Personalizing Educational Content through Retrieval Augmented Generation

ISEA Session 8

Min Sun and Shawon Sarkar
University of Washington
March 15, 2024

IES Institute of Education Sciences
AmplifyLearn.AI
UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS
UNIVERSITY OF OREGON

Housekeeping

- Hackweek lodging and travel information collection:

- **Claude—another LLM model that gains popularity**

Overview of today

1. Sentiment analysis performed by LLMs
2. RADAR for personalized search and content generation
3. Model evaluation

Sentiment Analysis

Domain-specific definition of sentiment

prompt_base = f"""
Act as a policy researcher, you will classify the sentiment in the interviews of educational policy stakeholders as: “Positive”, “Negative”, or “Neutral”. Here is a statement from a policy stakeholder:
[TextGoHere]
To warrant “Positive” sentiment, the statement has to: (1) include the interviewee’s satisfaction about an educational policy (policies) and program(s), or (2) express an enhancement or potential to enhance the quality or equity of student learning or school system, or (3) identify an improvement from past practice. To warrant “Negative”, the statement describes the interviewees’ dissatisfactions, or identifies problems/issues/challenges, or suggests areas needed for further improvement. When the interviewee just states the fact without expressing either positive or negative sentiment, you can classify as “neutral”. When multiple sentiments are observed in one statement, identify the most prevailing sentiment. Explain your reasoning for your analysis."""

Lexical-based sentiment analysis:
nltk.sentiment.vader package in Python

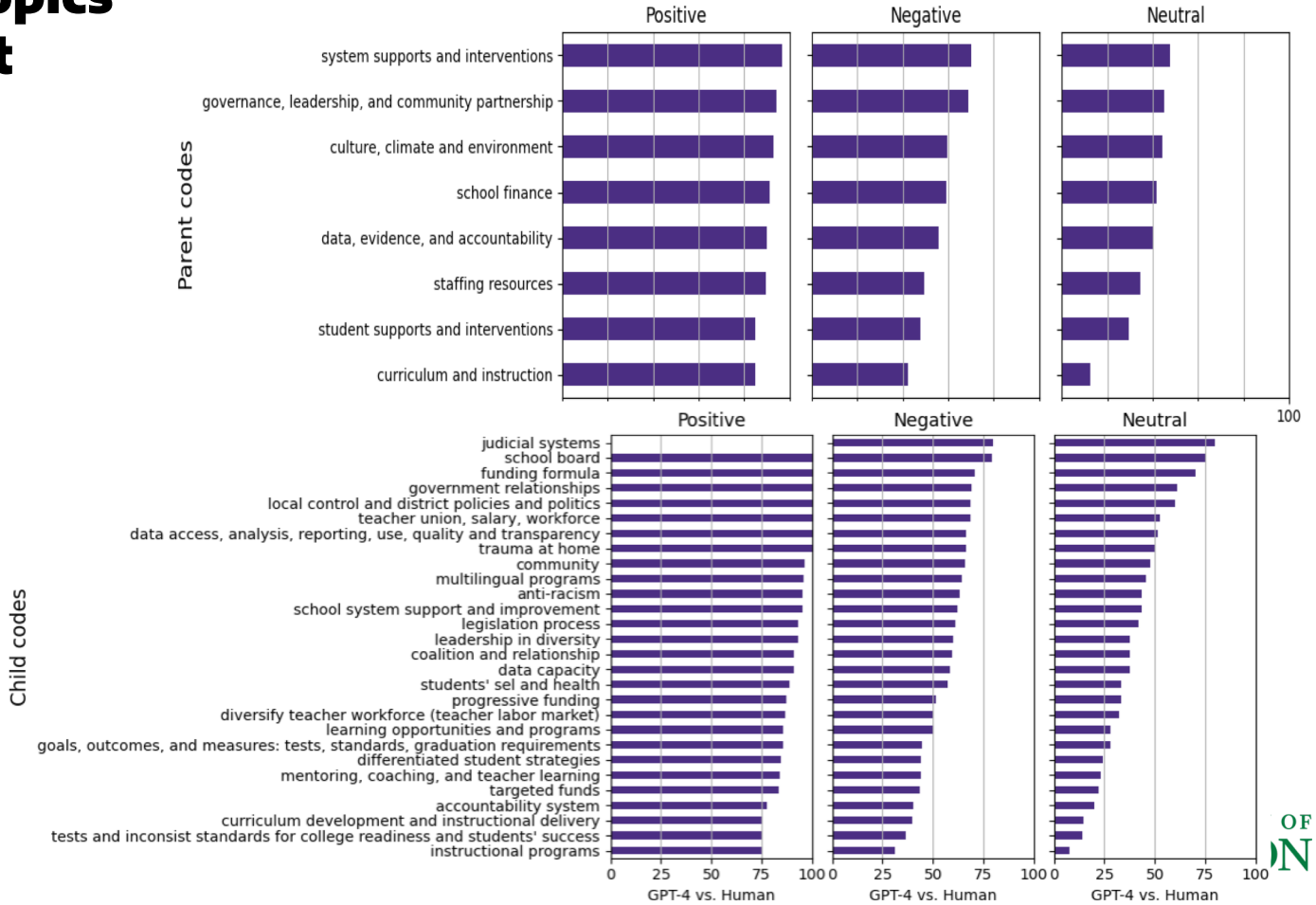
Sentiment Analysis

Confusion Matrix and Performance Metrics for Sentiment Analysis

	GPT-4 →			Lexicon →			Performance Metrics		
Human ↓	Positive	Negative	Neutral	Positive	Negative	Neutral		Accuracy	Cohen's κ
Positive	218	4	20	215	11	16	GPT-4 vs.Human	0.58	0.38
Negative	71	322	162	347	151	57	LDA vs. Human	0.31	0.09
Neutral	31	31	215	405	64	43			

GPT-4 is doing much better job than lexicon-based approach.
Agarwal et al. [2019] saw $\kappa = 0.44$ for news sentiment

Associating topics and sentiment



IES Institute of Education Sciences

Associating with outcome measures

Sun, M., Liu, J., Zhu, J., & LeClair, Z. (2019). Using a text-as-data approach to understand reform processes: A deep exploration of school improvement strategies. *Educational evaluation and policy analysis*, *41*(4), 510-536.

TABLE 5

The Associations Between Reform Topics and School Average Student Absences

Reform topics	Full-day absences	Part-day absences	% chronic full-day absences
Topic 1. Interventions and supports for promoting positive student behaviors	1.599 (3.177)	−3.147 (3.560)	0.055 (0.082)
Topic 2. General parent and community outreach	−7.447 (4.423)	−2.375 (4.784)	−0.217 (0.120)
Topic 3. Engaging parents about student academic and behavioral learning in schools	−3.820 (2.408)	−4.656 (3.233)	−0.090 (0.067)
Topic 4. Planning, providing, and evaluating professional development for instructional improvement	−6.200 (3.361)	0.243 (5.518)	−0.153 (0.096)
Topic 5. Monitoring student progress and using data to develop interventions	9.895 (5.276)	5.699 (5.518)	0.210 (0.137)
Topic 7. Using assessment data to identify students for targeted support	3.652 (3.971)	1.387 (4.683)	0.168 (0.120)
Topic 8. Extending instructional time and aligning curriculum/ assessments to standards	7.178 (4.585)	0.229 (4.402)	0.164 (0.114)
Topic 9. Teacher team (e.g., grade-level team, PLC) activities (e.g., reviewing data, planning, aligning standards, interventions)	−10.330*** (2.737)	−16.750*** (4.364)	−0.241*** (0.068)
Topic 10. Administering common assessments and disaggregating data to differentiate interventions	−16.36** (5.867)	−14.06* (6.961)	−0.396** (0.144)
Topic 11. Leadership teams setting goals and reviewing data for school improvement	6.686 (5.018)	3.876 (5.153)	0.131 (0.106)
Topic 12. Teacher instructional improvement via walkthroughs, observations, and feedback	1.701 (2.955)	6.552 (4.111)	0.0736 (0.075)
Topic 15. Setting goals for and recognizing teachers' and students' growth	1.777 (3.524)	0.389 (4.316)	0.0472 (0.085)
Topic 17. Extending learning time (or opportunities) for students and staff	1.651 (4.304)	10.32 (5.552)	0.0414 (0.113)
Topic 18. Collecting, analyzing, and aligning student assessments	−2.080 (6.625)	7.939 (8.064)	0.038 (0.177)
Topic 19. Improving special education	5.024 (4.834)	−1.479 (6.434)	0.137 (0.123)
N	599	599	599

Note. Each reform topic was added to the model separately. The models control for schools' prior achievement, the number of tasks schools were implementing in a given year, and school characteristics. The standard errors are clustered at school level. PLC = professional learning communities.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

IES Institute of Education Sciences

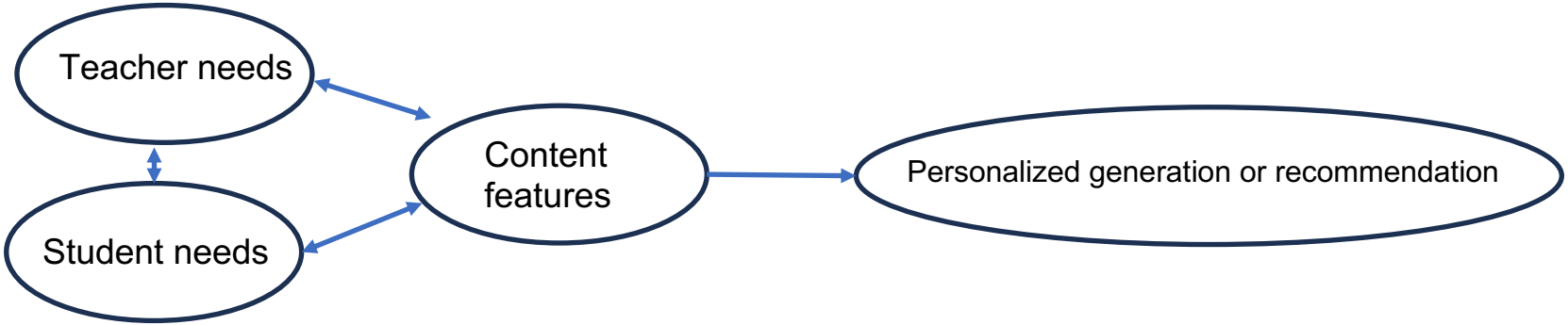
AmplifyLearn

AI's Use in Education: Motivate Today's Topic

- Generative AI's excitement centers on "Personalization" and "Adaptivity", which leads to efficiency in teaching and learning.
- To achieve this, needs tons of data analytics.
- Let's use Colleague.AI as a demo to see the problem space
- **Brainstorm: Decompose this task, what components do you need to personalize the recommendation?**

AI's Use in Education: Motivate Today's Topic

> Keep in mind the goal: "Personalization", "Adaptivity", "High Quality", and "Relevance" or "Accuracy".



RAG: The Problem to be Solved

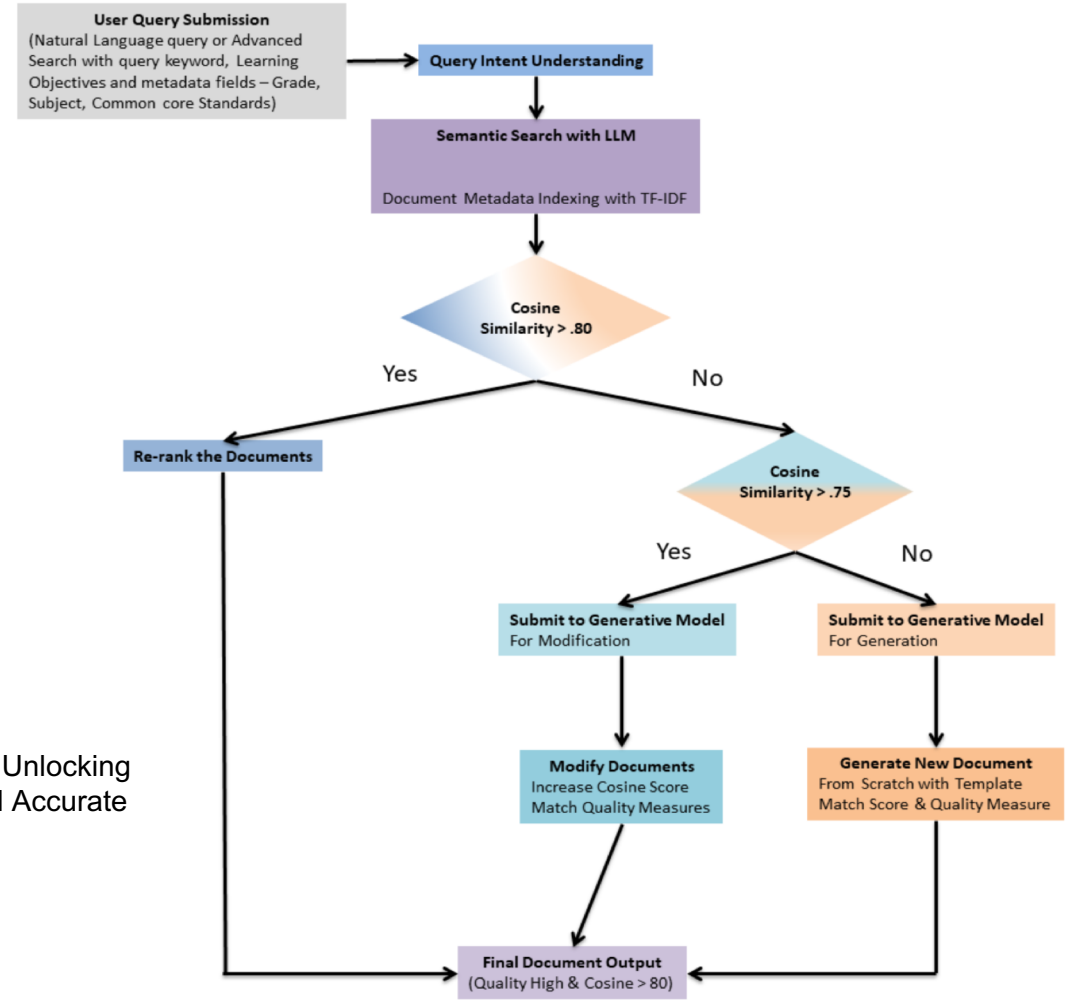
- Retrieved information may be more accurate and specific, but not be as relevant or personalized enough.
- Generated information can be relevant or personalized, dynamic, but may be less accurate or specific.
  - Hallucinate
  - Limited by context length
- RAG (Retrieval augmented generation)
  - Recommend relevant and personalized content
  - (May or may not) Reduce the computational cost
  - Combat hallucination and context length constraint
  - Embed domain knowledge into this process

RADAR

- Retrieve
- Assess (with cosine similarity)
- Decide (recommend, augment, or generate)
- Augment (if necessary)
- Regenerate (from scratch if below threshold)

RADAR in Colleague.AI

- Domain knowledge (quality math instruction) embedded in the model

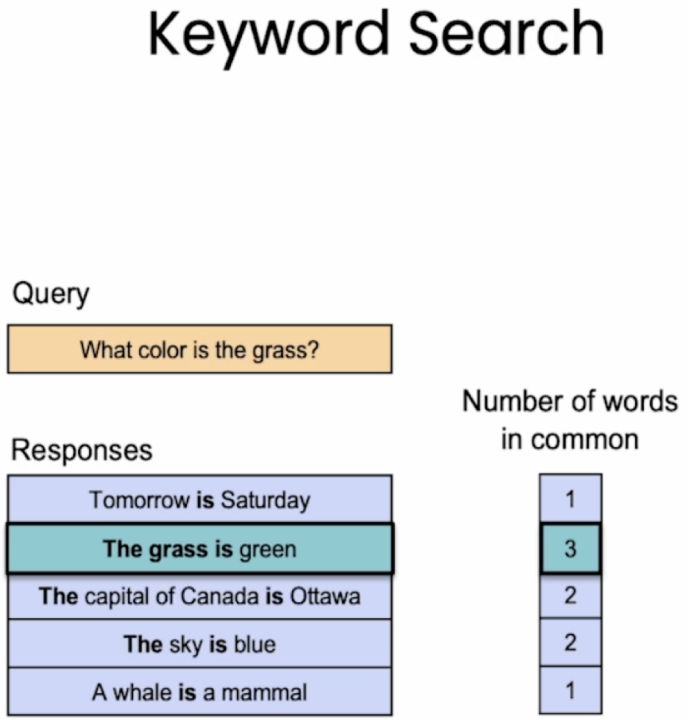


Citation: Sarkar, S., Sun, M., He, J., & Liu, A. (2024). Unlocking educational materials with RADAR: Personalized and Accurate math lesson plan search and generation.

IES Institute of Education Sciences

Semantic Search

- Doesn't account for the context
- Understanding Context and Intent: Utilizes the context surrounding a query and the user's intent to deliver more accurate results
  - Geographical location, search history, and the textual context of the query



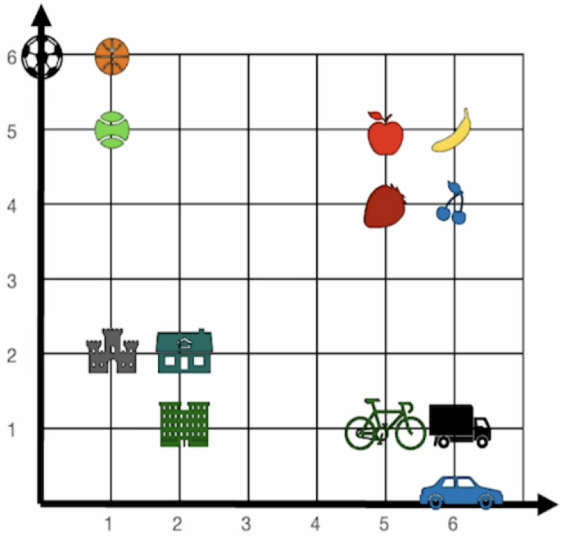
Cite: https://learn.deeplearning.ai/courses/large-language-models-semantic-search/lesson/1/introduction

Semantic Search

Vector Search and Embeddings: Converts queries and documents into vectors, numerical representations that allows for the comparison of their semantic similarities.

Embeddings

Quiz: Where would you put the word “apple”?



Word	Numbers	
Apple	?	?
Banana	6	5
Strawberry	5	4
Cherry	6	4
Soccer	0	6
Basketball	1	6
Tennis	1	5
Castle	1	2
House	2	2
Building	2	1
Bicycle	5	1
Truck	6	1
Car	6	0

Cite: https://learn.deeplearning.ai/courses/large-language-models-semantic-search/lesson/1/introduction

Text Embeddings

Sentence	Numbers				
Hello, how are you?	0.39	0.49	…	-1.01	-0.72
I'm going to school today	-0.79	-0.05	…	-0.94	2.71
…	…	…	…	…	…
Once upon a time	3.23	-0.23	…	-1.45	0.82
Hi, how's it going?	0.41	0.48	…	-0.98	-0.66

Cite: https://learn.deeplearning.ai/courses/large-language-models-semantic-search/lesson/1/introduction

Embeddings

Embedding methods refer to techniques used to represent data in a dense vector space, where similar items are mapped to nearby points (vectors) and dissimilar items are mapped to distant points. **Word2Vec, GloVe (Global Vectors for Word Representation), and FastText**

1. BERT (Bidirectional Encoder Representations from Transformers): A transformer-based model that learns contextual embeddings through self-attention and bidirectional training.
    - Pros: State-of-the-art performance on many NLP tasks, captures long-range dependencies
    - Cons: Large model size, requires significant computational resources
    - Best for: High-performance NLP tasks like text classification, sentiment analysis
2. GPT (Generative Pre-trained Transformer): A transformer-based language model that generates contextual embeddings and can be fine-tuned for various tasks.
    - Pros: Excels at text generation, can be fine-tuned for various tasks
    - Cons: Large model size, can be resource-intensive
    - Best for: Text generation, summarization, translation, question answering

https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

Cosine Similarity in Semantic Search

Using algorithms like k-nearest neighbors (kNN) or cosine similarity to find and rank the most conceptually relevant documents

- **Cosine Similarity Calculation**: The cosine similarity between the vector of the search query and the vectors of each document in the dataset is calculated. This similarity score ranges from -1 to 1, where 1 indicates identical direction (high similarity), 0 indicates orthogonality (no similarity), and -1 indicates opposite direction (high dissimilarity).
- **Ranking**: Documents are ranked based on their cosine similarity scores with respect to the search query vector. Higher scores indicate more relevant documents to the query's semantic content.
- **Retrieval:** The ranked list of documents is presented as search results, with the most conceptually relevant documents appearing at the top.

Retrieval with LLM

Data: A collection of documents or information items we want to search through

Preprocess data: Clean and prepare the data (e.g., normalization, tokenization)

Embedding generation: Use a LLM to generate embeddings for each document

Index embeddings: Store the generated embeddings and metadata in an index

Process query: Clean and preprocess query (similar to data preprocessing)

Generate query embedding: Use the same LLM to embed the query

Similarity calculation: Compare the query embedding to the document embeddings using a similarity measure (e.g., cosine similarity) to determine relevance

Rank retrieve results: Order the documents based on their similarity scores, with the highest scores indicating the most relevant documents

Return results: Present the ranked list of documents to the user

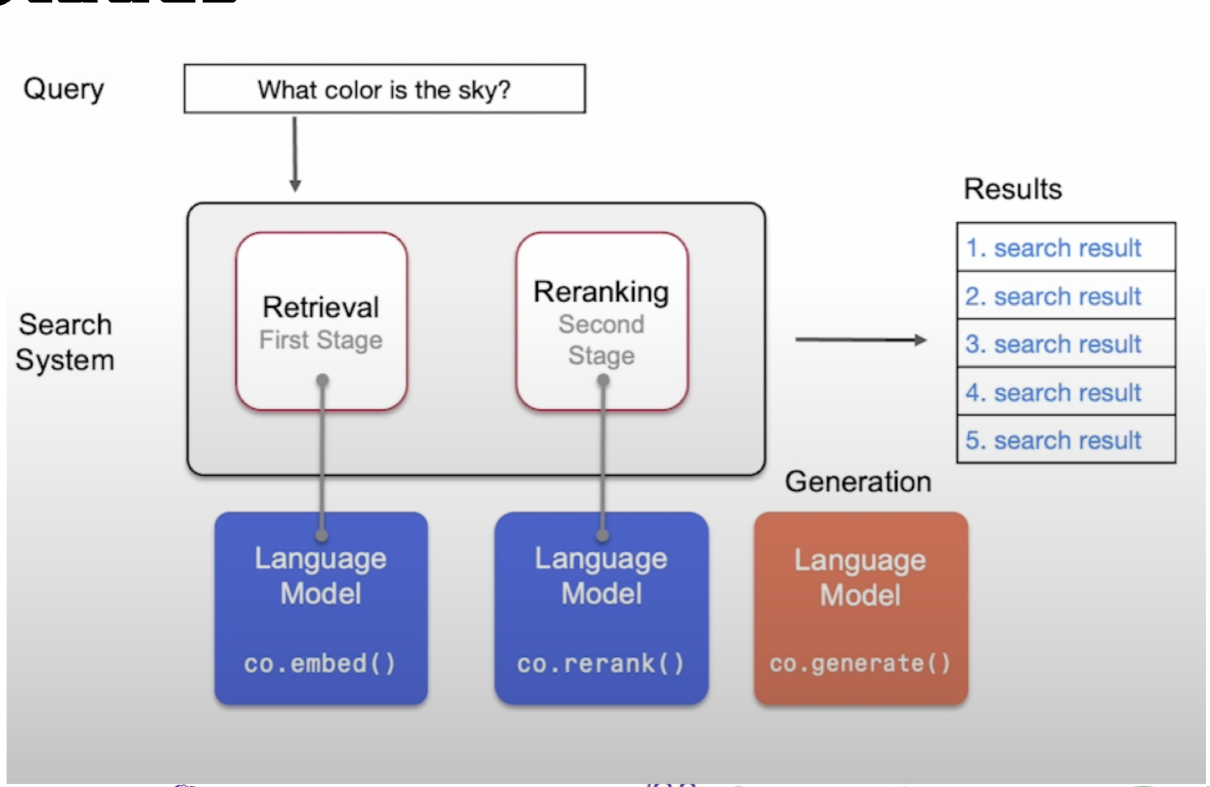
```
embedding_model = "text-embedding-3-small"
embedding_encoding = "cl100k_base"
max_tokens = 8000
```

Source: OpenAI

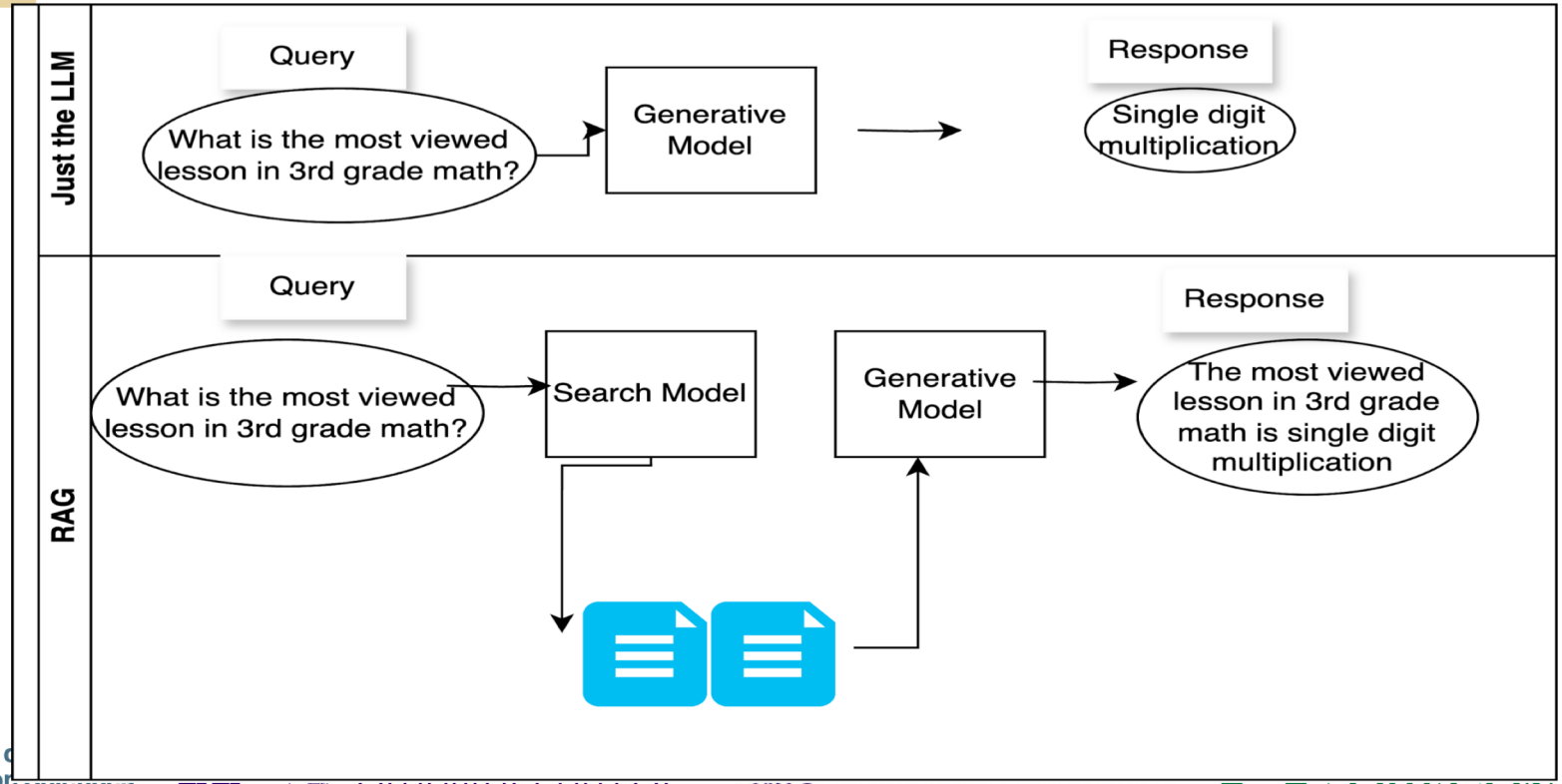
```
distances = distances_from_embeddings(query_embedding, doc_embedding, distance_metric="cosine")
```

Source: OpenAI

Language Models can Improve both Search Stages



RAG with LLM


Just the LLM
Query
What is the most viewed lesson in 3rd grade math?
Generative Model
Response
Single digit multiplication
RAG
Query
What is the most viewed lesson in 3rd grade math?
Search Model
Generative Model
Response
The most viewed lesson in 3rd grade math is single digit multiplication

Modification with Prompt-tuning LLM

Choose a pre-trained LLM suitable for the task

Design prompt: Design and experiment with different prompts

Input to a generative model: A user query, a document template, a custom prompt and a similarity threshold

Similarity Calculation: If the document's similarity score meets or exceeds the similarity threshold, finalize and return the lesson plan

Continue Refinement: Refine the document by reiterating generation and assessment steps until it meets the threshold or the process is manually stopped

```
prompt = f"Refine the following document to improve its similarity with the query: \nDocument: {top_document}\nQuery: {user_query}\n"

refined_document = interact_with_gpt3.5(prompt, dataset, user_query, threshold_high)
```

Generation with Prompt-tuning LLM

Choose a pre-trained LLM suitable for the task

Design prompt: Design and experiment with different prompts

Input to a generative model: A user query, a document template, a custom prompt and a similarity threshold

Similarity Calculation: If the document's similarity score meets or exceeds the similarity threshold for query, finalize and return the document

Continue Refinement: Refine the document by reiterating generation and assessment steps until it meets the threshold or the process is manually stopped

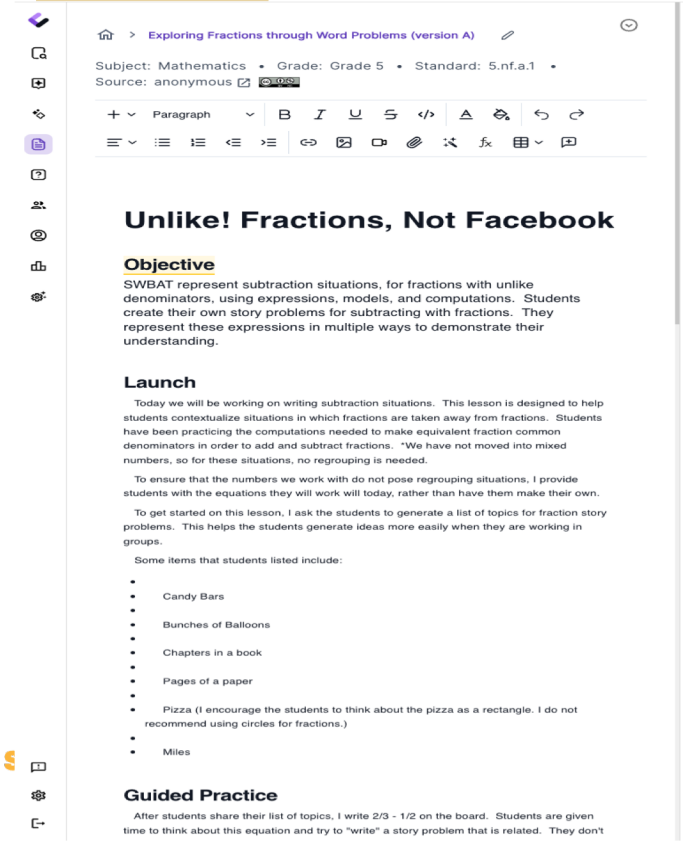
```
prompt = "Create a new document based on the following information:\n"
template = "Title: [Title of the document]\nLearning Objectives: [Main learning objectives]\nSection1: [....]\nSection2: [....]"

new_document_info = generate_gpt3_response(prompt + template)
```

Evaluation – Performance Statistics

Table 1: Evaluation results for different IR models and Colleague-IRGen

IR Model	Precision	Recall	F1-Score	Accuracy
BM25	0.82	0.83	0.84	0.84
TF-IDF	0.82	0.89	0.85	0.87
ColBERT	0.87	0.90	0.92	0.92
Colleague-IRGen (GPT-3.5-turbo)	0.92	0.95	0.94	0.95
Colleague-IRGen (GPT-4)	0.95	0.97	0.96	0.97

Evaluation -2: Human Judgement and Feedback



Exploring Fractions through Word Problems (version A)

Subject: Mathematics • Grade: Grade 5 • Standard: 5.nf.a.1 •
Source: anonymous

Paragraph

Unlike! Fractions, Not Facebook

Objective

SWBAT represent subtraction situations, for fractions with unlike denominators, using expressions, models, and computations. Students create their own story problems for subtracting with fractions. They represent these expressions in multiple ways to demonstrate their understanding.

Launch

Today we will be working on writing subtraction situations. This lesson is designed to help students contextualize situations in which fractions are taken away from fractions. Students have been practicing the computations needed to make equivalent fraction common denominators in order to add and subtract fractions. *We have not moved into mixed numbers, so for these situations, no regrouping is needed.

To ensure that the numbers we work with do not pose regrouping situations, I provide students with the equations they will work will today, rather than have them make their own.

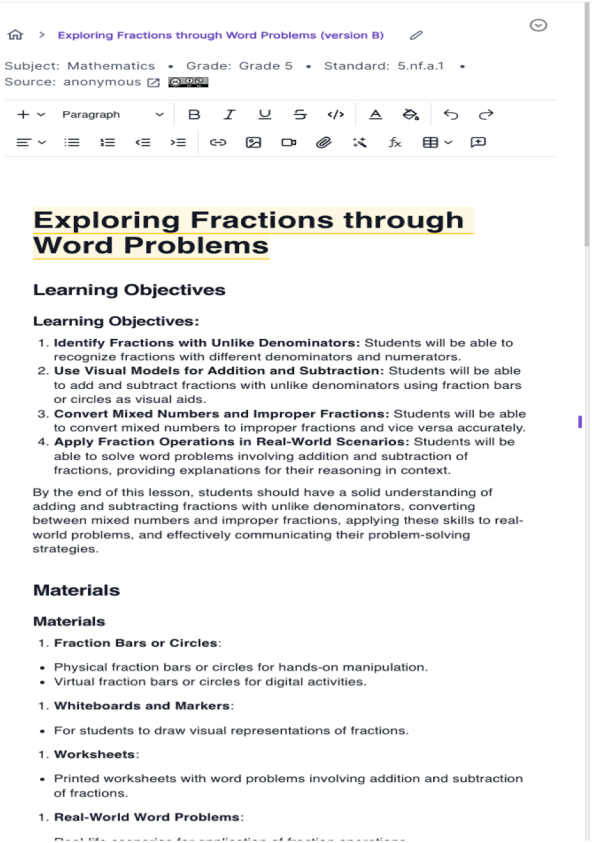
To get started on this lesson, I ask the students to generate a list of topics for fraction story problems. This helps the students generate ideas more easily when they are working in groups.

Some items that students listed include:

- Candy Bars
- Bunches of Balloons
- Chapters in a book
- Pages of a paper
- Pizza (I encourage the students to think about the pizza as a rectangle. I do not recommend using circles for fractions.)
- Miles

Guided Practice

After students share their list of topics, I write 2/3 - 1/2 on the board. Students are given time to think about this equation and try to "write" a story problem that is related. They don't



Exploring Fractions through Word Problems (version B)

Subject: Mathematics • Grade: Grade 5 • Standard: 5.nf.a.1 •
Source: anonymous

Paragraph

Exploring Fractions through Word Problems

Learning Objectives

Learning Objectives:

1. **Identify Fractions with Unlike Denominators:** Students will be able to recognize fractions with different denominators and numerators.
2. **Use Visual Models for Addition and Subtraction:** Students will be able to add and subtract fractions with unlike denominators using fraction bars or circles as visual aids.
3. **Convert Mixed Numbers and Improper Fractions:** Students will be able to convert mixed numbers to improper fractions and vice versa accurately.
4. **Apply Fraction Operations in Real-World Scenarios:** Students will be able to solve word problems involving addition and subtraction of fractions, providing explanations for their reasoning in context.

By the end of this lesson, students should have a solid understanding of adding and subtracting fractions with unlike denominators, converting between mixed numbers and improper fractions, applying these skills to real-world problems, and effectively communicating their problem-solving strategies.

Materials

Materials

1. **Fraction Bars or Circles**:

- Physical fraction bars or circles for hands-on manipulation.
- Virtual fraction bars or circles for digital activities.

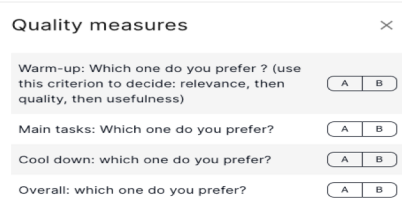
1. **Whiteboards and Markers**:

- For students to draw visual representations of fractions.

1. **Worksheets**:

- Printed worksheets with word problems involving addition and subtraction of fractions.

1. **Real-World Word Problems**:



Quality measures

Warm-up: Which one do you prefer ? (use this criterion to decide: relevance, then quality, then usefulness)	A B
Main tasks: Which one do you prefer?	A B
Cool down: which one do you prefer?	A B
Overall: which one do you prefer?	A B

Code Demo

> [Github repo]for some illustrative code, additional readings, and toy data.

Assignment

1. Review the first six weeks' materials.
2. Particularly if you do not have time to try out the codes and programs, please use this time to do so.
3. Catch up on additional readings.
4. Prepare for the session after Spring Break