorithm 1 is a capped geometric variable, with success probability $1 - \alpha$ and cap $\gamma + 1$, and the expected number of tokens generated by Algorithm 1 satisfies Equation (1). See Figure 2.

Leviathan et al. ‘22



(a) Best-case scenario for speculative execution

(b) Wasted work due to excessive drafting

(c) Wasted work due to insufficient context provided from large model

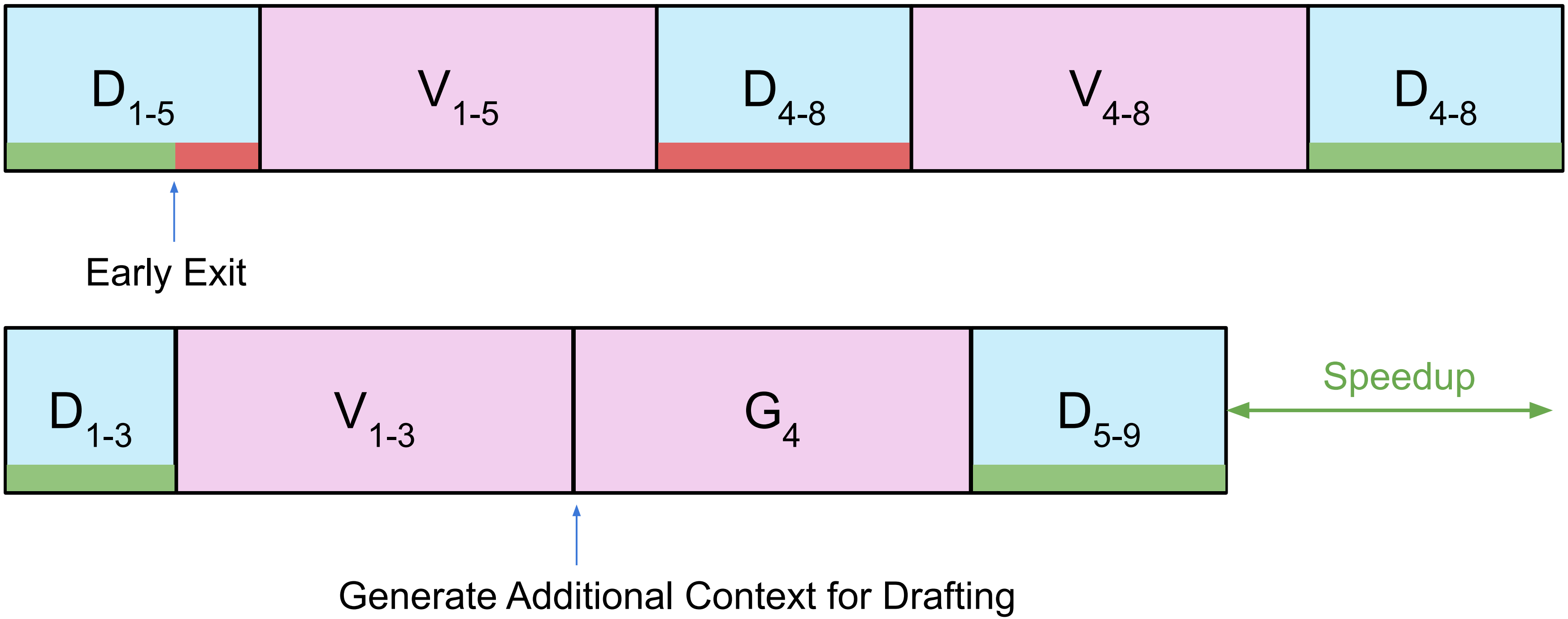
Context-aware Speculative Decoding 🌠

1️⃣The acceptance rate depends on context - intuitively, rejects are more likely to occur at phrase boundaries, numbers, etc.

2️⃣If we can more estimate the variation in acceptance rate based on context, we can avoid wasted work in speculative decoding.

3️⃣Rather than using a fixed draft length, we can more generally define this as an optimization problem where the control hand-offs between the drafter and verifier models must be optimized to minimize wasted work.

Speculative Execution with context-aware oracle



D_{1-5} V_{1-5} D_{4-8} V_{4-8} D_{4-8}

Early Exit

D_{1-3} V_{1-3} G_4 D_{5-9}

Speedup

Generate Additional Context for Drafting

Execution Plan 📅

1️⃣**October 10th:** Collect token acceptance data across different datasets, and perform analysis to identify common motifies.

2️⃣**October 25th:** Use acceptance traces to simulate the oracle system and measure maximum expected improvements.

3️⃣**November 10th:** Design and evaluate strategies to predict acceptance rate based on context.

4️⃣**November 25th:** Implement the context-aware policy using vLLM and evaluate the end-to-end performance gains.

4️⃣**December 5th:** Compile results and prepare the final report.

Summary 📝

1️⃣**Speculative execution** is used to speed up decoding phase in LLM inference.

2️⃣Existing techniques **assume constant acceptance rate** and use rigid heuristics.

3️⃣**Context-aware** speculative decoding can minimize the wasted work and reduce the end-to-end decoding latency.

Project Group 6

Introduction: Impact of Data Cleaning on Visualization Recommendations

- Visual recommendation (VisRec) systems have been proposed to lower the barrier of data visualization
- Lux introduces a framework that automatically suggests relevant data visualizations based on the characteristics of the dataset [1]
  - Integrates with pandas framework
  - Accelerates exploration and discovery
- Clean data cannot be taken for granted when dealing with real-world datasets
- Data visualization recommendations work on dirty data

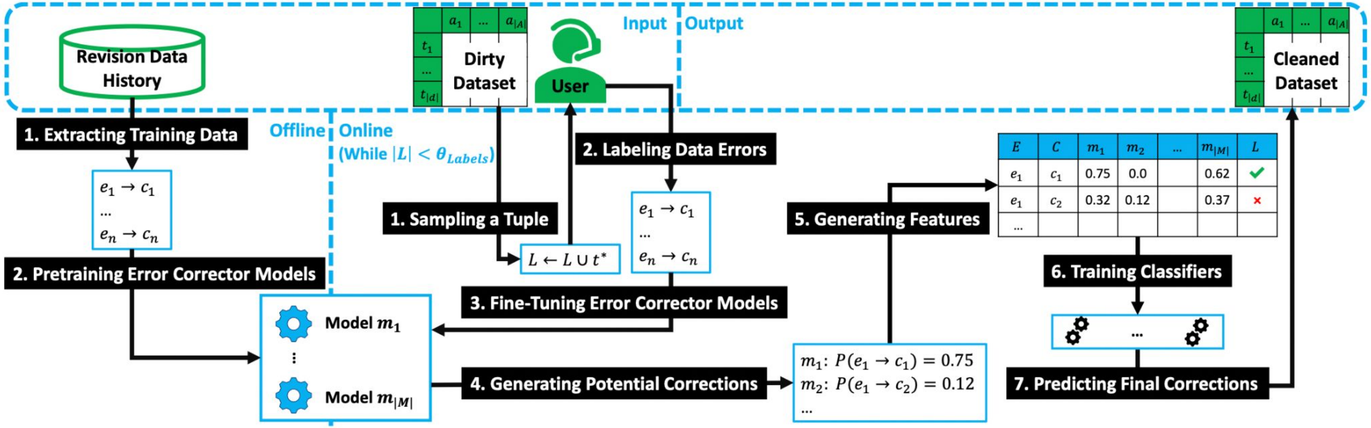
[1] Lee, Doris Jung-Lin, et al. "Lux: always-on visualization recommendations for exploratory dataframe workflows." *arXiv preprint arXiv:2105.00121* (2021).

Bit-Flip

- Bit
    - Existing VisRec systems assume that input datasets are **coherent**
- Flip
    - Examine how **different data cleaning approaches** and dirty data mode influence the results of visualization recommendations tools like Lux
        - Dirty data patterns such as missing values, duplicate data, formatting issues, outliers, and inconsistent data can lead to inaccurate or skewed recommendations
        - Different data cleaning methods can result in different datasets after cleaning

Solution

- Clean the data using different data cleaning scripts before visualizations
    - Conditional cleaning scripts: Custom detectors and repair functions
    - Automatic cleaning scripts: BoostClean [1], Baran [2]



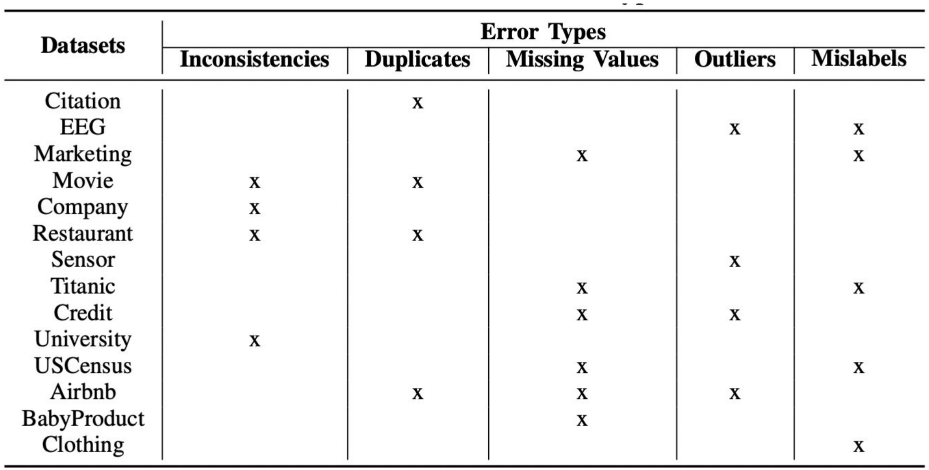
E	C	m_1	m_2	…	$m_{\vert M\vert}$	L
e_1	c_1	0.75	0.0		0.62	✓
e_1	c_2	0.32	0.12		0.37	×
…						

Figure 1: The workflow of Baran.

[1] Krishnan, Sanjay, et al. "Boostclean: Automated error detection and repair for machine learning." *arXiv preprint arXiv:1711.01299* (2017).
[2] Mahdavi, Mohammad, and Ziawasch Abedjan. "Baran: Effective error correction via a unified context representation and transfer learning." *Proceedings of the VLDB Endowment*13.12 (2020): 1948-1961.

Evaluation

- Compare the recommendation results training by clean data and new recommendations on dirty data with various data cleaning methods
    - Obtain 10-15 real-world datasets with vary error types and error percentages
- Evaluation metrics
    - Diversity: Number of distinct visualization types recommended
    - Consistency: Jaccard similarity or cosine similarity



Datasets	Error Types				
	Inconsistencies	Duplicates	Missing Values	Outliers	Mislabels
Citation		x			
EEG				x	x
Marketing			x		x
Movie	x	x			
Company	x				
Restaurant	x	x			
Sensor				x	
Titanic			x		x
Credit			x	x	
University	x				
USCensus			x		x
Airbnb		x	x	x	
BabyProduct			x		
Clothing					x

Table 1: CleanML summarizes 14 real-world datasets with varying error types and error rates [1]

[1] Li, Peng, et al. "CleanML: A study for evaluating the impact of data cleaning on ml classification tasks." *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021.

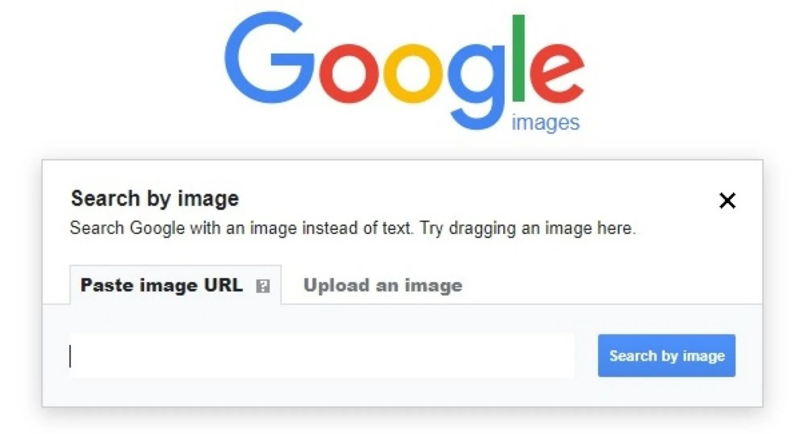
Project Group 7

Interactive Content-Based Image Retrieval

Introduction

- ❏ In today's digital age, there's a need for user-friendly image retrieval systems because of the large amount of image data available.
- ❏ Reverse image technology allows users to search images similar to a given query image

- ❏ Most reverse image searching methods have tried to perform similarity search on whole images or identify individual objects and lack an interactive interface.


Search by image

Search Google with an image instead of text. Try dragging an image here.

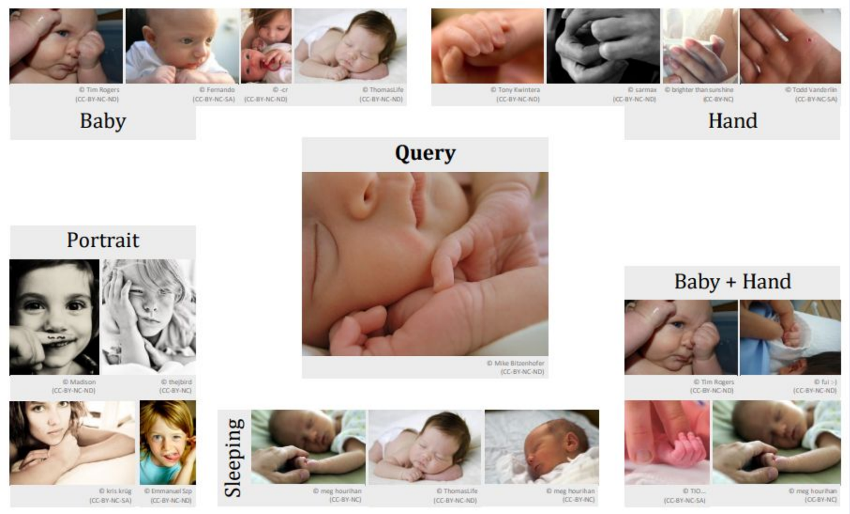
Bit Flip

Bit

- ❑ The entire process of reverse image searching solely depends on the image provided by the user and is thus non-interactive. It doesn’t take into account the intent behind the user's query, giving rise to inaccurate results.

Flip

- ❑ We propose the ability for the user to interactively query objects and the relations between them from a query image, on an image dataset.

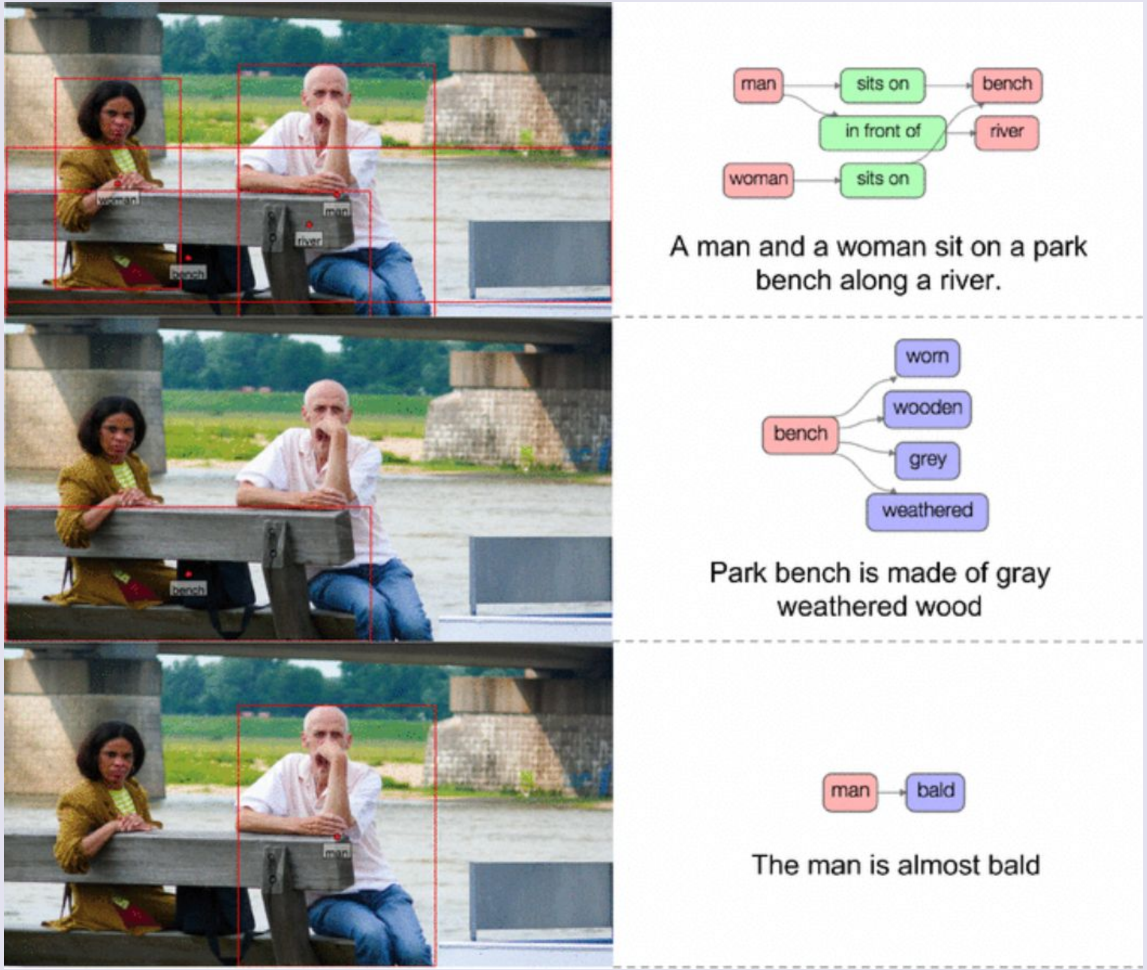

© Tim Rogers (CC-BY-NC-ND)
© Fernando (CC-BY-NC-SA)
© -cr (CC-BY-NC-ND)
© ThomasLife (CC-BY-NC-ND)
Baby
© Tony Kwintera (CC-BY-NC-ND)
© sarmax (CC-BY-NC-ND)
© brighter than sunshine (CC-BY-NC)
© Todd Vanderlin (CC-BY-NC-SA)
Hand
Query
© Mike Bitzenhofer (CC-BY-NC-ND)
Portrait
© Madison (CC-BY-NC-ND)
© thejbird (CC-BY-NC)
© kris krüg (CC-BY-NC-SA)
© Emmanuel Szp (CC-BY-NC-ND)
Sleeping
© meg hourihan (CC-BY-NC)
© ThomasLife (CC-BY-NC-ND)
© meg hourihan (CC-BY-NC)
Baby + Hand
© Tim Rogers (CC-BY-NC-ND)
© fui :-) (CC-BY-NC-ND)
© TIO... (CC-BY-NC-SA)
© meg hourihan (CC-BY-NC)

Solution

- ❏ We propose an interactive content-based image retrieval approach based on scene graph indexing wherein a user can choose the objects and relationships of his interest within the image.
- ❏ By harnessing the power of scene graphs, our system can pinpoint specific objects and their relationships that hold relevance to the user's query, thus allowing better interpretation of the user’s search intent.

Evaluation

- ❏ We propose a survey method for evaluating the system’s performance, a popular evaluation method for similar object detection and computer vision research.
- ❏ In these surveys, users assigned relevance scores to the query results based on their specific queries



man → sits on → bench

man → in front of → river

woman → sits on

A man and a woman sit on a park bench along a river.

bench → worn

bench → wooden

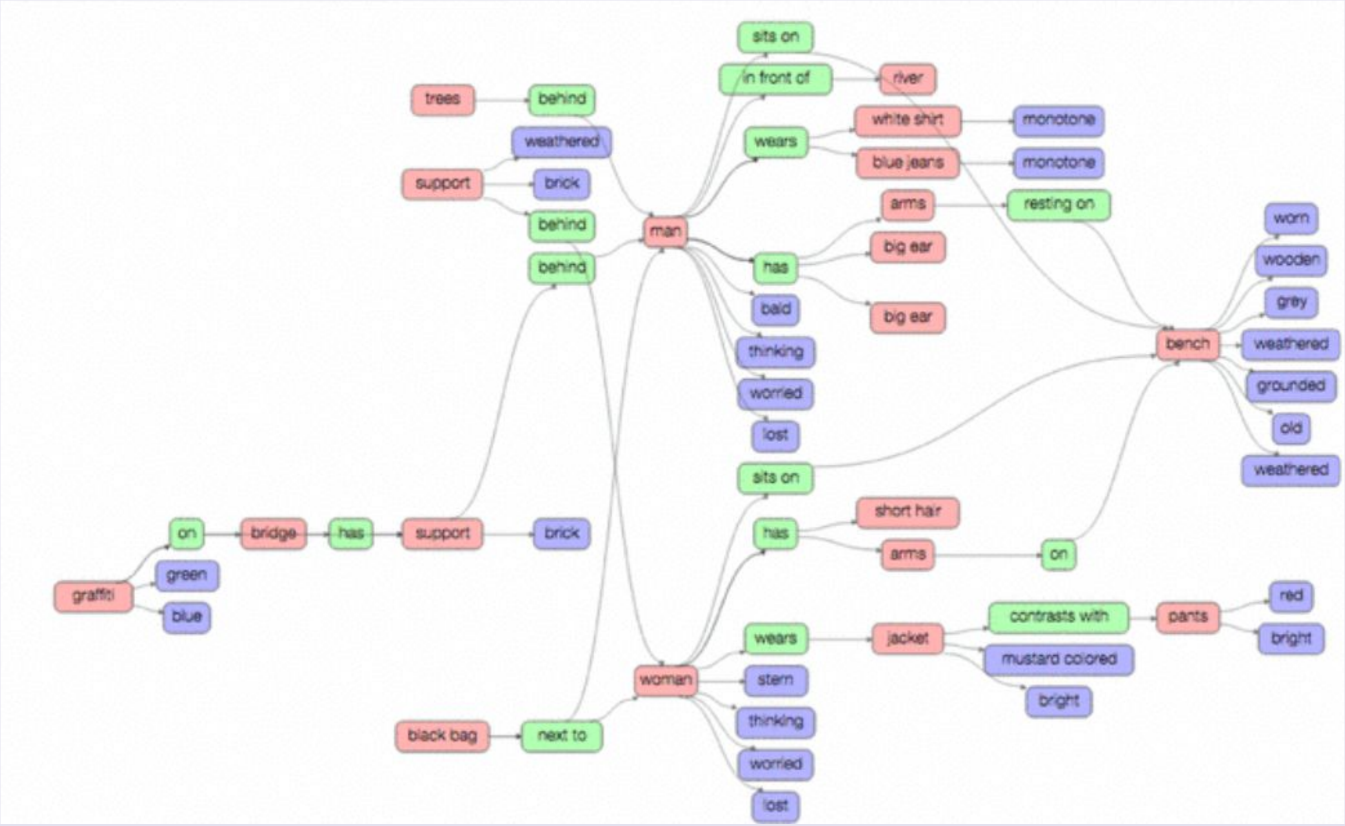
bench → grey

bench → weathered

Park bench is made of gray weathered wood

man → bald

The man is almost bald


sits on
in front of
river
trees
behind
white shirt
monotone
wears
weathered
blue jeans
monotone
support
brick
arms
resting on
behind
man
worn
has
big ear
wooden
behind
bald
big ear
grey
thinking
bench
weathered
worried
grounded
old
lost
weathered
sits on
short hair
on
bridge
has
support
brick
has
green
arms
on
graffiti
red
blue
contrasts with
pants
wears
jacket
bright
mustard colored
woman
stern
bright
thinking
black bag
next to
worried
lost

Project Group 8

Academic Optimization: Personalized Course Recommender

Daniel Lyczak

CS8803: 'Team' 8

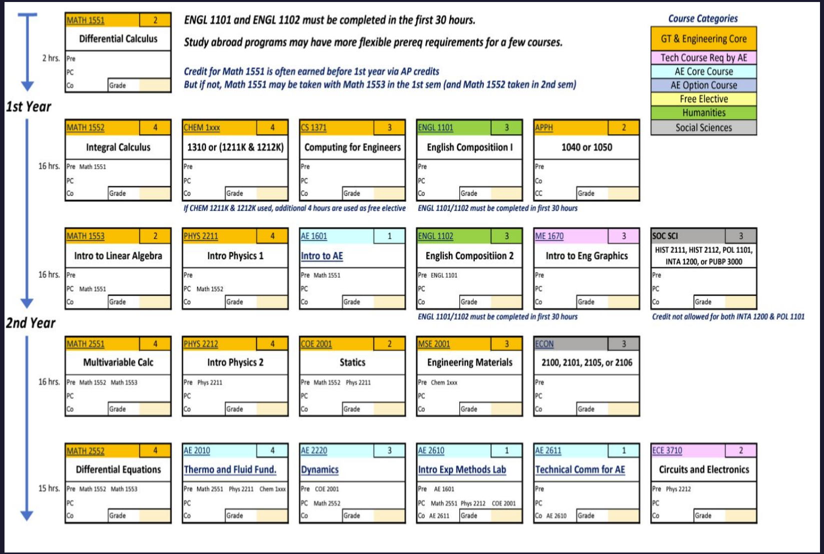
Introduction

Course Enrollment is static and generalized

More personalized tools are not oriented around the target student

Course interaction influence student satisfaction

Georgia Tech has data, a known problem, and…?



ENGL 1101 and ENGL 1102 must be completed in the first 30 hours.

Study abroad programs may have more flexible prereq requirements for a few courses.

Credit for Math 1551 is often earned before 1st year via AP credits
But if not, Math 1551 may be taken with Math 1553 in the 1st sem (and Math 1552 taken in 2nd sem)

Course Categories

- GT & Engineering Core
- Tech Course Req by AE
- AE Core Course
- AE Option Course
- Free Elective
- Humanities
- Social Sciences

1st Year

2 hrs.

MATH 1551	2
Differential Calculus	
Pre	
PC	
Co	Grade

16 hrs.

MATH 1552 4	CHEM 1xxx 4	CS 1371 3	ENGL 1101 3	APPH 2
Integral Calculus	1310 or (1211K & 1212K)	Computing for Engineers	English Compositiion I	1040 or 1050
Pre Math 1551	Pre	Pre	Pre	Pre
PC	PC	PC	PC	Co
Co Grade	Co Grade	Co Grade	Co Grade	CC Grade

If CHEM 1211K & 1212K used, additional 4 hours are used as free elective

ENGL 1101/1102 must be completed in first 30 hours

16 hrs.

MATH 1553 2	PHYS 2211 4	AE 1601 1	ENGL 1102 3	ME 1670 3	SOC SCI 3
Intro to Linear Algebra	Intro Physics 1	Intro to AE	English Compositiion 2	Intro to Eng Graphics	HIST 2111, HIST 2112, POL 1101, INTA 1200, or PUBP 3000
Pre	Pre	Pre Math 1551	Pre ENGL 1101	Pre	Pre
PC Math 1551	PC Math 1552	PC	PC	PC	PC
Co Grade	Co Grade	Co Grade	Co Grade	Co Grade	Co Grade

ENGL 1101/1102 must be completed in first 30 hours

Credit not allowed for both INTA 1200 & POL 1101

2nd Year

16 hrs.

MATH 2551 4	PHYS 2212 4	COE 2001 2	MSE 2001 3	ECON 3
Multivariable Calc	Intro Physics 2	Statics	Engineering Materials	2100, 2101, 2105, or 2106
Pre Math 1552 Math 1553	Pre Phys 2211	Pre Math 1552 Phys 2211	Pre Chem 1xxx	Pre
PC	PC	PC	PC	PC
Co Grade	Co Grade	Co Grade	Co Grade	Co Grade

15 hrs.

MATH 2552 4	AE 2010 4	AE 2220 3	AE 2610 1	AE 2611 1	ECE 3710 2
Differential Equations	Thermo and Fluid Fund.	Dynamics	Intro Exp Methods Lab	Technical Comm for AE	Circuits and Electronics
Pre Math 1552 Math 1553	Pre Math 2551 Phys 2211 Chem 1xxx	Pre COE 2001	Pre AE 1601	Pre	Pre Phys 2212
PC	PC	PC Math 2552	PC Math 2551 Phys 2212 COE 2001	PC	PC
Co Grade	Co Grade	Co Grade	Co AE 2611 Grade	Co AE 2610 Grade	Co Grade

Proposal

Use enrollment data to build a recommender

Match students on academic indicators and goals

Similarity Index to recommendation pool

Leverage Markov Decision Processes and Set Theory

Recommend a schedule to meet target student goals

Dynamic, personalized, adaptable, and analytical

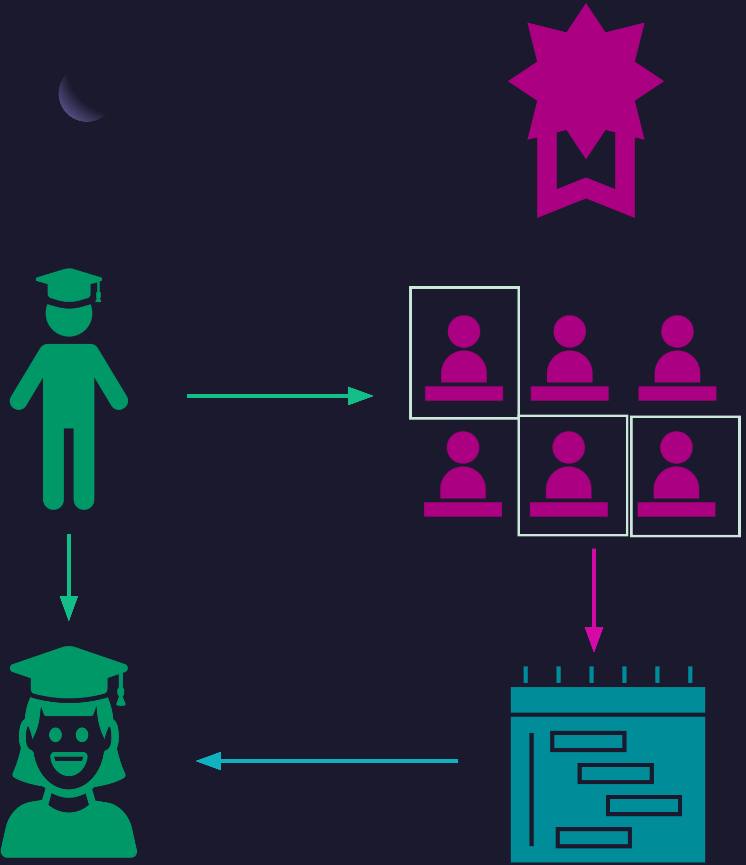
Tested and Evaluated

Check if algorithm can:

i) Match each student to a pool

ii) Generate a full-time schedule from pool

Attempt a historical 'what if'



Images taken from the noun project

Why it Matters

Course interaction leading indicator of college satisfaction

Academic achievement, most notably GPA, biggest contributor to student retention

Course offerings influence graduation rates, retention rates, departmental budgets, resources, rankings, family decisions, and many more

Most research in this area limited to simulation or narrow scope of courses

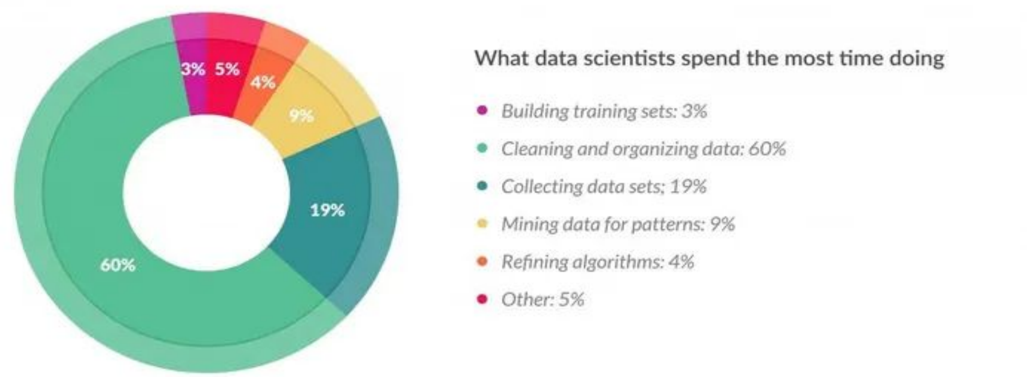
Thank you

Transferability of Data Pre-processing Pipelines

Project Group 9

Problem Statement

- Crafting effective data pre-processing pipelines can be time consuming and the results inefficient.
- Challenges:
  - A big discrete search space of possible pipelines
  - Need to avoid negative influence on the downstream model
  - Finding the right candidate given a resource budget



Bit

- Extensive work was done to automate the search for candidate pipeline, but with a limited and discrete search space.
- DiffPrep [1] solved this by considering a larger, continuous search space, but at the expense of training overhead.

Flip

- Reusability has always been a big boon for overall cost reduction
- We propose to reuse data pre-processing pipelines across different machine learning tasks and datasets based on the extent of transferability, thereby reducing the overhead involved during training.

Solution

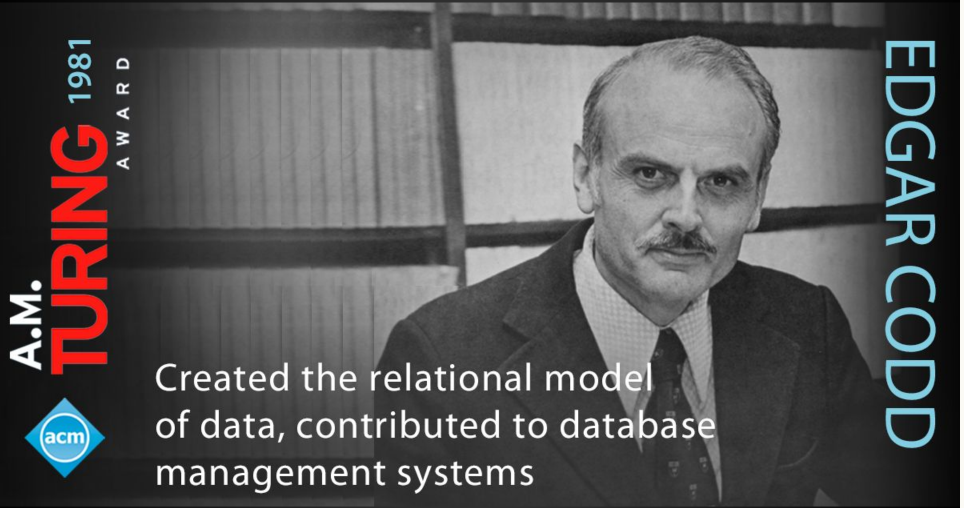
- Four datasets, two for each task, regression, and classification, and two from each domain, health and price/salary prediction, will be used to train a 3-layer Neural Network (NN) using DiffPrep[1].
- Experiments will be performed to evaluate the extent of transferability of the pre-processing pipeline, considering all combinations of tasks and data set domains.
- The model accuracy post-transfer and time spent fine-tuning the 3-layer NN using DiffPrep[1] will be used to calculate a Affinity matrix [2].
- This will provide a quantifiable measure of the extent of transferability between the source task, data set, and target task, data set.

[1]Peng Li, Zhiyi Chen, Xu Chu, and Kexin Rong. 2023. DiffPrep: Differentiable Data Preprocessing Pipeline Search for Learning over Tabular Data. Proc. ACM Manag. Data 1, 2, Article 183 (June 2023), 26 pages. https://doi.org/10.1145/3589328

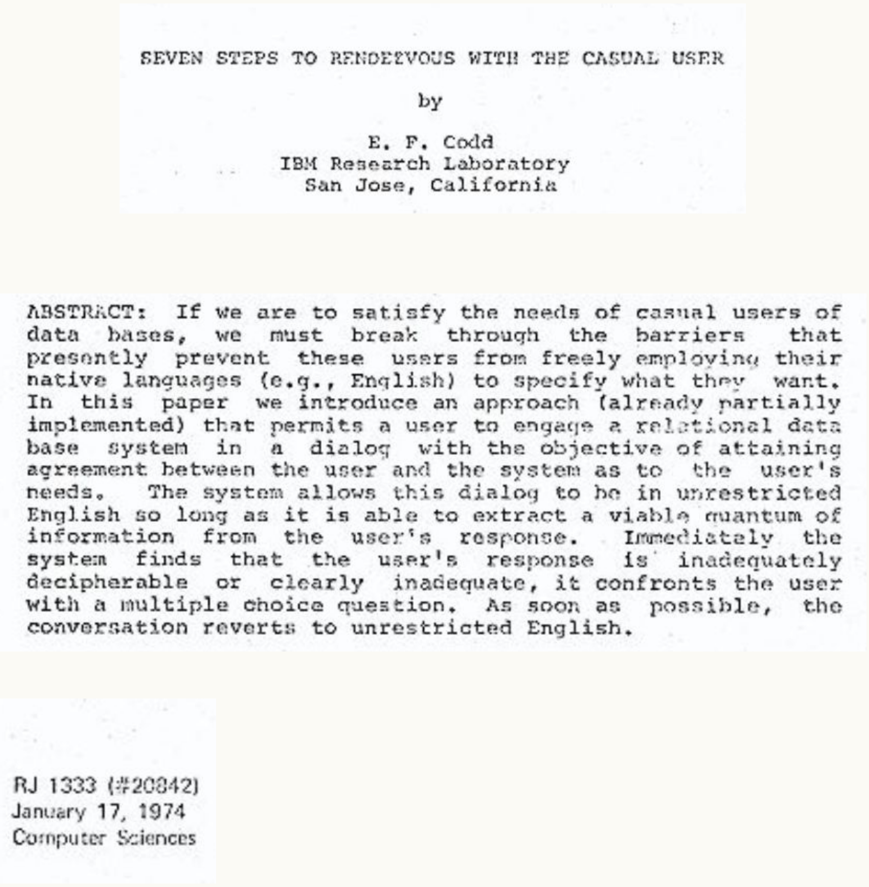
[2] Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling Task Transfer Learning.

Large Language Models as Commonsense Knowledge for Generalizable Natural Language Interface to Database

Project Group 10


A.M. TURING 1981 AWARD
acm
EDGAR CODD
Created the relational model of data, contributed to database management systems

"Honoring 50 Years of Visionaries and Their Enduring Legacies." *Spotlight on Turing Laureates*, awards.acm.org/about/turing-laureates-spotlight.



SEVEN STEPS TO RENDEZVOUS WITH THE CASUAL USER

by

E. F. Codd
IBM Research Laboratory
San Jose, California

ABSTRACT: If we are to satisfy the needs of casual users of data bases, we must break through the barriers that presently prevent these users from freely employing their native languages (e.g., English) to specify what they want. In this paper we introduce an approach (already partially implemented) that permits a user to engage a relational data base system in a dialog with the objective of attaining agreement between the user and the system as to the user's needs. The system allows this dialog to be in unrestricted English so long as it is able to extract a viable quantum of information from the user's response. Immediately the system finds that the user's response is inadequately decipherable or clearly inadequate, it confronts the user with a multiple choice question. As soon as possible, the conversation reverts to unrestricted English.

RJ 1333 (#20842)
January 17, 1974
Computer Sciences

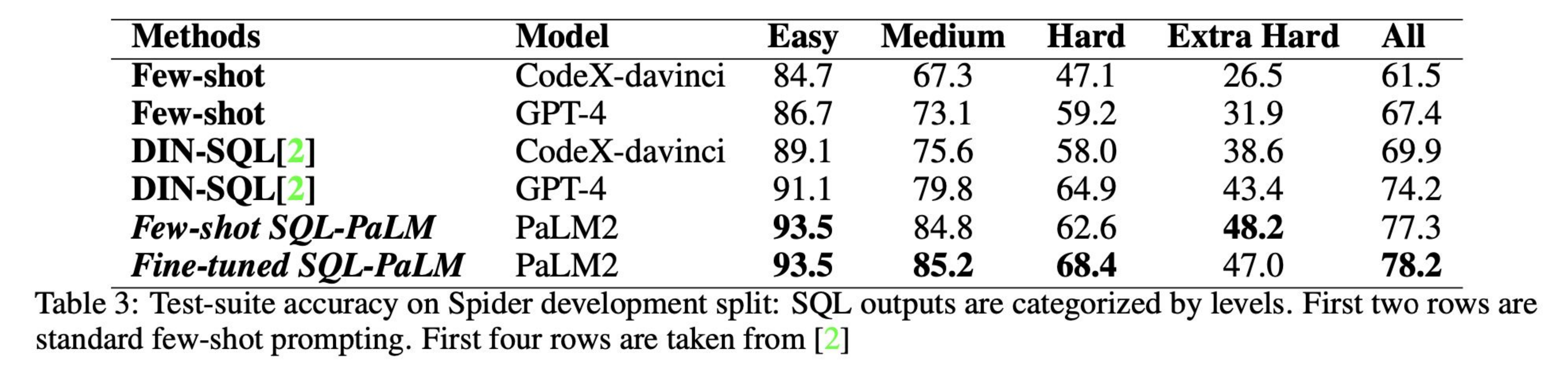
Codd, Edgar F. Seven Steps to Rendezvous with the Casual User. IBM Corporation, 1974.

Problem

- Real life enterprise data warehouses possess large and complex schemas
  - Case study with Credit Suisse
    - Requires days or weeks of collaboration between business users and database administrators need to…
      - ask ad-hoc queries
      - generate new reports
      - launch a new service
- A system that understands and translates natural language to SQL could save a lot of time.

Blunschi, Lukas, et al. “Soda: Generating Sql for Business Users.” ArXiv Preprint ArXiv:1207.0134, 2012.

The Bit

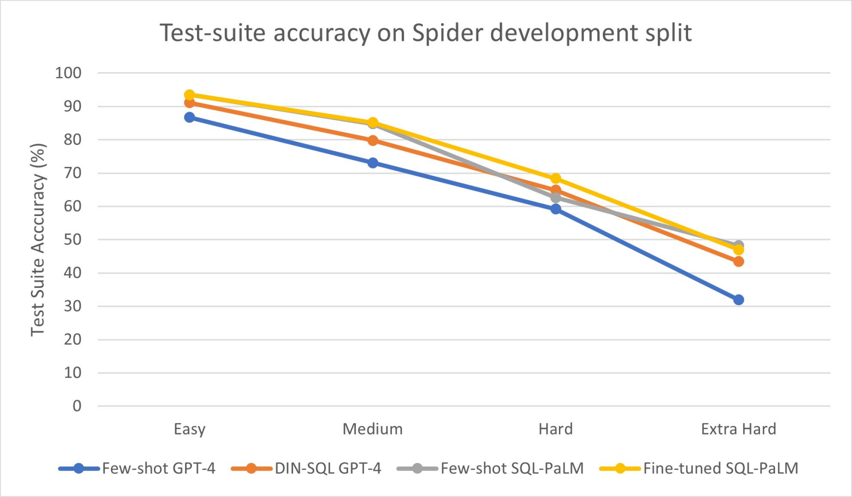


Methods	**Model**	**Easy**	**Medium**	**Hard**	**Extra Hard**	**All**
Few-shot	CodeX-davinci	84.7	67.3	47.1	26.5	61.5
Few-shot	GPT-4	86.7	73.1	59.2	31.9	67.4
DIN-SQL[2]	CodeX-davinci	89.1	75.6	58.0	38.6	69.9
DIN-SQL[2]	GPT-4	91.1	79.8	64.9	43.4	74.2
Few-shot SQL-PaLM	PaLM2	**93.5**	84.8	62.6	**48.2**	77.3
Fine-tuned SQL-PaLM	PaLM2	**93.5**	**85.2**	**68.4**	47.0	**78.2**

Table 3: Test-suite accuracy on Spider development split: SQL outputs are categorized by levels. First two rows are standard few-shot prompting. First four rows are taken from [2]

Sun, Ruoxi, et al. “SQL-PaLM: Improved Large Language Model Adaptation for Text-to-SQL.” ArXiv Preprint ArXiv:2306.00739, 2023.

The Bit



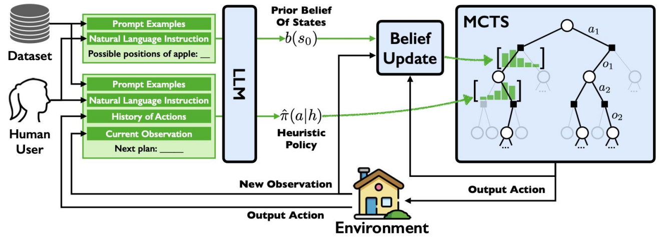
- Divided into four levels of difficulty based on whether solution requires
  - any nested sub-queries
  - column selections
  - aggregations

Yu, Tao, et al. "Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-Sql Task." ArXiv Preprint ArXiv:1809.08887, 2018.

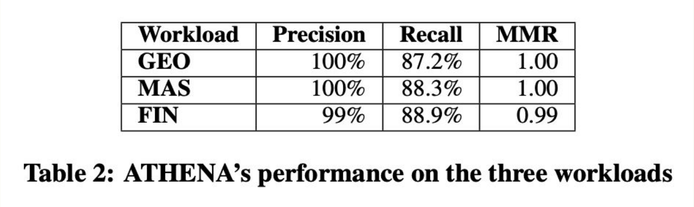
Bit-Flip

- The Bit
  - Use large language models as end-to-end neural machine translators
- The Flip
  - Leverage the knowledge within LLMs for extending prior rule-based systems that required humans to encode domain knowledge

The Flip



+



Workload	Precision	Recall	MMR
GEO	100%	87.2%	1.00
MAS	100%	88.3%	1.00
FIN	99%	88.9%	0.99

Table 2: ATHENA's performance on the three workloads

Ahn, Michael, et al. “Do as i Can, Not as i Say: Grounding Language in Robotic Affordances.” ArXiv Preprint ArXiv:2204.01691, 2022.
Zhao, Zirui, et al. “Large Language Models as Commonsense Knowledge for Large-Scale Task Planning.” ArXiv Preprint ArXiv:2305.14078, 2023.

Plan

- Integrate LLama 2 as a source of common knowledge to rule-based systems such as Sqlizer (handwritten repair rules), ATHENA (predefined ontology), and TEMPLAR (query log information)
    - Llama 2 - Meta's large language model pre-trained on 2 trillion tokens (access to weights available)
- Compare execution accuracy and and test-suite accuracy with current SOTA models using Bird-Bench, and Spider variants—Spider- Syn and Spider-Realistic.

Baik, Christopher, et al. “Bridging the Semantic Gap with SQL Query Logs in Natural Language Interfaces to Databases.” 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 374–85.
Saha, Diptikalyan, et al. “ATHENA: An Ontology-Driven System for Natural Language Querying over Relational Data Stores.” Proceedings of the VLDB Endowment, vol. 9, no. 12, VLDB Endowment, 2016, pp. 1209–20.
Yaghmazadeh, Navid, et al. “SQLizer: Query Synthesis from Natural Language.” Proceedings of the ACM on Programming Languages, vol. 1, no. OOPSLA, ACM New York, NY, USA, 2017, pp. 1–26.
Touvron, Hugo, et al. “Llama 2: Open Foundation and Fine-Tuned Chat Models.” ArXiv Preprint ArXiv:2307.09288, 2023.
Rubin, Ohad, and Jonathan Berant. “SmBoP: Semi-Autoregressive Bottom-up Semantic Parsing.” ArXiv Preprint ArXiv:2010.12412, 2020.
Li, Jinyang, et al. “Can Llm Already Serve as a Database Interface? A Big Bench for Large-Scale Database Grounded Text-to-Sqls.” ArXiv Preprint ArXiv:2305.03111, 2023.
Gan, Yujian, et al. “Towards Robustness of Text-to-SQL Models against Synonym Substitution.” ArXiv Preprint ArXiv:2106.01065, 2021.
Deng, Xiang, et al. “Structure-Grounded Pretraining for Text-to-SQL.” CoRR, vol. abs/2010.12773, 2020, https://arxiv.org/abs/2010.12773.

Thank You!

References

- “Honoring 50 Years of Visionaries and Their Enduring Legacies.” Spotlight on Turing Laureates, awards.acm.org/about/turing-laureates-spotlight. Accessed 25 Sept. 2023.
- Codd, Edgar F. Seven Steps to Rendezvous with the Casual User. IBM Corporation, 1974.
- Blunschi, Lukas, et al. “Soda: Generating Sql for Business Users.” ArXiv Preprint ArXiv:1207.0134, 2012.
- Sun, Ruoxi, et al. “SQL-PaLM: Improved Large Language Model Adaptation for Text-to-SQL.” ArXiv Preprint ArXiv:2306.00739, 2023.
- Rubin, Ohad, and Jonathan Berant. “SmBoP: Semi-Autoregressive Bottom-up Semantic Parsing.” ArXiv Preprint ArXiv:2010.12412, 2020.
- Yu, Tao, et al. “Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-Sql Task.” ArXiv Preprint ArXiv:1809.08887, 2018.
- Li, Yunyao, et al. Natural Language Interfaces to Databases. Springer Nature, 2023, https://nlidb.github.io/book/.
- Zhao, Zirui, et al. “Large Language Models as Commonsense Knowledge for Large-Scale Task Planning.” ArXiv Preprint ArXiv:2305.14078, 2023.
- Ahn, Michael, et al. “Do as i Can, Not as i Say: Grounding Language in Robotic Affordances.” ArXiv Preprint ArXiv:2204.01691, 2022.
- Baik, Christopher, et al. “Bridging the Semantic Gap with SQL Query Logs in Natural Language Interfaces to Databases.” 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 374–85.
- Saha, Diptikalyan, et al. “ATHENA: An Ontology-Driven System for Natural Language Querying over Relational Data Stores.” Proceedings of the VLDB Endowment, vol. 9, no. 12, VLDB Endowment, 2016, pp. 1209–20.
- Yaghmazadeh, Navid, et al. “SQLizer: Query Synthesis from Natural Language.” Proceedings of the ACM on Programming Languages, vol. 1, no. OOPSLA, ACM New York, NY, USA, 2017, pp. 1–26.
- Touvron, Hugo, et al. “Llama 2: Open Foundation and Fine-Tuned Chat Models.” ArXiv Preprint ArXiv:2307.09288, 2023.
- Li, Jinyang, et al. “Can Llm Already Serve as a Database Interface? A Big Bench for Large-Scale Database Grounded Text-to-Sqls.” ArXiv Preprint ArXiv:2305.03111, 2023.
- Gan, Yujian, et al. “Towards Robustness of Text-to-SQL Models against Synonym Substitution.” ArXiv Preprint ArXiv:2106.01065, 2021.
- Deng, Xiang, et al. “Structure-Grounded Pretraining for Text-to-SQL.” CoRR, vol. abs/2010.12773, 2020, https://arxiv.org/abs/2010.12773.