Zeta Alpha

OpenSource Connections

FROM LLM-AS-A-JUDGE TO HUMAN-IN-THE-LOOP

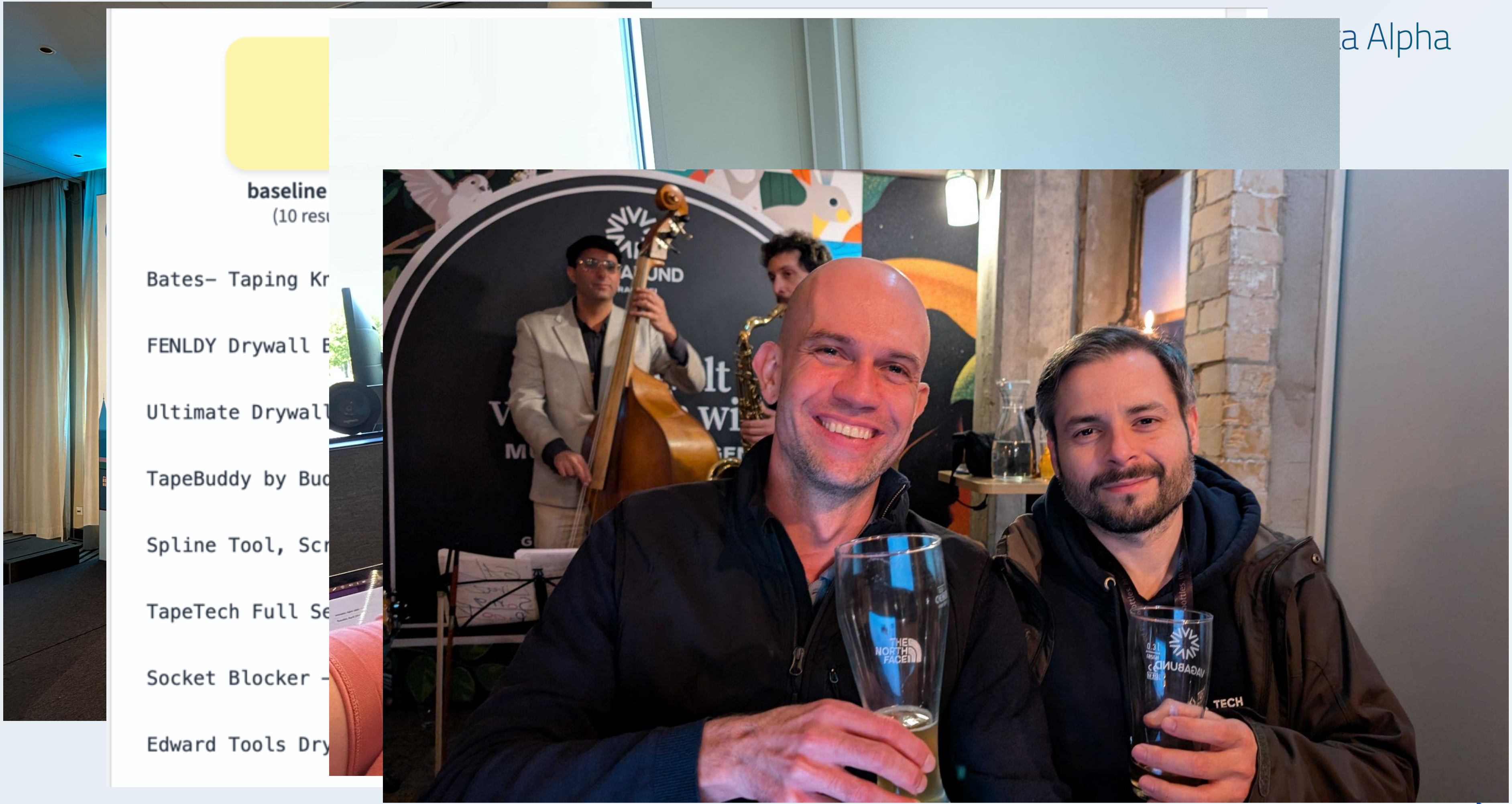
Fernando Rejon
CTO, Zeta Alpha

Daniel Wrigley
Senior Search Consultant,
OpenSource Connections

September 24th 2025 | Haystack | Berlin

FROM LLM-AS-A-JUDGE TO HUMAN-IN-THE-LOOP

LET'S SET THE CONTEXT FOR THIS TALK...

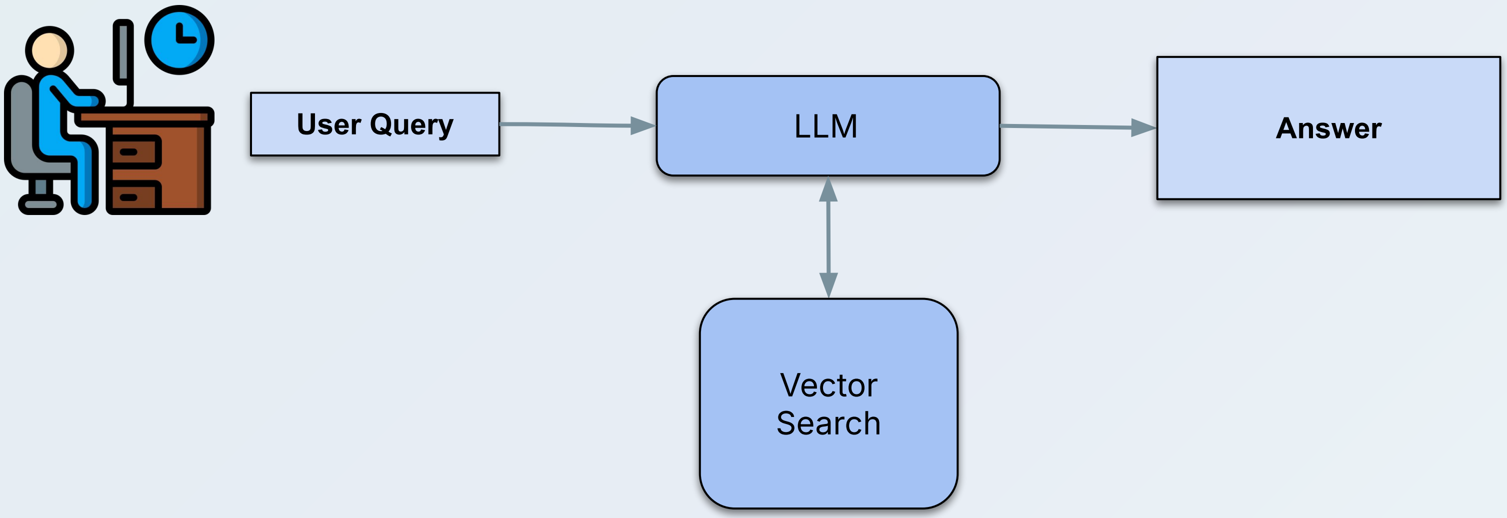

baseline
(10 resu
Bates– Taping Kr
FENLDY Drywall E
Ultimate Drywall
TapeBuddy by Bud
Spline Tool, Scr
TapeTech Full Se
Socket Blocker -
Edward Tools Dry
THE NORTH FACE
VAGABUND
TECH

HOW CAN WE INTRODUCE EYEBALLING INTO RAG EVALUATION?

Zeta Alpha

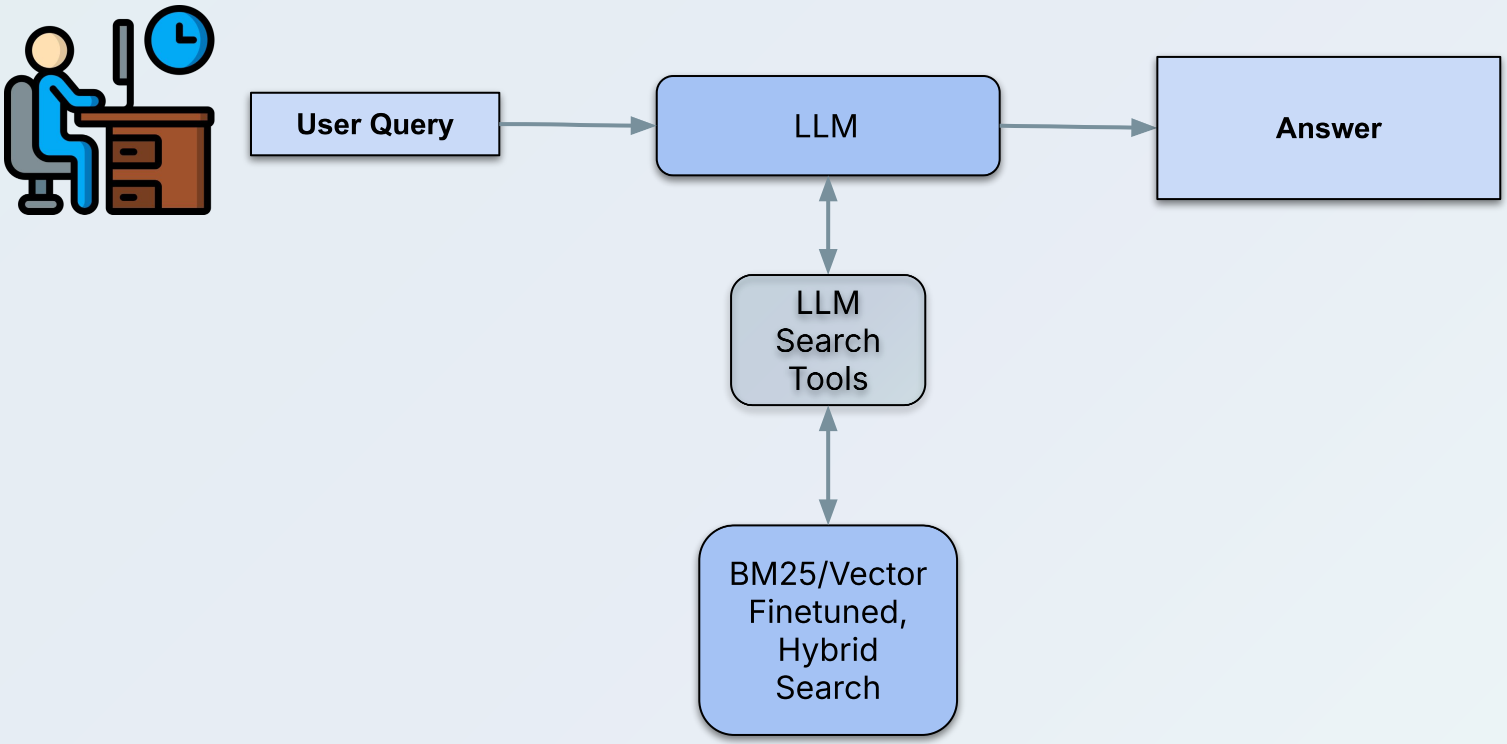
SIMPLE RAG FROM 2022

Zeta Alpha



MODERN RAG

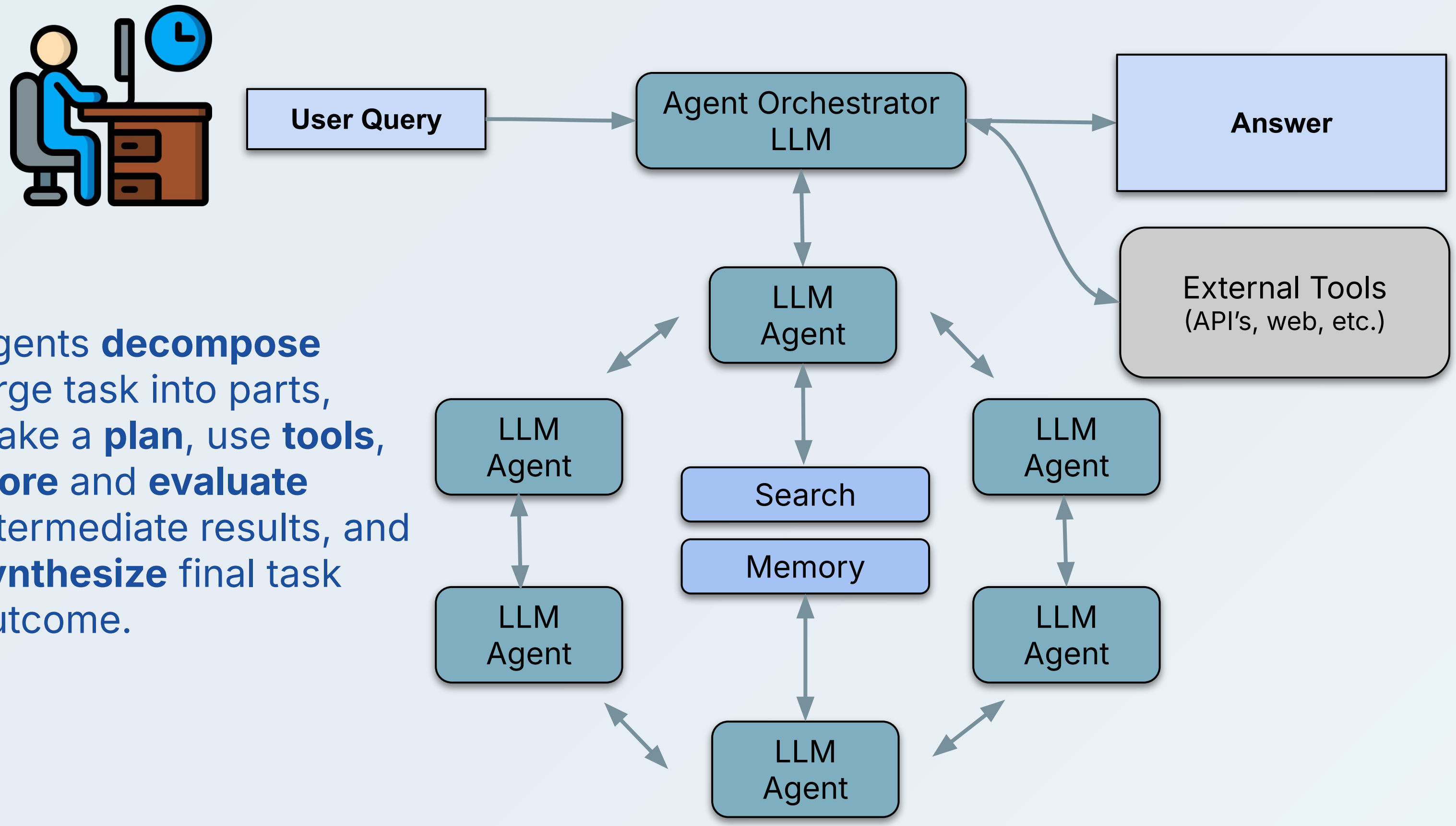
Zeta Alpha



FROM RAG TO AI AGENTS

Zeta Alpha

Agents **decompose** large task into parts, make a **plan**, use **tools**, **store** and **evaluate** intermediate results, and **synthesize** final task outcome.



EVALUATION...

Zeta Alpha

With Gen AI, **prototyping is easy, evaluation is hard**.

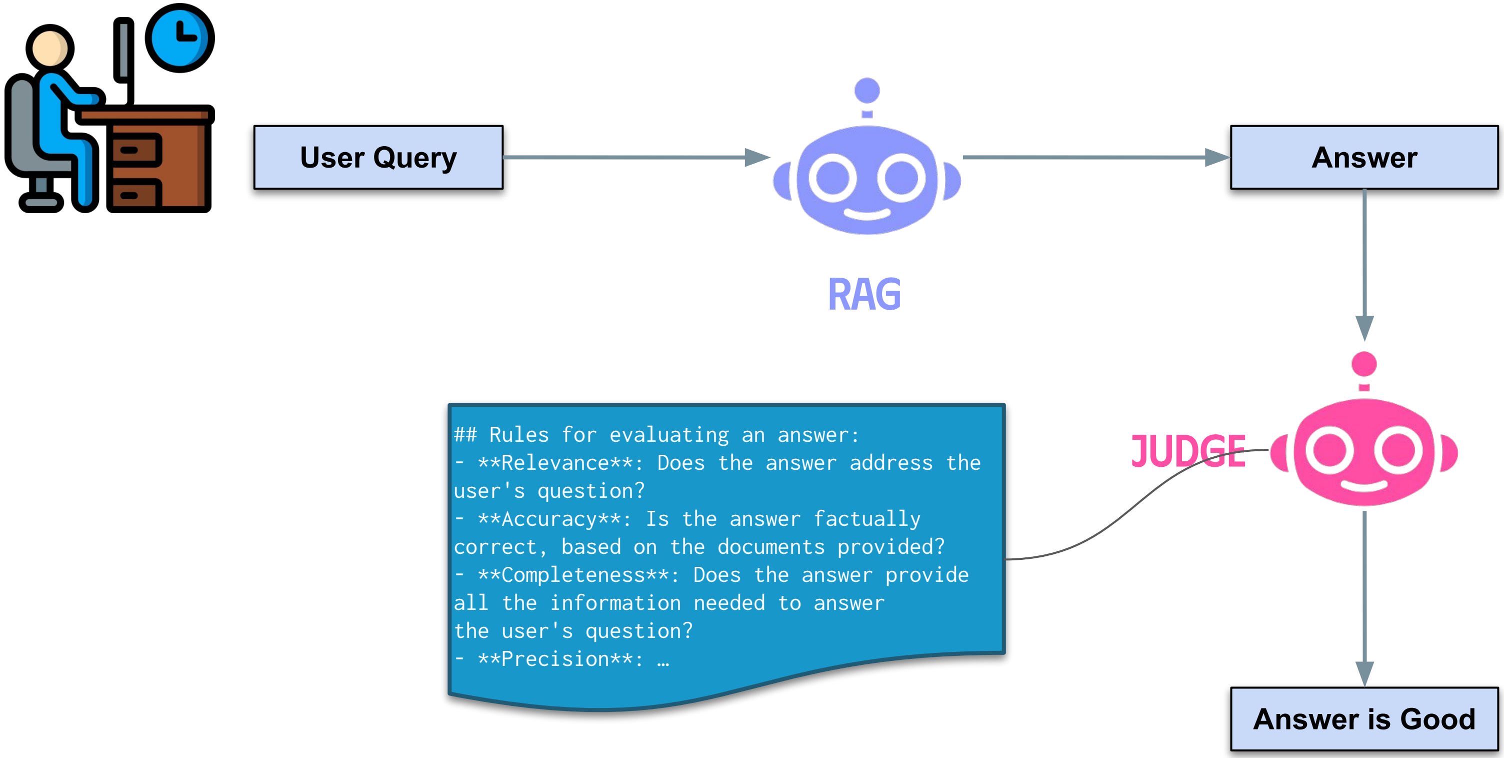
Agent Variation:

- **Retrieval pipeline** - models, processing pipelines, retrieval params
- **Agent architecture** - Multi-Agent, GraphRAG, Memory, Tools, Prompts
- **LLM** choice

EVALUATION IN THE ENTERPRISE

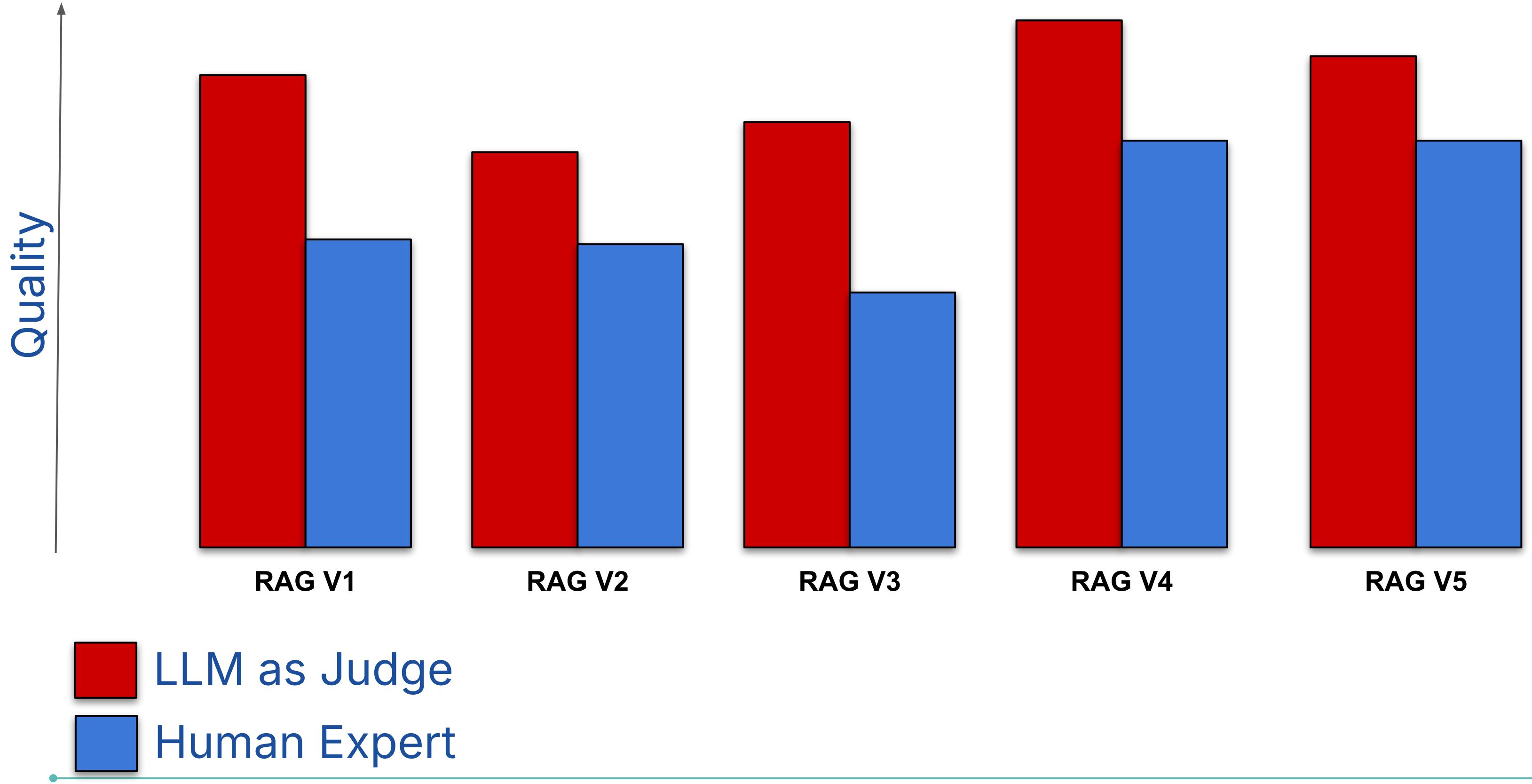
LLM AS A JUDGE: POINTWISE

Zeta Alpha



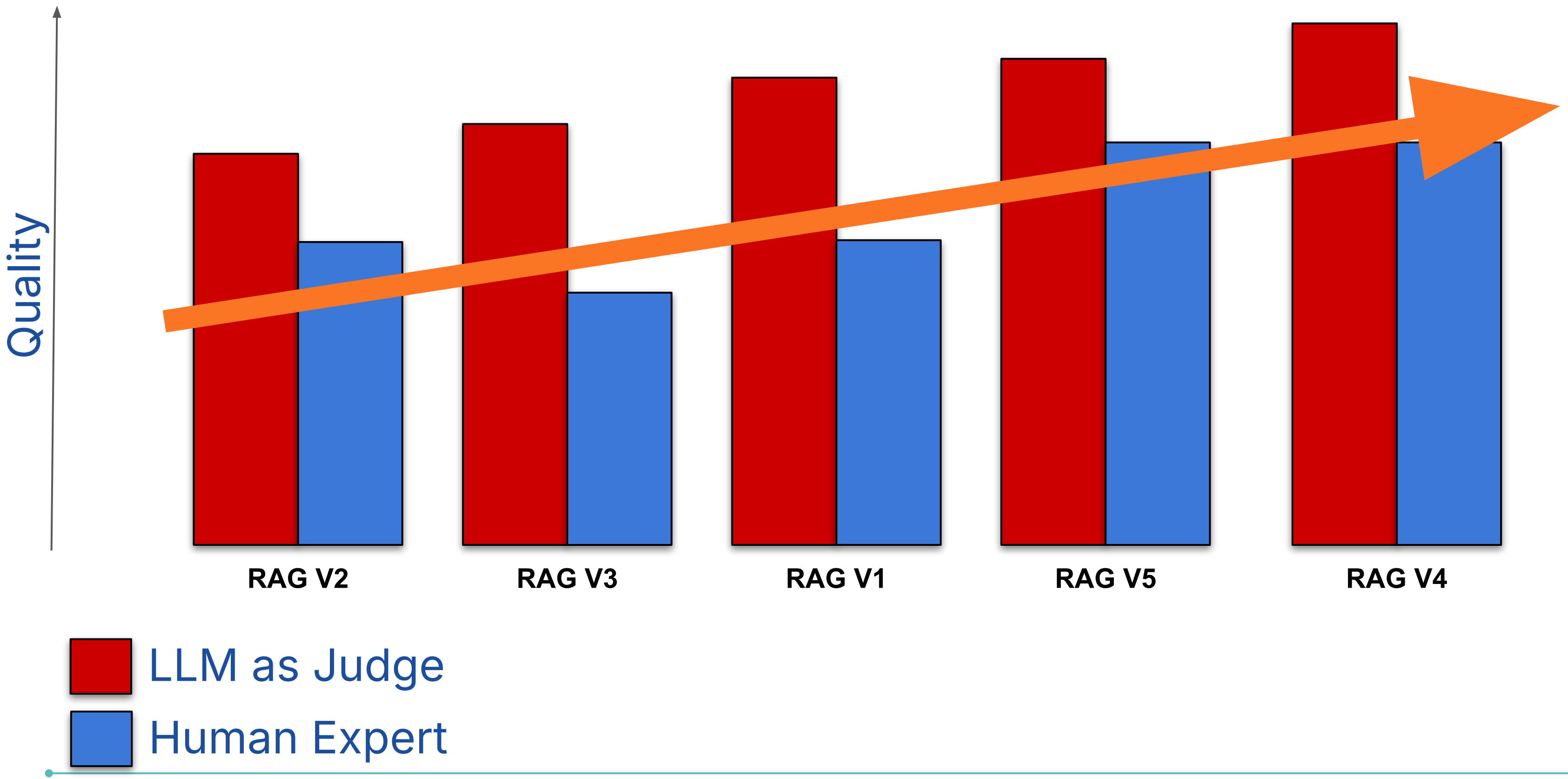
LLM AS A JUDGE VS EXPERTS

Zeta Alpha



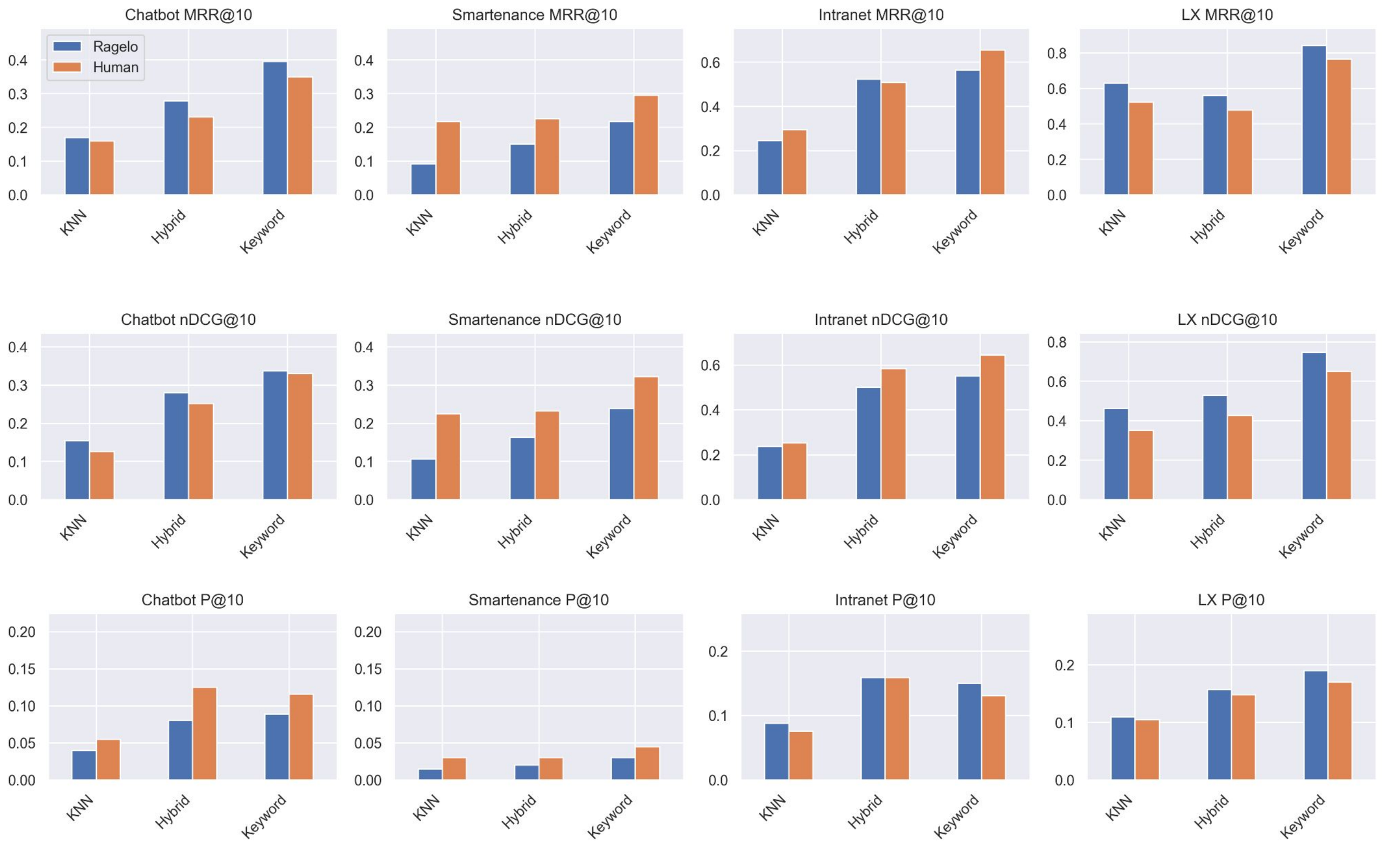
LLM AS A JUDGE VS EXPERTS

Zeta Alpha



LLM AS JUDGE VS HUMAN ANNOTATION

Zeta Alpha

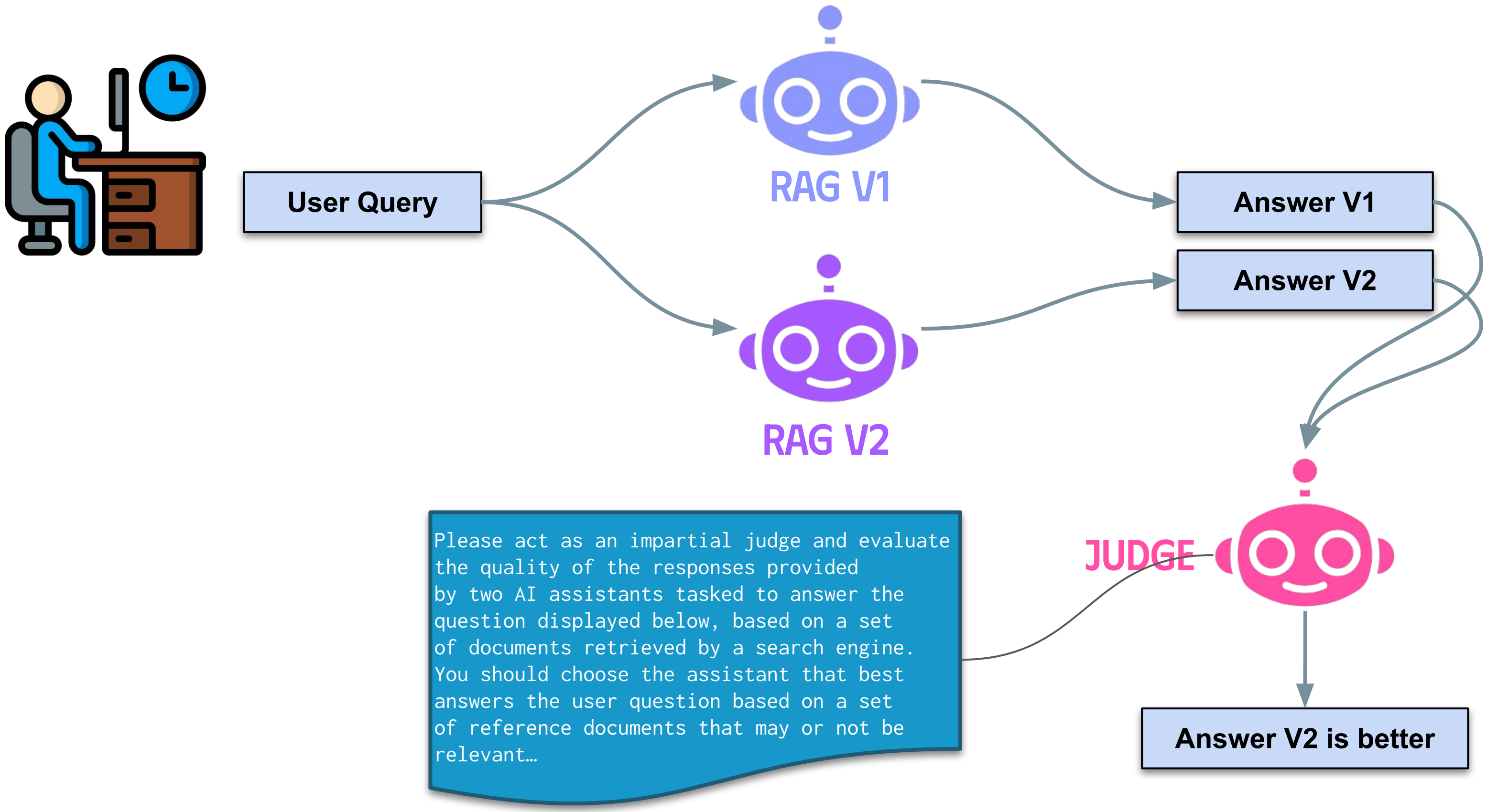


Zeta Alpha

WHAT IF WE ASK THE LLM TO COMPARE TWO ANSWERS?

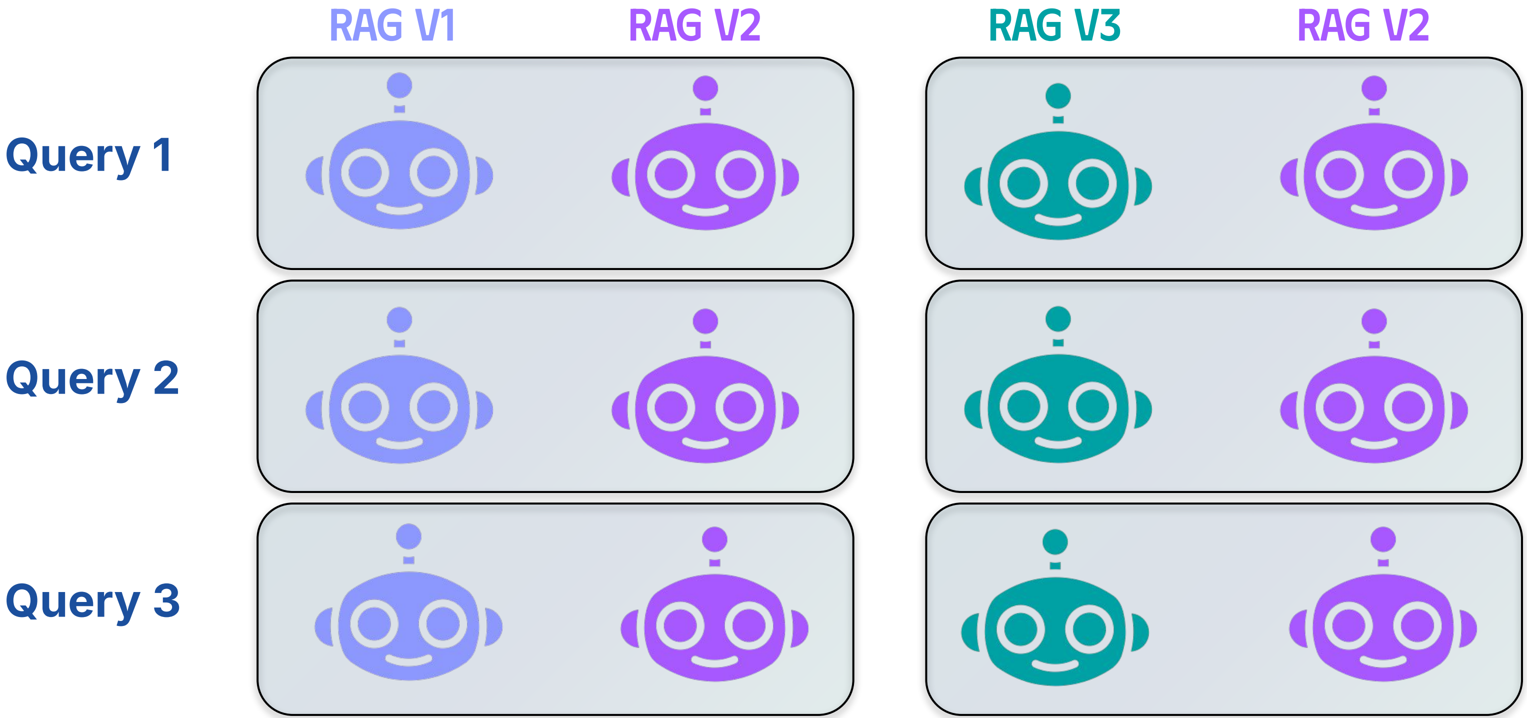
LLM AS A JUDGE: PAIRWISE

Zeta Alpha



HOW DO WE RANK PAIRWISE?

Zeta Alpha



THE ELO RANKING SYSTEM

Zeta Alpha

- Each **agent** (or RAG system) is a **player** in a tournament with an initial rank.
- Each **query** is **game** played between two **agents.**
- A **game** between **Agent A** and **Agent B** is played by prompting an LLM to select which answer to the same **query** is better.
- If **A** wins and its ranking is **higher** than **B**:
  - ↗ Score of **A** increases a bit.
  - ↘ Score of **B** decreases a bit.
- If **A** wins and its ranking is **lower** than **B**:

↖ Score of **A** increases more.

↘ Score of **B** decreases more.

ELO: EXAMPLE GAMES

Zeta Alpha

UPDATE RULE

New Rating = Old Rating + K * (Actual Score - Expected Score)

- ***Actual Score***: win=1, draw=0.5, loss=0
- ***Expected Score***: $\frac{1}{1+10^{(R_{Opp}-R_{You})/400}}$
- **Use *K* = 32** for the examples

NEW PLAYERS

Start Ratings: **A = 1500**, **B = 1500**

==A Wins==

Actual Scores: **A=1**, **B=0**
Expected Scores: **A=0.5**, **B=0.5**
Change: **ΔA = 32 * (1 - 0.5) = 16**, **ΔB = 32 * (0 - 0.5) = -16**
New Ratings: **A = 1516**, **B = 1484**

UPSET: YOU BEAT STRONGER PLAYER

Start Ratings: **A = 1500**, **B = 1700**

==A Wins==

Actual Scores: **A=1**, **B=0**
Expected Scores: **A=0.24**, **B=0.76**
Change: **ΔA = 32 * (1 - 0.24) = 24.3**, **ΔB = 32 * (0 - 0.76) = -24.3**
New Ratings: **A = 1524**, **B = 1676**

EXPECTED: YOU BEAT WEAKER PLAYER

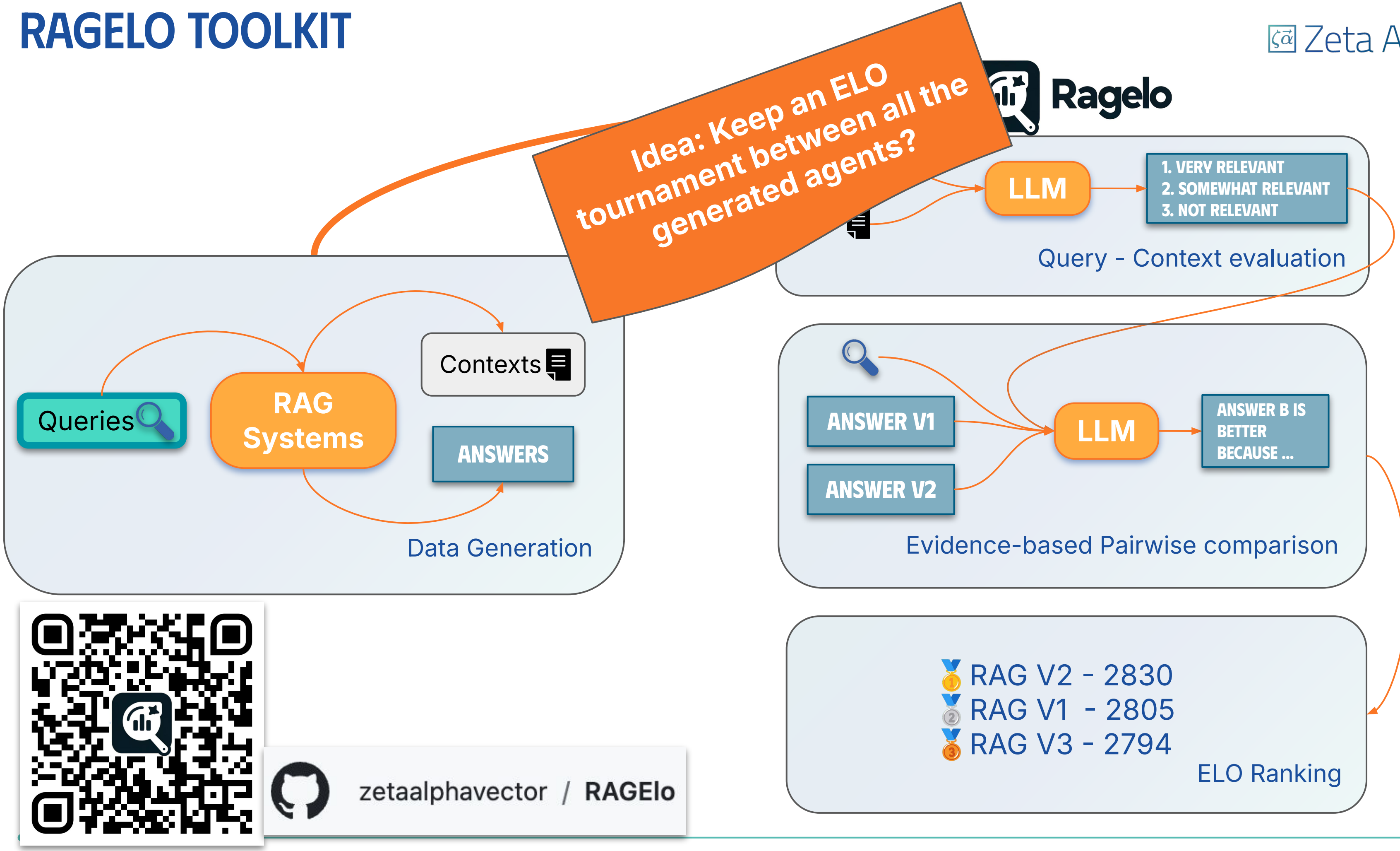
Start Ratings: **A = 1700**, **B = 1500**

==A Wins==

Actual Scores: **A=1**, **B=0**
Expected Scores: **A=0.76**, **B=0.24**
Change: **ΔA = 32 * (1 - 0.76) = 7.7**, **ΔB = 32 * (0 - 0.24) = -7.7**
New Ratings: **A = 1708**, **B = 1492**

RAGELO TOOLKIT

Zeta Alpha



RAGELO: DATA GENERATION FORMAT

Zeta Alpha

QUERIES

qid	query
1	What metrics can be used to evaluate text summarization?
2	What are all hardware accelerators used in AI?
3	Who are the main deep learning researchers in The Netherlands?

RETRIEVED DOCUMENTS

qid	did	document_text
1	d6e3da57be	And also ROUGE Precision, Recall and F-score [4]. ROUGE is a proxy metric for abstractive summarization...
1	dd2b71271	ROUGE Score: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a set of metrics
1	90314fbb41	Many popular summarization systems were evaluated with ROUGE...

ANSWERS

qid	agent	answer
1	rag_fusion_agent	In the quest to understand the metrics used to evaluate text summarization, we have explored various dimensions, including traditional metrics, novel approaches, and the potential for standardization within the field. Traditional metrics like ROUGE, BLEU, and METEOR have...
1	naive_rag	ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a commonly used package for the automatic evaluation of summaries...
1	rag_fusion	Metrics used to evaluate text summarization include ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [697133eeb1f7bcc5a3273607c2e28be1fc16c301_48], which has been the standard for nearly two decades. Other metrics...

RAGELO: QUERY–CONTEXT EVALUATION

Zeta Alpha

EXAMPLE PROMPT:

Evaluate if a document contains relevant information to answer a question submitted by a user.
You should write one sentence explaining why the document is relevant or not for the user question. A document can be:
- **Not relevant**: The document is not on topic.
- **Somewhat relevant**: The document is on topic but does not fully answer the user question.
- **Very relevant**: The document is on topic and answers the user question.

[user question]
What metrics can be used to evaluate text summarization?

[document content]
precision-based evaluation metric which considers exact n-gram matches. For a given value of n, the precision is computed as the fraction of n-grams in the generated hypothesis which match...

RETRIEVED DOCS EVALS

qid	did	answer
1	fc129c7f70	Very relevant: The document directly addresses the user question by discussing intrinsic and extrinsic metrics for evaluating text summarization, including summary coherence, informativeness, and the effect of summarization on other tasks.
1	4acbef3aaa	Somewhat relevant: The document discusses the BLEU score, which is a metric used to evaluate machine translation and can be applied to text summarization, but it does not cover other metrics that can also be used for evaluating text summarization.

RAGELO: PAIRWISE EVALUATION

Zeta Alpha

EXAMPLE PROMPT:

Evaluate the quality of the responses provided by two AI assistants tasked to answer the question displayed below, based on a set of documents retrieved by a search engine.
You should choose the assistant that best answers the user question based on a set of reference documents that may or not be relevant. Answers cite documents using square brackets.
For each reference document, you **will be provided with a reasoning explaining why the document is or is not relevant.**
Your evaluation should consider factors such as comprehensiveness, correctness, helpfulness, completeness, accuracy, depth, and level of detail of their responses. Answers are comprehensive if they show the user multiple perspectives in addition to but still relevant to the intent of the original question. .
Details are only useful if they answer the user question. If an answer contains non-relevant details, it should not be preferred over one that only use relevant information.
Begin your evaluation by **explaining why each answer correctly answers the user question**. Then, you should compare the two responses and **provide a short explanation on their differences**. 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. 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]
What metrics can be used to evaluate text summarization?
[Reference Documents]
[d6e3da57be79b9dba955ce0d1ff48c62893177e4_2] Very relevant: The document directly addresses the user's question by discussing ROUGE metrics (Precision, Recall, F-score, and longest common subsequence) which are used to evaluate text summarization, specifically focusing on their application and effectiveness in measuring the quality of summaries.
[dd2b71271f63518f0927b4d7fbbae837e69423bd_38] Very relevant: The document directly addresses the user question by detailing the ROUGE Score, a specific set of metrics used to evaluate the quality of text summarization, explaining its different variants and how they function.
... more reference documents ...
[The Start of Assistant A's Answer]
In the quest to understand the metrics used to evaluate text summarization, we have explored various dimensions, including traditional metrics, novel approaches, and the potential for standardization within the field. Traditional metrics like ...
[The End of Assistant A's Answer]
[The Start of Assistant B's Answer]
In evaluating text summarization, several metrics are employed to assess different aspects of the generated summaries. The ROUGE metric, which stands for ...
[The End of Assistant B's Answer]

PAIRWISE ANSWER EVALUATIONS

qid	agent_a	agent_b	raw_answer	answer
1	naive_rag	rag_fusion	Assistant A's response is very brief and only mentions... Assistant B's response is more comprehensive... Final verdict: [[B]]	B
1	rag_fusion_agent	naive_rag	Assistant A provides a comprehensive response that discusses various aspects... Assistant B, on the other hand, offers... [[A]]	A

RAGELO: ELO RANKING

Zeta Alpha

number games: 5792 (Prompt Type - ELO score)	
gpt4_qa_chain_expanded_reasoning_rerank	1256
gpt4_qa_chain	1204
gpt4_qa_chain_filter	1197
chatgpt_qa_chain_filter	1119
gpt4_qa_chain_compressor	1106
davinci_qa_chain_compressor	1104
davinci_qa_chain	1071
gpt4_zeta_alpha_qa_chain	1069
chatgpt_qa_chain_compressor	1013
chatgpt_qa_chain_expanded_reasoning_rerank	980
chatgpt_qa_chain	969
davinci_qa_chain_expanded_reasoning_rerank	917
chatgpt_zeta_alpha_qa_chain	886
gpt4_zeta_alpha_qa_chain_v2	873
davinci_zeta_alpha_qa_chain	871
davinci_zeta_alpha_qa_chain_v2	827
davinci_qa_chain_filter	779
chatgpt_zeta_alpha_qa_chain_v2	749

QUIS CUSTODIET IPSOS CUSTODES?

WHO IS EVALUATING THE JUDGE?

WINNING IS NUANCED

Zeta Alpha

Who are the main deep learning researchers in The Netherlands?

In The Netherlands, some notable researchers in the field of deep learning and related areas include:

1. **Geert Litjens**: Known for his work in applying deep learning to medical image analysis at Radboud University Medical Center [83d301b0].
2. **Peter A. N. Bosman**: A group leader at the Evolutionary Intelligence research group at the Dutch National Research Institute for Mathematics and Computer Science (Centrum Wiskunde & Informatica) and a professor at Delft University of Technology, focusing on scalable model-based evolutionary algorithms [3b5f7e39].
3. **Maarten de Rijke**: Engaged in research on fairness, accountability, and transparency in artificial intelligence, supported by several Dutch and European research grants [b42e2f79].

Do you mean “main” as in the most prominent PIs/groups (short list), or a more comprehensive directory by university and research area? I can do either. Below is a short curated list of well-known deep-learning researchers and groups based in the Netherlands (by institution), with a one-line note on area — I can expand each entry with links/publications if you want.

University of Amsterdam (UvA) / Amsterdam Machine Learning Lab (AMLAB)

- Max Welling — probabilistic deep learning, graph models, variational methods; leads AMLAB and several industry labs [2d73f75f48474b5de6675a29e17d32fa1b63001d_54].
- Wouter Kool — deep RL / neural combinatorial optimization (routing problems) (co-author with Welling) [9ad9c6ada071048305fb513045e2e31bb582acbd_33].
- Herke van Hoof — reinforcement learning, learning for combinatorial problems (works with Welling) [ed050789].
- Maarten de Rijke — deep learning for information retrieval / NLP (large IR and NLP group

Verdict: B

Explanation: Both Assistant A and Assistant B provide lists of deep learning researchers in the Netherlands, but they approach the task differently and with varying levels of detail and grounding in the provided documents.

Comparing the two responses, Assistant B provides a more comprehensive and informative answer, detailing not only the names but also the specific areas of expertise and institutional affiliations of the researchers. This level of detail is particularly useful for understanding the scope and focus of deep learning research in the Netherlands.

Zeta Alpha

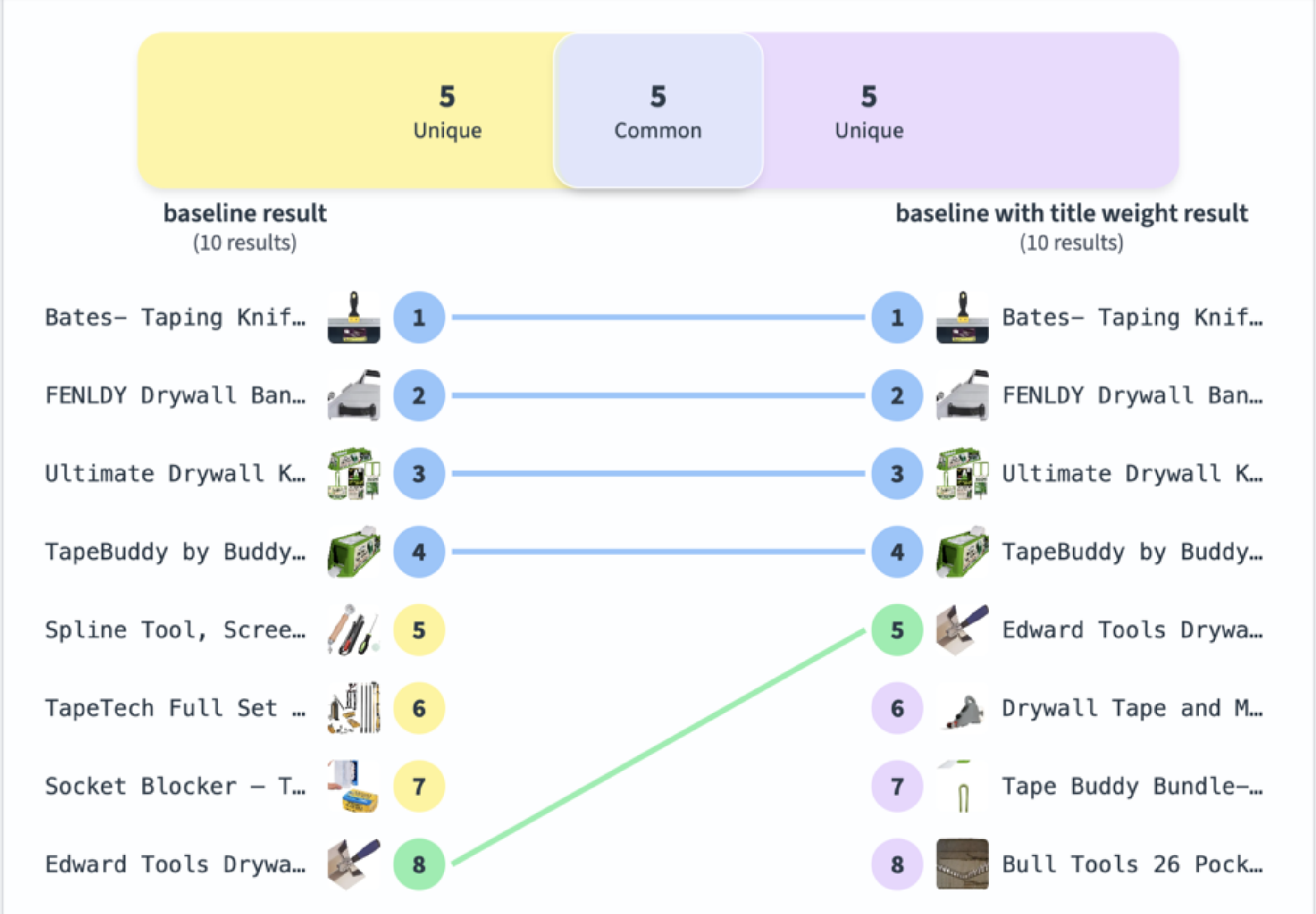
WHAT IF WE COMBINE THE EVALUATION POWER OF RAGELO WITH THE EYEBALLING POWER OF SRW?

+ Ragelo =

RAG EVALUATION SUPERLOVE?

SEARCH RELEVANCE WORKBENCH

Zeta Alpha



Zeta Alpha

DEMO!

LESSONS LEARNED IN THIS EFFORT

- **SRW** is very good for eyeballing **search results**, but using it for RAG requires some **changes**.
- **RBO/Jaccard** signals **hit** discrepancies. What is the equivalent for **RAG answers**?
- How do we eyeball **tool calls**?
- ELO tournaments involve multiple RAG versions, how to **eyeball** the whole **tournament**?
- SRW APIs are very useful. Should we have vibe coded our UI instead?