

# AI BUILDERS SUMMIT

Workshop

RAG



## Stefan Webb

Developer Advocate  
Zilliz

# Evaluating Retrieval-Augmented Generation and LLM-as-a-Judge Methodologies

VIRTUAL JAN 15 - FEB 6

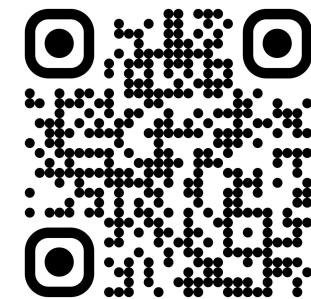


# Stefan Webb

Developer Advocate, Zilliz



[linkedin.com/in/stefan-webb](https://linkedin.com/in/stefan-webb)





# Prerequisites (Apple Silicon)

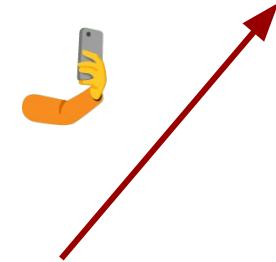
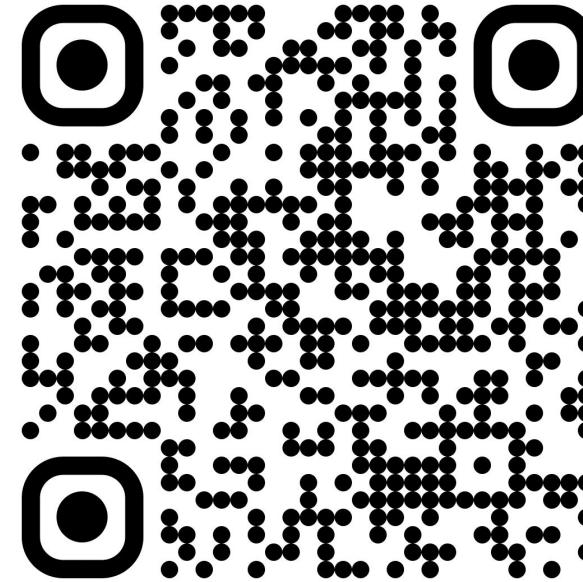
▲ Important ▲

```
> pip install -U ragas pymilvus llama-index  
transformers sentence-transformers mlx-lm==0.20.6
```

▲ Also Important ▲

Slides, Notebooks, Resources

QR Code to  
Webinar Material  
on Every Slide!



- 01 Introduction**
- 02 Evaluating Foundation Models**
- 03 Challenges and Limitations**
- 04 Coding Exploration**



# 01

## Introduction





# How does a vector database work?

## Sources

What is a Vector Database? - Elastic

elastic · 1

What is a Vector Database & How Does it Work? Use Cases + ...

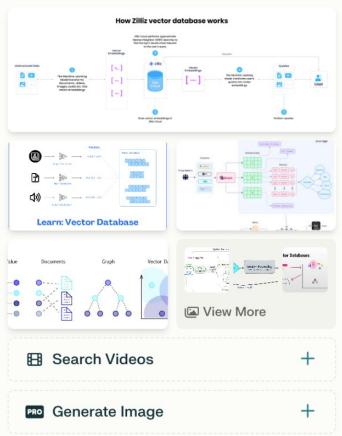
pinecone · 2

What are Vector Databases and How Do They Work? Use Cases + ...

zilliz · 3



View 4 more



## Perplexity

A vector database works by storing, indexing, and searching data represented as high-dimensional vectors. Here's an overview of how vector databases function:

### Vector Representation

Vector databases store data as vector embeddings - numerical representations of data objects like text, images, or audio [1](#). These embeddings are created using machine learning models that capture semantic relationships between data points [5](#).

### Indexing

Vector databases use specialized indexing algorithms to organize the vector data for efficient retrieval:

- Algorithms like HNSW (Hierarchical Navigable Small World), LSH (Locality-Sensitive Hashing), or PQ (Product Quantization) are used to index vectors [3](#).
- Indexing maps vectors to data structures that enable fast similarity searches [5](#).

### Querying

When a query is made:

- The query is converted into a vector embedding using the same model used for indexing [1](#).
- The database performs an Approximate Nearest Neighbor (ANN) search to find the

[Ask follow-up](#)

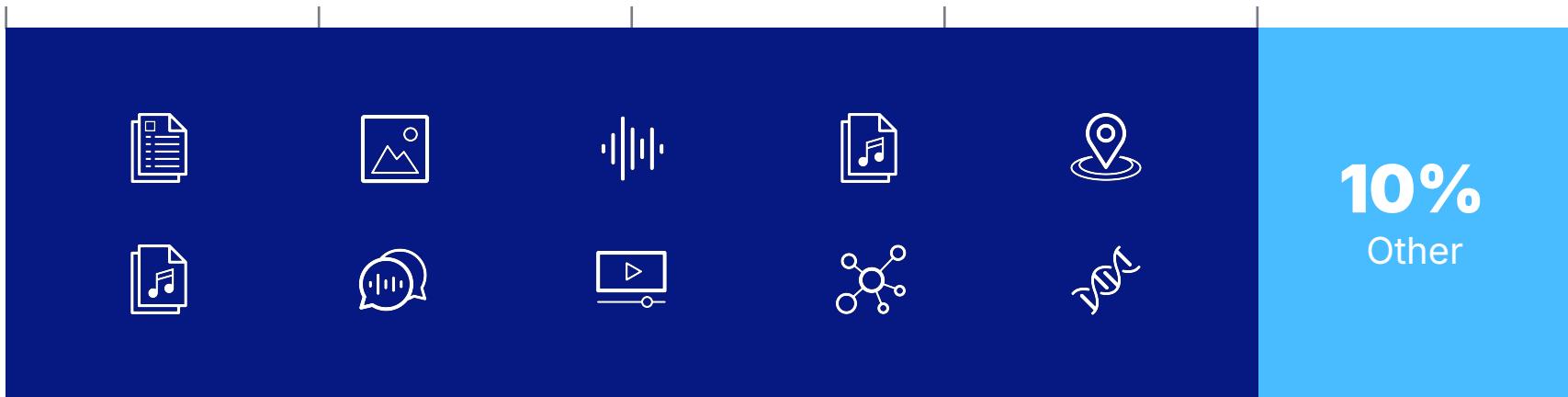
Pro





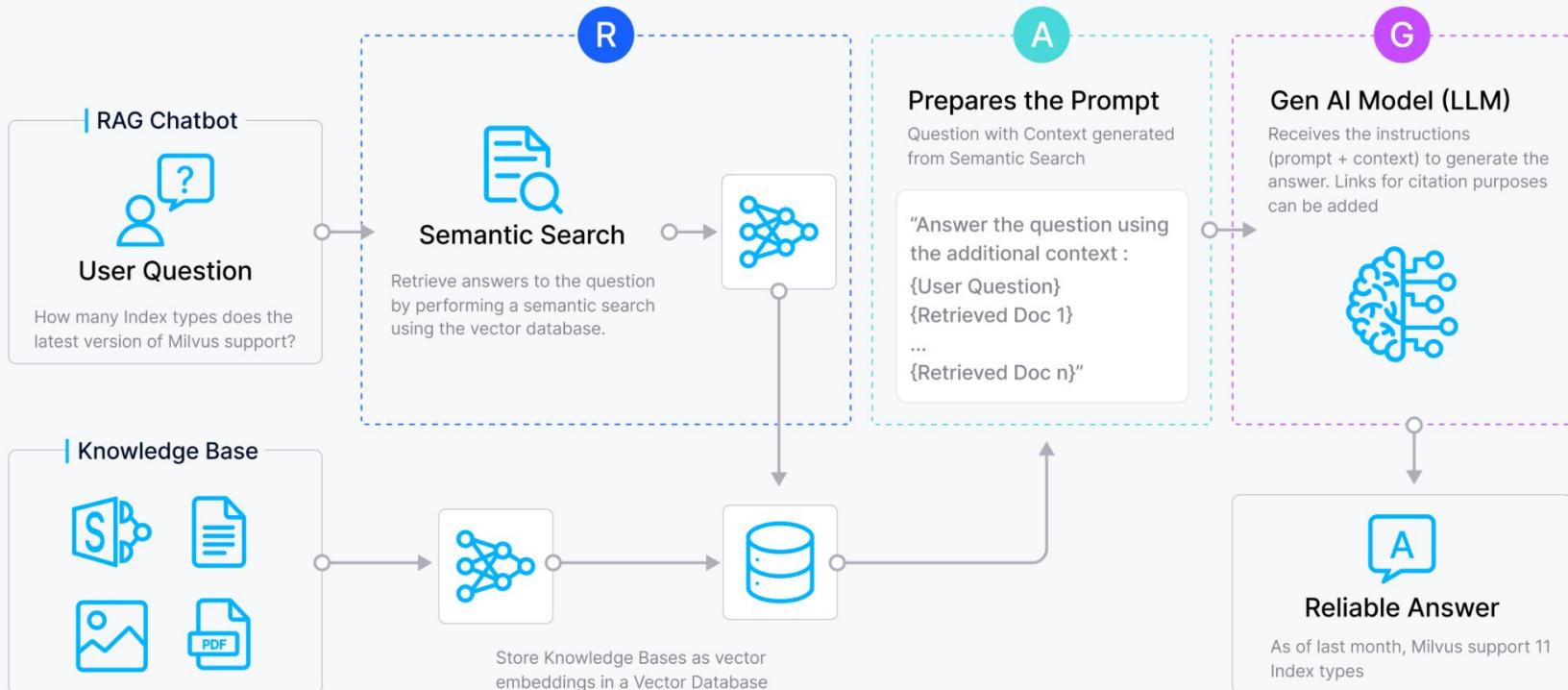
# Why is Semantic Search Important?

**90%** newly generated data in 2025  
will be unstructured data



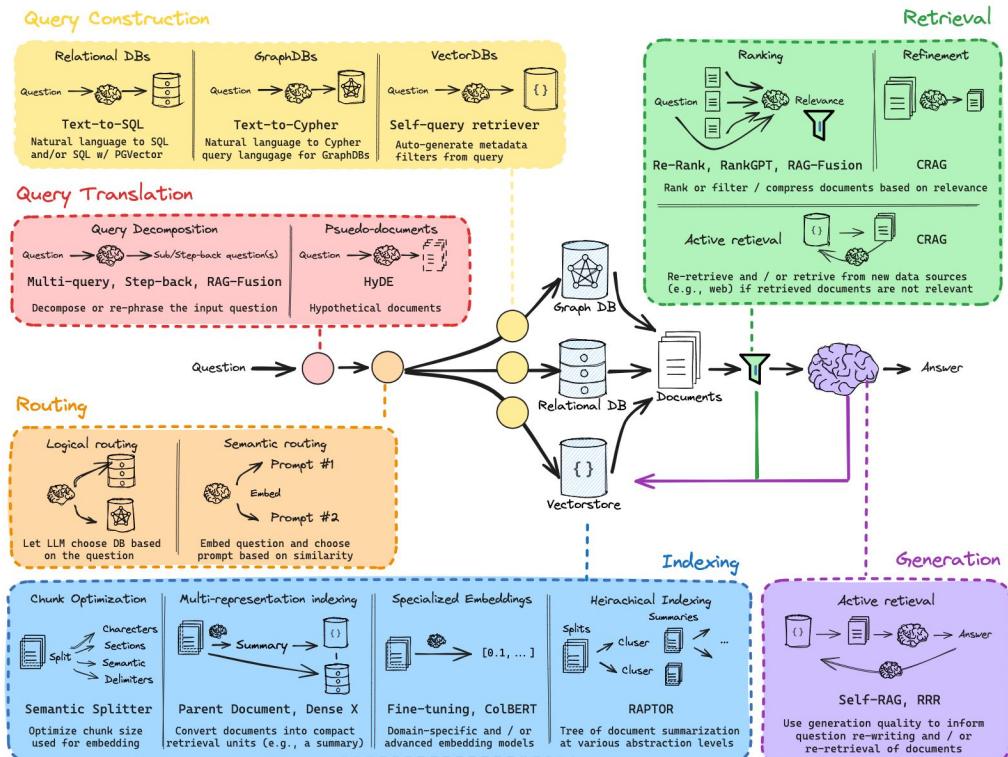


# RAG Refresher





# RAG Pipeline Options



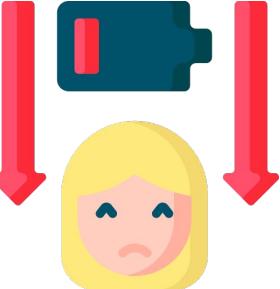
[<https://github.com/langchain-ai/rag-from-scratch>]



# RAG Pipeline Options

TABLE I  
SUMMARY OF RAG METHODS

Method	Retrieval Source	Retrieval Data Type	Retrieval Granularity	Augmentation Stage	Retrieval process
CoG [29]	Wikipedia	Text	Phrase	Pre-training	Iterative
FactoX [30]	FactoWiki	Text	Phrase	Inference	Once
EAT [31]	Dataset-base	Text	Sentence	Tuning	Once
UPRISE [20]	Dataset-base	Text	Sentence	Tuning	Once
RAST [32]	Dataset-base	Text	Sentence	Tuning	Once
Self-RATE [7]	Dataset-base	Text	Sentence	Tuning	Once
RA-TRE [24]	Search Engine,Wikipedia	Text	Sentence	Tuning	Adaptive
PORa [33]	Wikipedia	Text	Sentence	Inference	Once
FILCO [34]	Dataset-base	Text	Sentence	Inference	Once
RADA [35]	Dataset-base	Text	Sentence	Inference	Once
FinetuneR [36]	Synthesized dataset	Text	Sentence	Inference	Once
R-QQA [37]	Dataset-base	Text	Sentence Pair	Tuning	Once
LLM-R [38]	Dataset-base	Text	Sentence Pair	Inference	Iterative
LM-RETRIEVE [39]	Dataset-base	Text	Item	Pre-training	Once
LM-Inferer [40]	Dataset-base	Text	Item-base	Tuning	Once
BEQE [9]	Dataset-base	Text	Item-base	Tuning	Once
CT-RAG [41]	Synthesized dataset	Text	Item-base	Tuning	Iterative
RA-CHUNK [42]	Wikipedia, Common Crawl	Text	Chunk	Pre-training	Once
RAVEN [43]	Wikipedia	Text	Chunk	Pre-training	Once
RETRIO++ [44]	Pre-training Corpus	Text	Chunk	Pre-training	Iterative
INSTRUC-RETRO [45]	Pre-training corpus	Text	Chunk	Pre-training	Iterative
RRF [7]	Search Engine	Text	Chunk	Tuning	Once
RA-e2e [46]	Dataset-base	Text	Chunk	Tuning	Once
BEIR [47]	BEIR	Text	Chunk	Tuning	Once
PROMPTAGATOR [21]	MSMARCO,Wikipedia	Text	Chunk	Tuning	Once
RA-REF [48]	Common Crawl,Wikipedia	Text	Chunk	Tuning	Once
RA-DFT [27]	Wikipedia	Text	Chunk	Tuning	Once
RAG-Robust [48]	Dataset-base	Text	Chunk	Tuning	Once
RA-Form [49]	CoCo	Text	Chunk	Tuning	Once
CoCo [50]	Wikipedia	Text	Chunk	Tuning	Once
Self-RAG [25]	Wikipedia	Text	Chunk	Inference	Adaptive
BGM [26]	Wikipedia	Text	Chunk	Inference	Once
CoQ [51]	Wikipedia	Text	Chunk	Inference	Iterative
Tokenization [52]	Wikipedia	Text	Chunk	Inference	Once
PaperQA [53]	Arxiv,Online Database,PubMed	Text	Chunk	Inference	Iterative
NoiseRAG [54]	FactoWiki	Text	Chunk	Inference	Once
IAE [55]	Dataset-base	Text	Chunk	Inference	Once
NeuroACL [56]	Search Engine,Wikipedia	Text	Chunk	Inference	Once
Tac [57]	Search Engine,Wikipedia	Text	Chunk	Inference	Recursive
SKR [58]	Dataset-base,Wikipedia	Text	Chunk	Inference	Adaptive
RA-LongContext [59]	Wikipedia	Text	Chunk	Inference	Recursive
RAG-LongContext [60]	Dataset-base	Text	Chunk	Inference	Once
ITER-RETRIEVE [14]	Wikipedia	Text	Chunk	Inference	Iterative
IRC4J [61]	Wikipedia	Text	Chunk	Inference	Recursive
LLM-KnowledgeTransfer [62]	Wikipedia	Text	Chunk	Inference	Once
RAPTOR [63]	Dataset-base	Text	Chunk	Inference	Recursive
RECITE [22]	LLMs	Text	Chunk	Inference	Once
ICRALM [64]	Pile,Wikipedia	Text	Chunk	Inference	Iterative
Retrieval-Augmented [65]	Dataset-base	Doc	Doc	Tuning	Once
Zenit [66]	C4	Text	Doc	Tuning	Once
CRAG [67]	Arxiv	Text	Doc	Inference	Once
1-Step [68]	Wikipedia	Text	Doc	Inference	Once
PRCA [69]	Dataset-base	Text	Doc	Inference	Once
QLM-Doc-ranking [70]	Dataset-base	Text	Doc	Inference	Once
Dump [71]	Wikipedia	Text	Doc	Inference	Once
ReptiLIG [72]	Wikipedia	Text	Doc	Inference	Once
ARM-RAG [73]	Dataset-base	Text	Doc	Inference	Once
CoCo [51]	LLMs	Text	Doc	Inference	Iterative
Unsup-RAG [74]	Dataset-base	Text	Doc	Inference	Iterative
CREA-ICL [19]	Dataset-base	Crosslingual,Text	Sentence	Inference	Once
PKG [75]	LLM	Tabular,Text	Chunk	Inference	Once
SANTA [76]	Dataset-base	Code,Text	Item	Pre-training	Once
SLURP [77]	Dataset-base	Text	Sub-Graph	Tuning	Once
MK-ToD [78]	Dataset-base	KG	Entity	Tuning	Once
Dual-Feedback-ToD [79]	Dataset-base	KG	Entity Sequence	Tuning	Once
Knowledge-ToD [15]	Dataset-base	KG	Triplet	Inference	Multi-stage
PARLIB [80]	Dataset-base,Graph	KG	Entity	Inference	Once
HyKE [81]	CMKG	KG	Triplet	Inference	Once
KALMV [82]	Wikipedia	KG	Triplet	Inference	Iterative
KG-RC [83]	Freebase	KG	Triplet	Inference	Once
G-Retriever [84]	Dataset-base	Text,Graph	Sub-Graph	Inference	Once

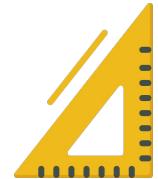




# How to Make RAG Design Choices?

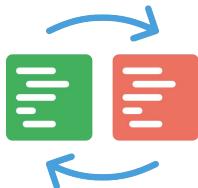


① Consider alternatives



② Measure effects

How??



③ Update and repeat



# Scope of Webinar



- Fundamentals of LLM-as-a-Judge
- Evaluating output:
  - LLMs
  - RAG
  - Offline
- Frameworks:
  - Milvus
  - RAGAS



- Deeper discussion of limitations of LLM-as-a-Judge
- Further / alternative methods
- Evaluating output:
  - Agents
  - Multimodal models
  - Online
- Evaluating latency
- Adversarial attacks

02

## Evaluating Foundation Models



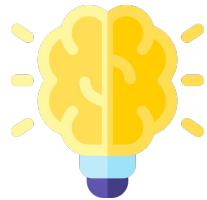


# What is “Performance”?

- Distinction between evaluation on a task vs evaluation on itself
- Comparing answer to ground-truth vs. comparing output to context
- Evaluating retrieval vs. evaluating LLM output
- With ground-truth and without ground-truth
- Human evaluation is “gold standard” but doesn’t scale well
- **Surprisingly, (strong) LLMs can evaluate LLMs when there is no ground truth**



# Task-Based Evaluation



## Knowledge-based

- MMLU
- [HellaSwag](#)
- ARC



## Conversational

- [CoQA](#)
- MMDialog
- OpenAssistant

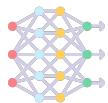


## Instruct following

- Flan
- [Self-instruct](#)
- NaturalInstructions



# "Introspection"-based Evaluation



## Generation-based

- *Faithfulness / groundedness*
  - “factual consistency of the generated answer against the given context.”
  - “claims in the answer that cannot be deduced from context [are] penalized.”
- *Answer Relevancy*
  - “response directly addresses and is appropriate for a given question or context.”

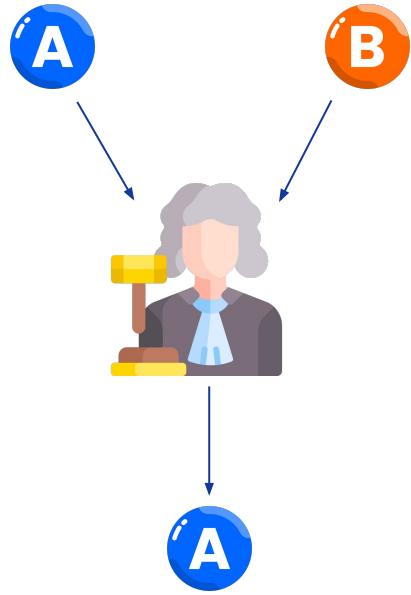


## Retrieval-based

- *Context Relevance*
  - “how relevant are retrieved contexts to the question.”
  - “context should only contain information necessary to answer the question.”
- *Context Recall / retrieval-based metrics*
  - recall of the retrieved context using the annotated answer as ground truth.



# LLM-as-a-Judge: Pairwise Comparison



[System]

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

[User Question]

{question}

[The Start of Assistant A's Answer]

{answer\_a}

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

{answer\_b}

[The End of Assistant B's Answer]

[\[Zheng et al., 2023; Fig 5\]](#)



# LLM-as-a-Judge: Single Answer Grading



[System]  
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of the response. Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please rate the response on a scale of 1 to 10 by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]  
{question}

[The Start of Assistant's Answer]  
{answer}  
[The End of Assistant's Answer]

[[Zheng et al., 2023](#); Fig 6]



# LLM-as-a-Judge: Multi-turn Pairwise Comparison

```
[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two
AI assistants to the user question displayed below. You should choose the assistant that
follows the user's instructions and answers the user's question better. Your evaluation
should consider factors such as the helpfulness, relevance, accuracy, depth, creativity,
and level of detail of their responses. Begin your evaluation by comparing the two
responses and provide a short explanation. Avoid any position biases and ensure that the
order in which the responses were presented does not influence your decision. Do not allow
the length of the responses to influence your evaluation. Do not favor certain names of
the assistants. Be as objective as possible. After providing your explanation, output your
final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]"
if assistant B is better, and "[[C]]" for a tie.

<|The Start of Assistant A's Conversation with User>
### User:
{question 1}

### Assistant A:
{answer 1}

### User:
{question 2}

### Assistant A:
{answer 2}
<|The End of Assistant A's Conversation with User>

<|The Start of Assistant B's Conversation with User>
### User:
{question 1}

### Assistant B:
{answer 1}

### User:
{question 2}

### Assistant B:
{answer 2}
<|The End of Assistant B's Conversation with User>
```

[[Zheng et al., 2023](#); Fig 9]

# 03

## Challenges and Limitations



# Position Bias



**Question:** What are some business etiquette norms when doing business in Japan?

## Assistant A:

Here are some business etiquette norms to keep in mind when doing business in Japan:

1. Bowing: Bowing is a common greeting in Japan. It is a sign of respect and should be reciprocated.
2. Business cards: Business cards are exchanged at the beginning of a meeting. It is important to receive and give business cards with both hands and to take a moment to read the card before putting it away.
3. Punctuality: Being on time is important in Japan. Arrive a few minutes early to show respect for your colleagues.
4. Dress code: Dress conservatively and professionally. Men should wear dark suits and ties, while women should wear conservative business attire.
5. Language: Learn a few basic Japanese phrases to show respect for the culture. It is also important to speak clearly and avoid using slang or idioms.
6. Gift giving: Gift giving is a common practice in Japan. Bring a small gift, such as a box of chocolates or a small souvenir from your home country.
7. Dining etiquette: When dining with colleagues, wait for the host to begin eating before you start. Use chopsticks correctly and avoid leaving food on your plate.

Remember, showing respect for Japanese culture and customs is key to building strong business relationships in Japan.

## GPT-4 Judgment (when A is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant A's answer is more detailed and organized, covering a wider range of topics such as dress code and language, which are not mentioned in Assistant B's response. Assistant A also provides more specific advice on how to handle certain situations, such as exchanging business cards and dining etiquette. Therefore, Assistant A's answer is more comprehensive and useful for someone looking to understand Japanese business etiquette.

## GPT-4 Judgment (when B is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant B's answer is more detailed and covers a wider range of topics, such as addressing people by their titles, avoiding direct eye contact, and being mindful of communication styles. Assistant A's answer is more concise but lacks some of the depth found in Assistant B's response. Therefore, Assistant B's answer is better in this case.

## Assistant B:

When doing business in Japan, it is important to be aware of and respect the country's business etiquette norms. Some of these norms include:

1. Greetings: A handshake and bow is the standard greeting in Japan. It is important to bow deeply and to offer a firm handshake.
2. Business cards: In Japan, business cards are exchanged after the initial greeting. It is important to handle the card with respect and to read it carefully before placing it on the table.
3. Meetings: Meetings in Japan are typically formal and conducted in a hierarchical manner. It is important to address people by their titles and to avoid direct eye contact.
4. Punctuality: Punctuality is highly valued in Japan. It is important to arrive on time for meetings and to be prepared in advance.
5. Gift giving: Gift giving is a common practice in Japan, but it is important to be mindful of the occasion and the relationship. Avoid giving gifts that are considered too personal or extravagant.
6. Dining: Dining is an important part of doing business in Japan. It is important to be aware of the proper table manners and to avoid slurping or making loud noises while eating.
7. Communication: Communication in Japan is typically indirect and reserved. It is important to listen carefully and to avoid confrontational language.

Overall, it is important to be respectful, punctual, and mindful of hierarchy when doing business in Japan.

# Verbosity Bias



**Question:** What is the central dogma of molecular biology? What processes are involved? Who named this?

**Assistant A:**

The central dogma of molecular biology is the fundamental principle that explains how genetic information flows within a biological system. It states that DNA is transcribed into RNA, which is then translated into proteins. This means that the genetic information stored in DNA is first copied into RNA, which then serves as a template for the synthesis of proteins.

The processes involved in the central dogma are:

1. Transcription rephrased: The act of copying genetic information from DNA to RNA.
2. Translation rephrased: The act of using genetic information in RNA to create proteins.
3. Transcription: The process by which the genetic information in DNA is copied into RNA.
4. Translation: The process by which the genetic information in RNA is used to synthesize proteins.

The term "central dogma" was coined by Francis Crick, one of the co-discoverers of the structure of DNA, in 1958.

**GPT-3.5 Judgment:**

Assistant A provides a more detailed and accurate response to the user's question. While both assistants correctly state the central dogma of molecular biology and who coined the term, Assistant A provides a more comprehensive explanation of the processes involved in the central dogma, including the rephrased definitions of transcription and translation. Therefore, I choose A as the better assistant.

**Claude-v1 Judgment:**

Both assistants answer the question by defining the central dogma and detailing the two processes involved, transcription and translation. However, assistant A provides a more in-depth and detailed explanation by briefly rephrasing and elaborating on these processes. Assistant A also mentions who coined the term "central dogma", which provides useful context and shows a more comprehensive understanding of the topic. Overall, assistant A gives a fuller and more insightful response, so I have chosen it as the better assistant for this question.

**GPT-4 Judgment:**

Both Assistant A and Assistant B provide accurate and relevant information about the central dogma of molecular biology, the processes involved, and who named it. However, Assistant A's answer contains a minor redundancy by rephrasing transcription and translation twice. Assistant B's answer is more concise and avoids this redundancy. Therefore, Assistant B's answer is better.

**Assistant B:**

The central dogma of molecular biology is the fundamental principle that explains how genetic information flows within a biological system. It states that DNA is transcribed into RNA, which is then translated into proteins. This means that the genetic information stored in DNA is first copied into RNA, which then serves as a template for the synthesis of proteins.

The processes involved in the central dogma are:

1. Transcription: The process by which the genetic information in DNA is copied into RNA.
2. Translation: The process by which the genetic information in RNA is used to synthesize proteins.

The term "central dogma" was coined by Francis Crick, one of the co-discoverers of the structure of DNA, in 1958.

# Can Answer But Can't Judge



**Question:** Each problem consists of three statements. Based on the first two statements, the third statement may be true, false, or uncertain.

1. Oranges cost more than apples.
2. Oranges cost less than bananas.
3. Bananas cost more than apples and bananas cost more than orange.

If the first two statements are true, then the third statement is

## Assistant A:

false.

If oranges cost more than apples (statement 1) and oranges cost less than bananas (statement 2), then it is not possible for bananas to cost less than oranges and also cost more than both apples and oranges (statement 3). Therefore, statement 3 is false.

## Assistant B:

False.

If the first two statements are true, then the third statement is false. The first statement says that oranges cost more than apples, and the second statement says that oranges cost less than bananas. This means that oranges cost less than bananas, which contradicts the third statement that bananas cost more than apples and more than oranges. Therefore, the third statement must be false.

## GPT-4 Judgment:

Both Assistant A and Assistant B provided accurate and relevant responses to the user's question. They both concluded that the third statement is false based on the first two statements. However, Assistant B's response is slightly clearer and more concise in explaining the contradiction between the statements. Therefore, I choose Assistant B as the better response.

[[B]]

## GPT-4's Answer (if asked separately):

True.

If oranges cost more than apples (statement 1) and oranges cost less than bananas (statement 2), it logically follows that bananas cost more than apples and bananas cost more than oranges, which is the third statement. So, if the first two statements are true, the third one is also true.



# C-o-T Failure Mode

**Question:** How many integers are in the solution of the inequality  $|x + 5| < 10$ ?

**Assistant A:**

There are 19 integers in the solution of the inequality  $|x + 5| < 10$ .

**Assistant B:**

To solve the inequality  $|x + 5| < 10$ , we need to consider two cases:

Case 1:  $x + 5$  is positive or zero

In this case, the inequality becomes  $x + 5 < 10$ , which simplifies to  $x < 5$ .

Case 2:  $x + 5$  is negative

In this case, the inequality becomes  $-(x + 5) < 10$ , which simplifies to  $-x - 5 < 10$ . Adding 5 to both sides gives  $-x < 15$ , and multiplying by -1 (and flipping the inequality) gives  $x > -15$ .

So the solution to the inequality is  $-15 < x < 5$ . The integers in this range are **-14, -13, -12, ..., 4**, for a total of 20 integers. Therefore, there are 20 integers in the solution of the inequality.

**GPT-4 Judgment:**

To solve the inequality  $|x + 5| < 10$ , we need to consider two cases:

Case 1:  $x + 5$  is positive or zero

In this case, the inequality becomes  $x + 5 < 10$ , which simplifies to  $x < 5$ .

Case 2:  $x + 5$  is negative

In this case, the inequality becomes  $-(x + 5) < 10$ , which simplifies to  $-x - 5 < 10$ . Adding 5 to both sides gives  $-x < 15$ , and multiplying by -1 (and flipping the inequality) gives  $x > -15$ .

So the solution to the inequality is  $-15 < x < 5$ . The integers in this range are **-14, -13, -12, ..., 4**, for a total of 20 integers. Therefore, there are 20 integers in the solution of the inequality.

Comparing the answers provided by the two AI assistants:

Assistant A: 19 integers

Assistant B: 20 integers

The correct answer is 20 integers, so Assistant B provided the correct and more helpful response.

Final verdict: [[B]]

**GPT-4's Answer (if asked separately):**

We can split the inequality into two cases:

1) If  $x+5 \geq 0$ , then  $|x+5|=x+5$ . So, we have  $x+5<10$ , which means  $x<5$ . Since  $x+5 \geq 0$ , we have  $x \geq -5$ . So in this case,  $x$  can be any integer from  $-5$  to  $4$ , inclusive, which gives us  $5 - (-5) + 1 = 10$  integers.

2) If  $x+5 < 0$ , then  $|x+5|=-(x+5)$ . So, we have  $-(x+5)<10$ , which means  $x+5>-10$ , which means  $x>-15$ . Since  $x+5<0$ , we have  $x<-5$ . So in this case,  $x$  can be any integer from  $-14$  to  $-6$ , inclusive, which gives us  $-6 - (-14) + 1 = 9$  integers.

In total, there are  $10 + 9 = \boxed{19}$  integers in the solution.

Figure 15: An example of GPT-4's wrong judgment with chain-of-thought prompt. We can see GPT-4 exactly copied Assistant B's answer (which contains arithmetic errors) and determined Assistant A's answer is incorrect. This suggest GPT-4's chain-of-thought process can be significantly influenced by the given answers despite we ask it to think independently.



# Data Quality

***Like in other parts of Gen AI, the main challenge is producing the right dataset!***



# Fine-tuned Judge Models

- [huggingface.co/nuclia/REMi-v0](https://huggingface.co/nuclia/REMi-v0)
- [huggingface.co/grounded-ai](https://huggingface.co/grounded-ai)
- [huggingface.co/prometheus-eval](https://huggingface.co/prometheus-eval)
- [huggingface.co/flowaicom/Flow-Judge-v0.1](https://huggingface.co/flowaicom/Flow-Judge-v0.1)
- [huggingface.co/facebook/Self-taught-evaluator-llama3.1-70B](https://huggingface.co/facebook/Self-taught-evaluator-llama3.1-70B)

# 04

## Let's Build with Open-Source!





**32K+**  
GitHub Stars



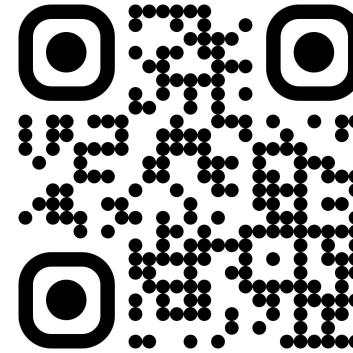
**3K+**  
Forks



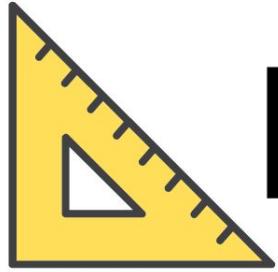
**66M+**  
Docker Pulls



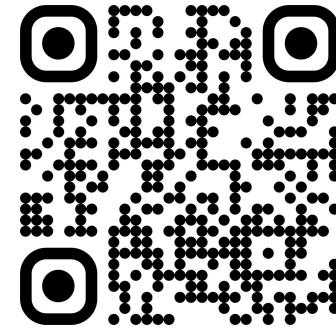
**400+**  
Contributors



[milvus.io/docs](https://milvus.io/docs) | [github.com/milvus-io/milvus](https://github.com/milvus-io/milvus)

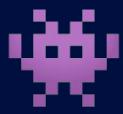


# ragas



[docs.ragas.io/en/stable/](https://docs.ragas.io/en/stable/) | [github.com/explodinggradients/ragas](https://github.com/explodinggradients/ragas)

# It's Coding Time!





# Summary



*There are many methods, models, etc.*

*We need quantitative eval for principled design*



*Tasks can evaluate LLMs when there's a ground truth*

*LLMs can evaluate LLMs when there's no ground truth (also, when there is a ground truth)*



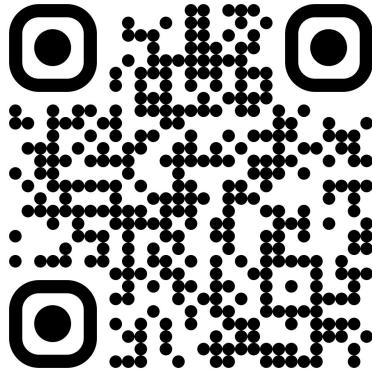
*There are several excellent open-source evaluation frameworks although it is a relative nascent field*

*Research into RAG eval methodology continues...*

# LET'S STAY CONNECTED!

**Stefan Webb**

Developer Advocate, Zilliz



<https://milvus.io/discord>



<https://github.com/milvus-io/milvus>



<https://x.com/milvusio>



<https://www.linkedin.com/company/the-milvus-project>



Book a free 1:1 session to get help with your production deployment  
[meetings.hubspot.com/chloe-williams1/milvus-office-hours](https://meetings.hubspot.com/chloe-williams1/milvus-office-hours)



Join us at our next meetup!  
[lu.ma/unstructured-data-meetup](https://lu.ma/unstructured-data-meetup)



Feb 27, San Francisco, AWS GenAI Loft  
Zilliz, Amazon AWS, [TBC]

THANK YOU

