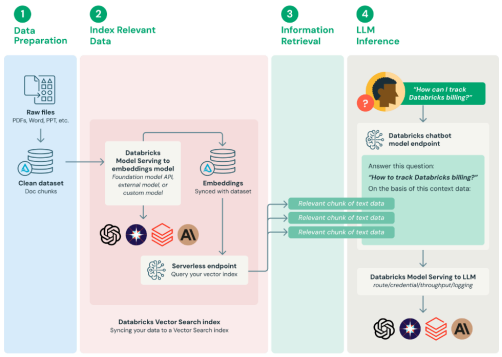
These answers serve as examples and may include additional details. For more accurate grading, please adhere to the grading rubrics.

Q1.



a. What is the above architecture related to? Just name it.

Solution:
RAG: Retrieval Augmentation, or 'external memory', or 'infinite memory'.

Rubrics:

- +1 for RAG or the other 2 terms
- -1 for any other answer/ any long explanations

b. Explain what it's for (its purpose, i.e. function, i.e. reason to exist).

Solution:
Retrieval-augmented generation (RAG) is optimizing the output of a large language model. Hence, it references external knowledge sources, such as documents or data relevant to a task or question, outside its training data sources before generating a response. This allows the LLM to create more accurate, context-aware, and transparent AI-generated content. i.e., **prevent hallucination**.

Rubrics:

- +1 for preventing hallucinations/ providing authentic sources for the LLM to refer
- -1 for irrelevant explanations

Q2.

Embedding data (text, images etc) usually occurs in orthogonal metric space, ie. defined by generalized XYZ... (to multiple, much higher # of dimensions or axes), where we use cosine similarity or Euclidean distance to measure similarity (ie. do search).

What are two applications of the above, in information retrieval? In other words, what can it enable, for end users? Describe each briefly.

Solution:

1. TF-IDF
2. Other forms of similarity search (eg using images etc)
3. Classification (via clustering) [eg of documents]
4. Other Applications of Semantic/Similarity Search

- Semantic Code Search and Documentation: Develop code search and documentation systems to understand the semantics of code snippets, functions, or APIs, enabling developers to find relevant code examples or documentation based on their intent or use case.
- Semantic Resume Matching and Job Recommendation: Build systems that can semantically match job descriptions with candidate resumes, skills, and experiences, enabling more accurate job recommendations and candidate sourcing.
- Semantic Search for...
    - Scientific Literature: Develop search engines for scientific publications and research papers that can understand the semantic relationships between concepts, methods, and findings, enabling researchers to discover relevant literature more effectively.
    - E-commerce Product Catalogs: Build search engines for e-commerce platforms that can understand the semantic relationships between product descriptions, features, and customer queries, enabling more accurate product searches and recommendations.
    - Legal and Regulatory Documents: Develop search systems for legal and regulatory documents to understand the semantic relationships between legal concepts, precedents, and cases, enabling lawyers and legal professionals to find relevant information more efficiently.

Rubrics:

- +1 for each relevant application (2) (Students can come up with other examples that are not listed above)
- -1 for irrelevant applications/ explanations of how embedded data can be used

Q3.
a. How do search engines (eg. Google's) currently serve spatial search results? Explain briefly.

Solution:

Currently, they serve a map, street view video, and also, possibly, additional facts via knowledge graphs (eg. photographs, videos, historic info). The search engine uses the following techniques to return [location-based or keyword-based] search results:

1. Using real-time geolocation (GPS) data, if location-based [eg by identifying keywords like 'near me', 'around me', 'in this area']
2. If keyword-based, identifying some location information like the place, city, country, etc, from the search query

Rubrics:

- +1 for any valid explanation
- -1 for wrong explanation
- 0 for any answer that seems too 'long', or 'copy-pasted from ChatGPT.'

b. ML-based advances are expected to lead to enhanced functionality for users doing spatial search. How specifically? Explain briefly.

Solution:

1. **'Photorealistic map tiles' (by Google) uses point clouds and rendering techniques, such as NeRF or 3D Gaussian Splatting, to output for the user, the ability to navigate the searched-for space, eg to 'go into' a restaurant, for ex**
2. ML can analyze user queries and past searches to understand the context of what users are looking for.
3. ML algorithms can learn user preferences and modify search results. If you frequently search for parks with playgrounds, ML can prioritize results that mention playgrounds in their descriptions.
4. Personalization based on user search history

Rubrics:

- +1 for any valid explanation (just one point is enough)
- -1 if it is not ML-based or futuristic.
- 0 for any answer that seems too 'long', or 'copy-pasted from ChatGPT.'

Q4.

'Code search' is a custom search domain, where developers perform searches specifically related to code (eg. on GitHub). The search engine for this would use a custom tf-idf structure, created using custom terms (ie. programming language constructs such as 'float' and 'switch'), and custom documents (source code files, eg. matrix.cpp).

List two other such 'custom' searches. For each, mention the terms and documents involved (similar to the 'code' example above).

Solution:

Example custom search domains-

- **Music: Searching audio based on small snippets of music (Shazam) or based on humming by user (Google Audio Search)**
  - **Terms: Musical genres, artist names, song titles etc**
  - **Documents: Music scores, lyrics, album reviews, artist biographies etc**
- **Legal Documents: Getting relevant legal analysis based on prior precedent**
  - **Terms: Legal terminology, case law references, etc.**
  - **Documents: Court cases, statutes, contracts, etc.**
- **Scientific Papers: Search research papers based on abstracts, specific domain and more**
  - **Terms: Scientific terminology, researchers, abstract etc.**
  - **Documents: Research papers, conference proceedings, academic dissertations, etc**

Rubrics:

- **other applications not mentioned above are ok too, if they make sense (eg searching through a collection of museum archives, or paintings, or rare books, etc), and mention relevant terms and returned document types**
- -1 for each missing application and explanation
- For each application
  - -0.5 if an application is correct, but no mention of terms and documents
  - -0.25 if an application is correct, but just mentioned terms
  - -0.25 if an application is correct, but just mentioned documents
- 0 for any answer that seems too 'long', or 'copy-pasted from ChatGPT.

Q5.

a. The current search engine advertising model is built around a linear (sequential) listing of result links. With the expected rise of LLM-based search summarization, this existing advertising model is in trouble.
How could the search 'vendors' (Google, Microsoft...) respond? In other words, what would they do, to continue advertising?

Solution:
They could explore integrating advertisements directly within the summarized search results generated by LLMs. These ads could be seamlessly woven into the summaries in a way that doesn't disrupt the user experience but still garners attention. For example if the user is searching for TVs, LLMs could directly suggest advertising links from Samsung/Sony. Or if a user is planning a trip, LLMs could serve AD links from airlines and hotels in a single response, creating a cohesive experience for the customer. Also, they could play a video [that can't be blocked, presumably] which shows an ad (or more), prior to the summarizing - which will be in the video as well, after the ad(s).

The answers could contain imaginative (novel) solutions, there are no good answers out there in the real-world just yet! We can accept any VALID answer [which needs to propose a workable solution].

Rubrics:

- +1 mark for valid explanation relevant to integrating ADs inside text generated using LLMs
- 0 for any answer that seems too 'long', 'irrelevent' or 'copy-pasted from ChatGPT.

b. How would AR-based search/retrieval aid in (revolutionize!!) education? Answer, by picturing students *looking* at things in the world...

Solution:
AR can revolutionize education by providing interactive, and contextual learning experiences. Sample answers-

- Overlay virtual elements onto the real world to study concepts - Human Anatomy for medical students, 3D blueprint visualization for civil students etc
- overlay animations and UI on math equations so students can interact with them - likewise for any symbolic 'notation' in general - molecular formulae, sheet music, architectural blueprints...

- AR Remote collaboration between teachers and students can help study concepts from any corner of the world. Can simultaneously interact with a virtual object and learn

Rubrics:

- +1 mark for each valid answer (just one point is enough)
- 0 for any answer that seems too ‘long’, or ‘copy-pasted from ChatGPT.

Q6.

In information retrieval, how do clustering and classification [which are two different sets of algorithms/techniques/applications) work together?
Explain.

Solution:
Clustering for document grouping to make groups of similar documents, *then* use classification for document tagging to place it in one of the clusters - ie. use this trained model to classify any new document.

Solution 2:
[Optional]
Apply clustering and classification iteratively to refine the grouping and classification of documents. For example, clustering algorithms can initially group documents into broad clusters, and then classification techniques can be used to further classify documents within each cluster into more specific categories or topics.

Solution 2 is similar to solution 1... it's ultimately about clustering first, then using the clusters to classify new items. Tangentially - it's a lot like NN, where the NN model (architecture + weights) is equivalent to the clustering - learn, then classify [instead of cluster, then classify].

Rubrics:
- -1 for not mentioning 'grouping of similar documents' through clustering algorithms
- -1 for not mentioning 'document tagging' or similar for classification
- +0 for any other answer

Q7.

a. What is the 'existential crisis' Google (and other search companies) had, when the web started experiencing sudden, massive growth? In other words, what was the problem?

Solution:

<Booming number of sites>

- major challenge in efficiently indexing and searching the increasingly large number of web pages. This made it difficult to maintain current indexes and deliver quick, relevant search results to users.

Rubrics:

- **+1 for correct answer**
- **0 for any other answer**

b. What technique did Google come up with, to address the issue? Explain briefly, IN YOUR OWN WORDS.

Solution:

MapReduce [aka Hadoop for the open source version]

- use map() to process data IN PARALLEL (eg. search for a query in a part (fragment) of the tf-idf index)
- use reduce() to aggregate the results (eg. rank, using PageRank and other metrics, eg user pref, site reputation etc)

Rubrics:

- **+1 for correct answer MUST mention map() for PARALLEL processing, reduce() for combining. (MUST mention MapReduce and/or Hadoop]**
- **0 for any other answer**

Q8.

a. How are knowledge graphs (KGs) constructed? Provide a high-level description.

Solution:

Knowledge graphs (KGs) are constructed by extracting entities and relationships from structured and unstructured data sources using techniques like entity linking and relation extraction. Additionally, they are built by collecting extensive data from various sources to identify and categorize key information, connecting concepts based on their relationships. [Optional - the technique is called 'NER', Named Entity Recognition].

Rubrics:

- +1 for correct answer based on the above
- 0 for wrong answer

b. How (where) do KGs and LLMs meet (get used together)?

Solution:

In RAG architecture, KGs and LLMs intersect in using KGs to guide LLM outputs with factual knowledge and constraints. In other words, a KG, rather than a pdf/text/.. or a relational DB or nonrelational DB, serves as 'external memory' used to answer the user query - the LLM then outputs the KG-retrieved results in natural language back to the user.

Rubrics:

- +1 for mentioning 'KG provided to LLM (ie used in conjunction with, ie in place of!) for retrieving factual knowledge'
- 0 for wrong answer

Q9.

a. *In your own words*, what is WordNet?

Solution:
WordNet is a hierarchical lexical database for the English language, where words are grouped into sets of synonyms called synsets and linked by semantic and lexical relations, providing a powerful resource for natural language processing.

It's ok if the answer doesn't mention synsets. The answer does need to mention a hierarchy of terms (can even say 'isA hierarchy'), with related words at each level in the hierarchy. Alternatively the answer can say that WordNet is a 'GRAPH of words' - where the graph nodes are the words themselves, and graph edges are word relationships (semantic - eg. similarity, opposites etc).

Any reasonable answer would mention words AND semantic relationships (eg synonyms) between them.

Rubrics:

- +1 for correct explanation of WordNet
- 0 for the wrong answer/ seems too 'long', or 'copy-pasted from ChatGPT.
- partial credit for a partial answer, eg. 'it is a hierarchy of words' [misses mentioning semantic relationships]

b. a. *In your own words*, what is ImageNet?

Solution:
ImageNet is a large-scale database of annotated images organized into a hierarchical structure similar to WordNet, widely used as a training and evaluation resource for computer vision and deep learning algorithms focused on image classification and object recognition tasks.

Or: ImageNet uses the structure of WordNet to create a comparable structure populated by images instead of words.

Rubrics:

- +1 for a correct explanation of ImageNet
- 0 for the wrong answer/ seems too 'long', or 'copy-pasted from ChatGPT.

Q10.

When AR (Augmented Reality) viewers become commonplace, commercial establishments (stores, restaurants...) can serve hyper-focused advertising to customers, matched to their buying/eating preferences. How exactly would such (advertising) work?

Solution:

1. AR overlay - Based on the user's preference, users can use their phone's camera as an AR display and overlay food content, shops, logos, and shopping places to guide them better.
2. 3D product view 'live' rendering - User can again place the virtual objects like couches, lamps, laptops, etc, on their floor/tables to see how would they look
3. Preference-based immersive advertising experience - for example, display a final look of the dish to users based on their preference even before they order, or create an experience where users can try out different clothes at home virtually.
4. virtual coupon labels overlaid on items, hyper-targeted to a user's past purchases

Rubrics:

- +2 for any valid explanation (just one point is enough)
- -1 partial point if explanation is vague and does not make much sense of how AR will be used
- -2 for a completely incorrect answer

Q11.

'Iteration' is a common solution technique employed in a variety of computations, eg. in the Newton-Raphson technique to calculate roots, in numerical formulae to calculate 'pi', etc.

In our course, we studied at least two algorithms that are based on iteration. Discuss which, making sure to state exactly WHAT is being iterated.

Solution:

[1] Breadth-First Search: iterate over web searches.
[2] K-Means: iterate over the centroid of each cluster.
[3] Page Rank: iterate over the page rankings.
[4] Spelling Correction: iterate over the misspelled word corrections.
[5] Snippet Generation: iterate over the potential snippets based on relevance score.
[6] Gradient Descent: iterate over the weights of layers.
[7] Crawler: iterate over the URL queues.
[8] Levenshtein Distance: iterate over modifications of one string into another.
[9] iterate (recurse) over links found in fetched web-pages, to populate a queue with links-to-crawl
...

Rubrics:

- +2 for mentioning two acceptable answers.
- -1 for mentioning just one acceptable answer.
- For each point
    - -0.5 for mentioning just the algorithm/method, and not what is being iterated
    - -0.5 for mentioning just what is being iterated, but not the name of the algorithm/method

Q12.

Information retrieval has had an astonishing evolution, beginning with the mainframe era, to what we have today.

Briefly trace the evolution, *in your own words*. You do need to mention salient changes along the way.

Solution:

Disk -> keyword search -> database -> query language -> search engine -> page ranking -> indexing and crawling -> cloud -> LLM -> SEMANTIC, multimodal, interactive retrieval

Rubrics:

- **+2 for mostly capturing these evolutions. (Student may write the above answer in their own words, please awards marks for those as well)**
- **+1 for barely capturing these evolutions.**
- **+0 for no answer.**

Q13, 1 point.

Create your own IR-related question, then answer it! Write the question, then the answer below it.

The answer CANNOT be, just one or two words, or a single line - it needs to be longer than that.

Solution:
Sample
Question:
What is a rich snippet? How is it different from a normal snippet?
Answer:

- A rich snippet is an enhanced search result that provides additional information beyond the basic title, URL, and description. It includes elements like ratings, images, or prices, making it more visually appealing and informative.
- Compared to a normal snippet, a rich snippet is more likely to attract user attention and clicks because it stands out in search results by offering more specific and relevant information about the content of the page.

Rubrics:

- Award full credits if any relevant question and answer is written that pertains to the content covered in class.
- -1 mark if the answer is less than 2 lines
- -1 mark if the question is irrelevant to the material covered

Q14, 1 point, bonus, Jeopardy-style: the answer is provided, you need to write a QUESTION for it that starts with 'What is' and ends with a '?' :)

Eg. A. CS572.
Q. What is the coolest course, at USC (or, ANYWHERE, EVERRR)?

A. This model encodes input data as a distribution in latent space.
Q.

Solution:
What is a VAE? or What is a Variational Auto Encoder?

Rubrics:
- Award full credits if the above question is written.
- -1 if anything else is mentioned.

Q15, 1 point, bonus, Jeopardy-style.

A. This is a more common name for an algorithm that Google invented in 2004. Q.

Solution:
What is MapReduce?

Rubrics:
- Award full credits if the above question is written.
- -1 if anything else is mentioned.
- Hadoop is incorrect - that came from Nutch, afterwards
- PageRank is incorrect - that was invented (mid/late 90s, around 1996) before Googe even existed

Q16, 1 point, bonus, Jeopardy-style.

A. This technique is an alternative, to the recommendation system based on others' preferences. Q.

Solution:
What is Content-Based Filtering? or What is CBF?

Rubrics:
- Award full credits if the above question is written.
- -1 if anything else is mentioned.

For Q14, Q15, Q16, it IS 1 or 0, because the answer is very specific!