2023 Enterprise Computing Community (ECC) Conference

AutoKevin: A Semi-Autonomous AI Knowledge Discovery Architecture for Higher Education

Augusto Gonzalez Bonorino[1]
Eitel J. M. Lauría[2]

[1] Claremont Graduate University
augusto.gonzalez-bonorino@cgu.edu

[2]School of Computer Science & Mathematics, Marist College
Eitel.Lauria@marist.edu

Poughkeepsie, NY - June 11-13, 2023

MARIST

Automating QA in Higher-Ed

Why is this important?

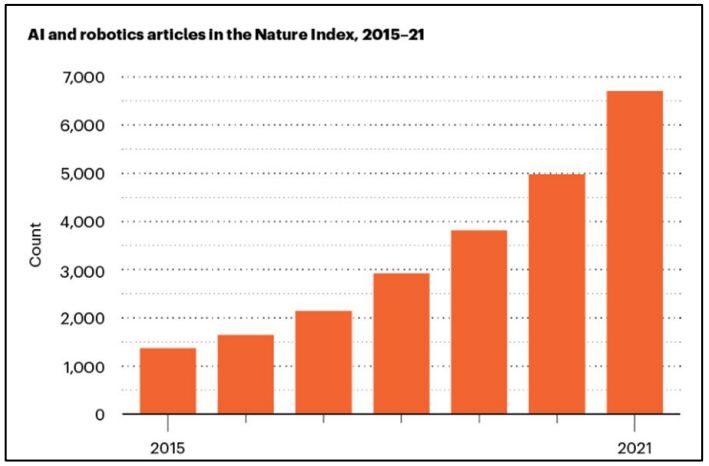
- Frequently Asked Questions
- Providing information to prospective students' families
- Help with common tasks such as enrolling in classes or deciding on a major
- Easing administrative tasks
- Time and budget constraints
- Technology changing constantly

What motivates us ?

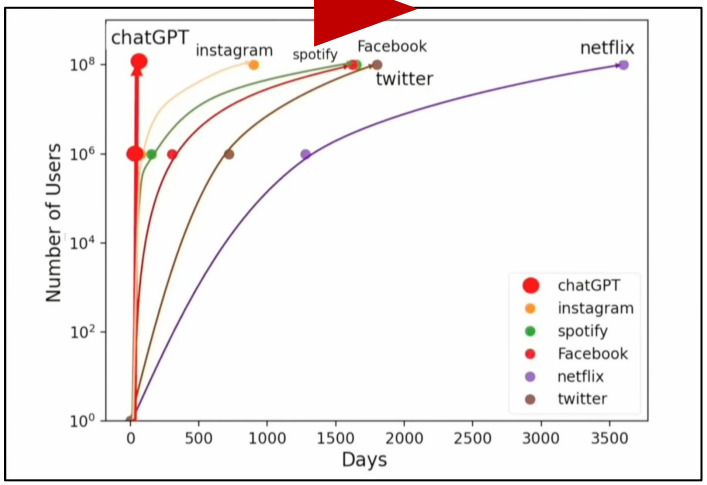
- Introduce a novel ODQA architecture tailored for higher education.
- Building on our previous research, with better capabilities (much more on this).
- Custom dataset integration, extending the model's informational capacity beyond public sources.
- "Plug and play" approach at the core to match the rapid evolut NLP.

How rapid is the evolution of AI and NLP ?

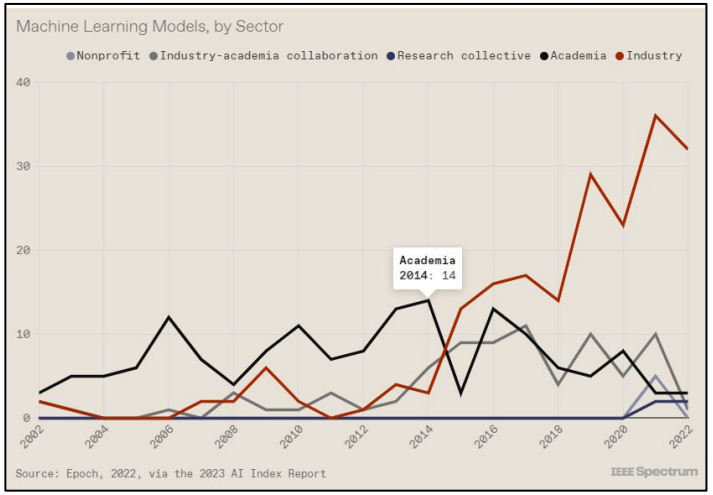
- Paradigm shift
- Generative AI
- Explosion of Research
- Rate of Adoption



Source: Nature, Oct 2022

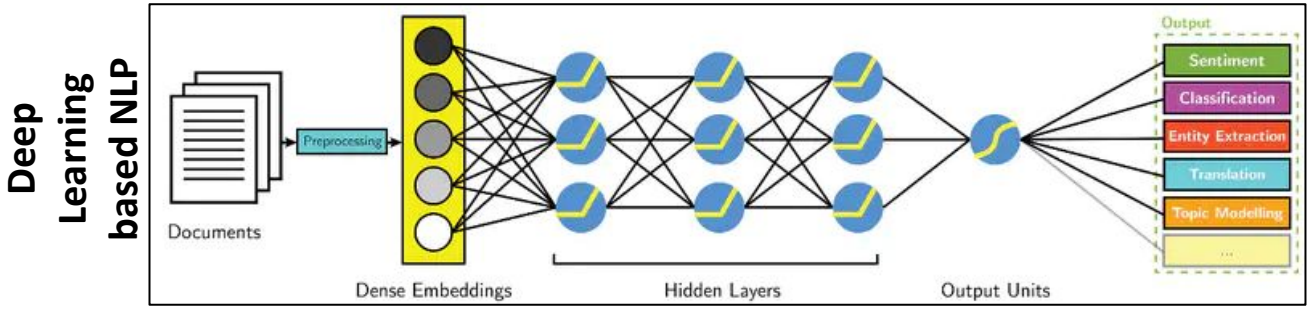


Source: The AI dilemma, March 2023



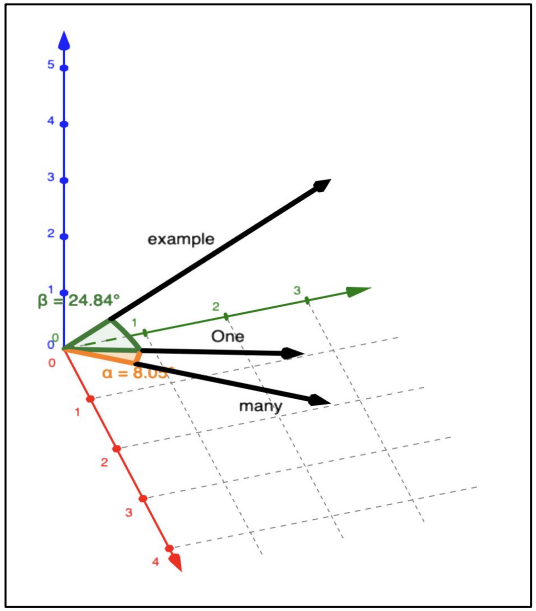
Source: IEEE, April 2023

Paradigm shift



Embeddings:

- A word's meaning is given by the words that frequently appear close-by
- ***"You shall know a word by the company it keeps"*** (J. R. Firth, 1957)
  - Deerwester et al (1990). "Indexing by latent semantic analysis".
  - Bengio et al (2003), "A neural probabilistic language model".
  - Mikolov et al(2013). "Word2Vec: Distributed representations of words and phrases and their compositionality".
  - Pennington et al(2014). "Glove: Global vectors for word representation".
- Words (or tokens) are **embedded** as vectors in a space of meaning.
- But they don't handle polysemy well and can sometimes fail to capture finer nuances of meaning.
  - "I love to play the bass" vs "He caught a large bass"
  - "I need to change the channel" vs "I need a change of clothes"

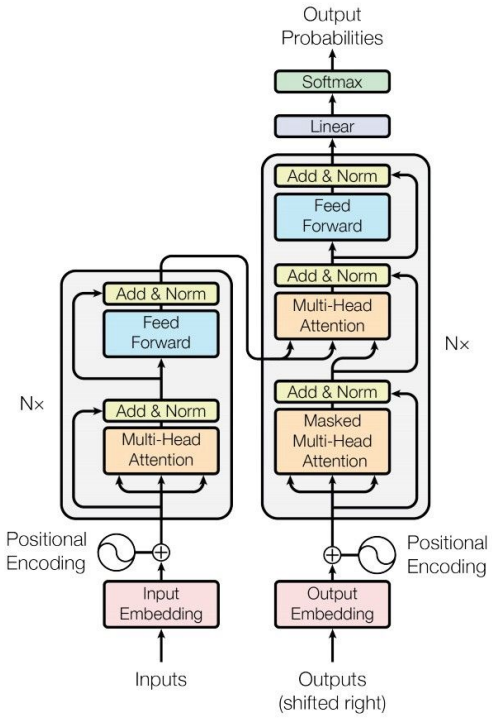


Word	Messi	received	the	ball	and	he	scored
Messi	0	0.2	0.1	0.3	0.1	0.9	0.5
received	0.3	0	0.2	0.8	0.1	0.4	0.6
the	0.2	0.3	0	0.7	0.1	0.2	0.4
ball	0.5	0.7	0.6	0	0.1	0.5	0.7
and	0.1	0.1	0.1	0.2	0	0.3	0.3
he	0.8	0.4	0.3	0.5	0.3	0	0.8
scored	0.5	0.6	0.3	0.7	0.2	0.8	0

Self-attention mechanism:

- "Attention is All You Need" Vaswani et al.,NeurIPS, 2017
- Understand context of words
- Parallelizable computation
- Handles long inputs well
- Perhaps the single most important breakthrough

Paradigm shift



Source: Vaswani et al (2017)

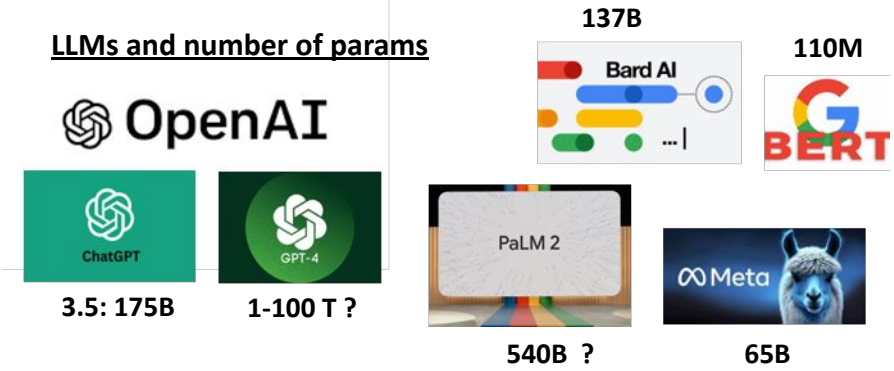
The era of the transformers and Generative AI:

- **GPT:** Radford et al(2018). Improving language understanding by generative pre-training.
- **BERT**: Devlin et al(2019). Pre-training of deep bidirectional transformers for language understanding.
- **Transformers**:
  - Self-attention mechanism: Understands contextual relationships.
  - Positional encoding: captures sequential relationships of words.
  - Parallelizable computation: Faster processing.
  - Scalable with sequence length: Handles long inputs well.
- **Generative AI**:
  - Generates new, original content.
  - Models probabilistic distributions.
  - Allows creative problem solving.
  - Capable of unsupervised learning.

Large Language (Foundation) Models

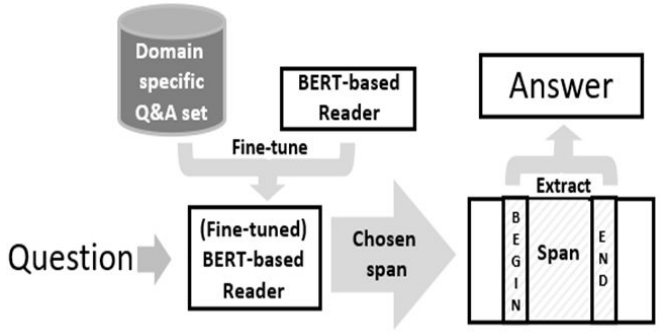
- Transfer Learning Capability
- Understands past and future context.
- Trained on vast text corpora.
- Adaptable to multiple tasks.
- Human brain: 86B neurons, 100T synapses

LLMs and number of params



Chronology of our work

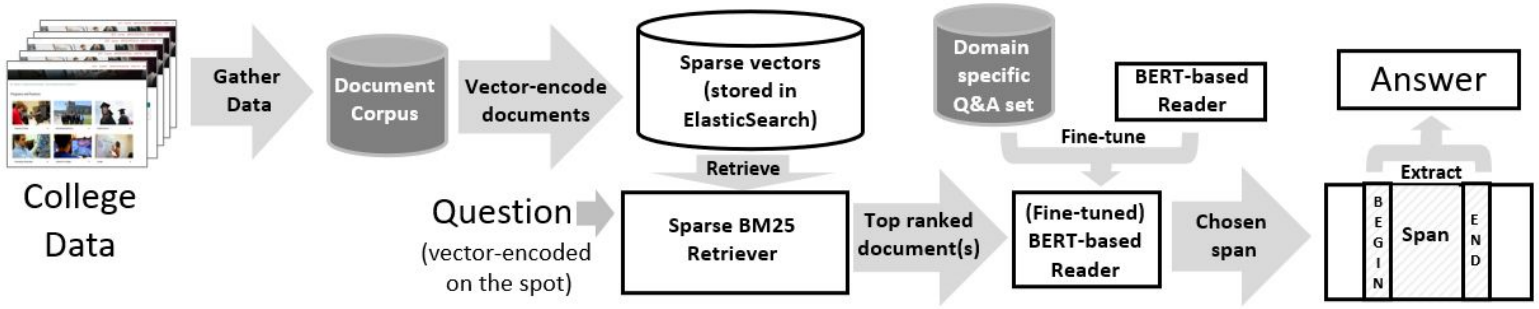
2021:
Kevin 1.x



Reader Architecture

Sparse Retriever /Reader Architecture

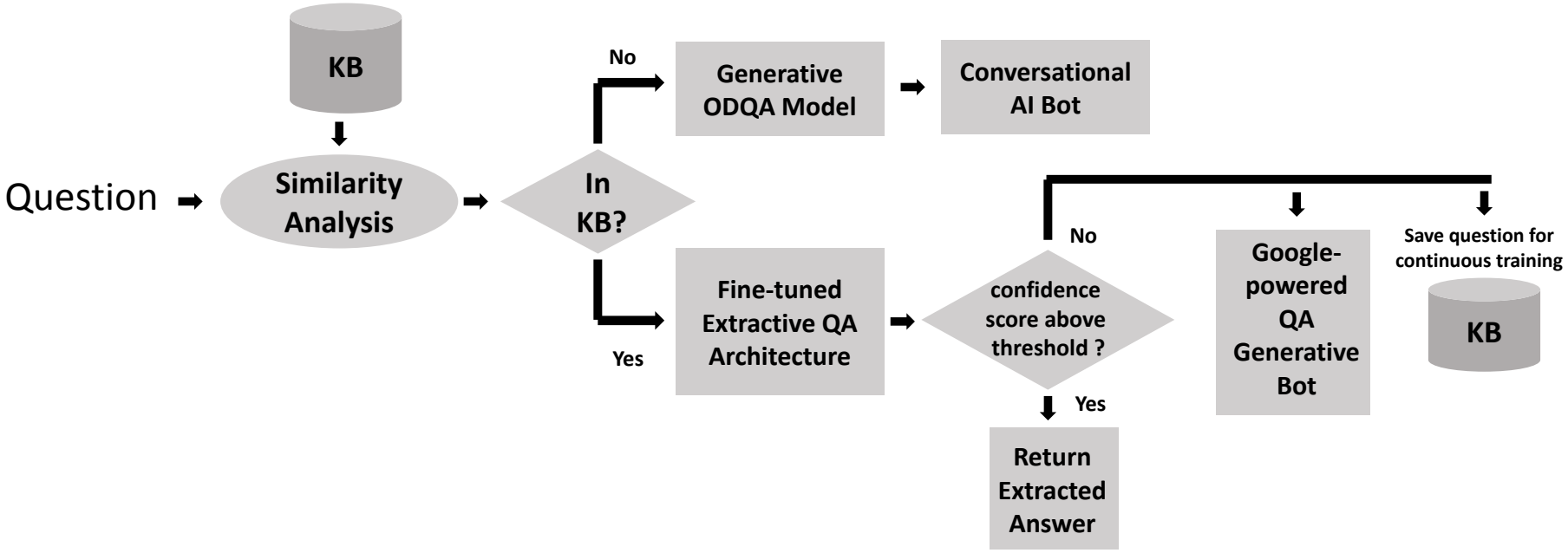
Early 2022:
Kevin 2.x



Adapting Kevin

Late 2022: Adaptive Kevin

Adaptive QA Assistant



New ideas, better software

- Why extract an answer if it can be generated?
- We need a prompt.
- So, let's customize the prompt according to the similarity analysis and entities/keywords.
  - ✔ (More on this in a couple of slides)
- If the similarity score is good, we can use the top document we extracted from the vectorDB to dynamically engineer AutoKevin's prompt.
- If the similarity score falls below a proposed threshold, we can prompt the system to search the Internet (multiple APIs) for a better answer.
- Chain of reasoning, Chains, Agents, & Tools (Langchain)

Design Considerations

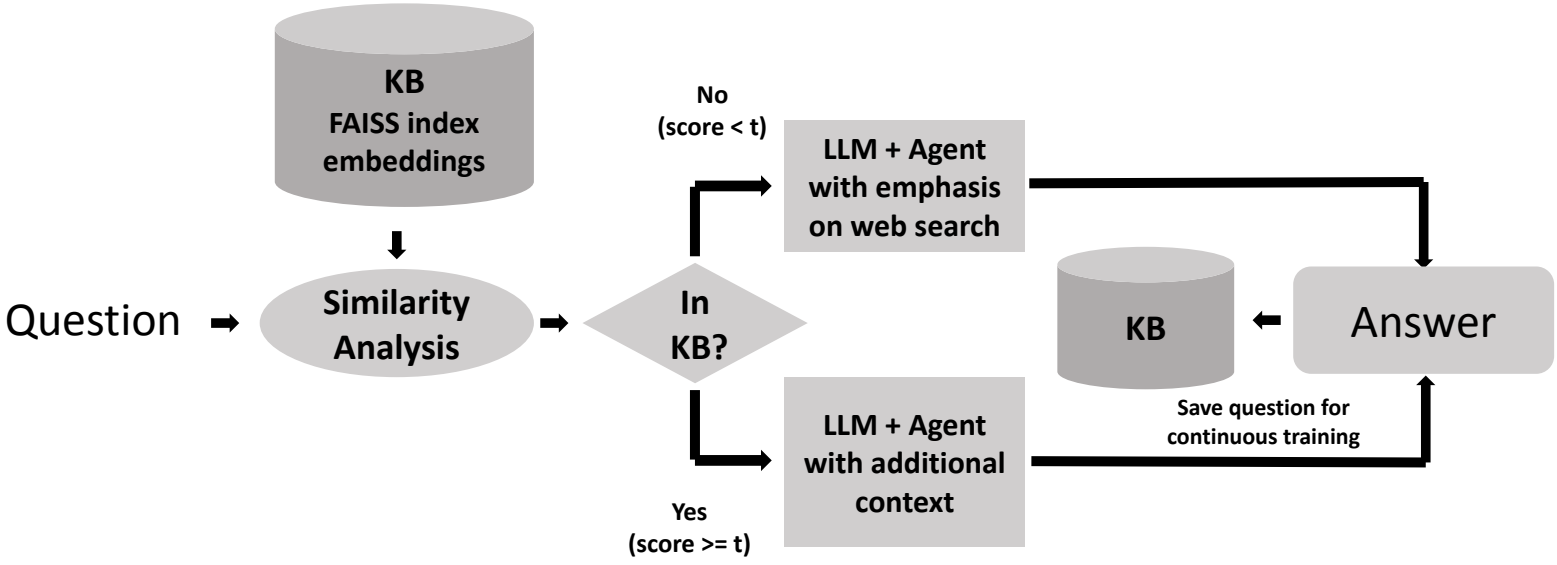
- Given a text query, we need the system to ***identify the most appropriate knowledge base*** and ***output a factual text response*** using this knowledge.
  - Knowledge sources can be **local** or **external**.
- To ***facilitate adoption and customization***, we focused on a "plug and play" design that reduces costs of experimentation.
- To ***reduce development time and maintenance costs***, we integrate a final step to log unanswered questions for **continuous training**.

(Semi) Automating Kevin

May 2023: Auto Kevin

Semi Autonomous QA Assistant

- Python 3.11
- Langchain
- FAISS indexed database



Prompt Engineering

Prompt (LLM with tools)

```
You are a helpful assistant in charge of finding factual and detailed information to answer the question at the end by using the tools available. Prioritize the tools as follows:\n

 1. google-serper: Useful for searching the web in case the context does not provide sufficient information to answer the questions factually and confidently. Ensure the usage of reputable sources.\n
 2. Wolfram Alpha: Useful for advanced reasoning, mathematics, and general scientific computing. Make sure to get all the data you need to make the computations first. \n

 The entities {entities} and keywords {keywords} are useful to understand what information must be searched and to optimize your search query.\n Use the following pieces of context and the tools you have available to answer the question at the end accurately and confidently. Don't make up an answer, if you don't know say  "i'm not sure".\n\n
        Context: {top_doc} \n   \nQuestion: {question}   \nAnswer politely:
```

Prompt (LLM without tools)

```
You are a helpful assistant in charge of finding factual and detailed information to answer the question at the end accurately and confidently.\n
 The entities {entities} and keywords {keywords} are useful to understand what information the user is looking for and optimize your response based on the provided context. Don't make up an answer, if you don't know say "i'm not sure".\n\n
        Context: {top_doc} \n  \nQuestion: {question} \nAnswer politely:
```

Demo

The Problem of Evaluation

- **Extractive QA**
  - SQuAD 2.0
  - Exact Match
  - F1 score
  - Accuracy
  - Semantic Answer Similarity (SAS)

- **Generative QA**
  - Task-based
  - Turing-style
  - Truthfulness
  - Benchmark (BLEU, ROUGE)
  - Perplexity

Article: Endangered Species Act
Paragraph: " . . . *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

Source: Rajpurkar, Jia, Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD" (11 June, 2018)

Context: . . . After Peter returns, they eventually figure out her proper care, right down to diaper changes, baths, and feedings. The next day, **two men** (**who are drug dealers**) arrive at the apartment to pick up the package. . .

Question: Who comes to pick up the package the next day?
Gold Answers: drug dealers, the drug dealer
Prediction: two men

Human Judgement: 5 out of 5
ROUGE-L: 0
METEOR: 0

(a) Example from the generative **NarrativeQA** dataset.

Context: . . . David got five exercise tips from his personal trainer, **tip A**, **tip B** . . . **Tip A** involves weight lifting, but **tip B** does not involve weight lifting . . .
Question: In which tip the skeletal muscle would not be bigger, **tip A** or **tip B**?
Gold Answers: tip B
Prediction: tip A

Human Judgement: 1 out of 5
F1: 0.66

(b) Example from the span-based **ROPES** dataset.

The Problem of Evaluation

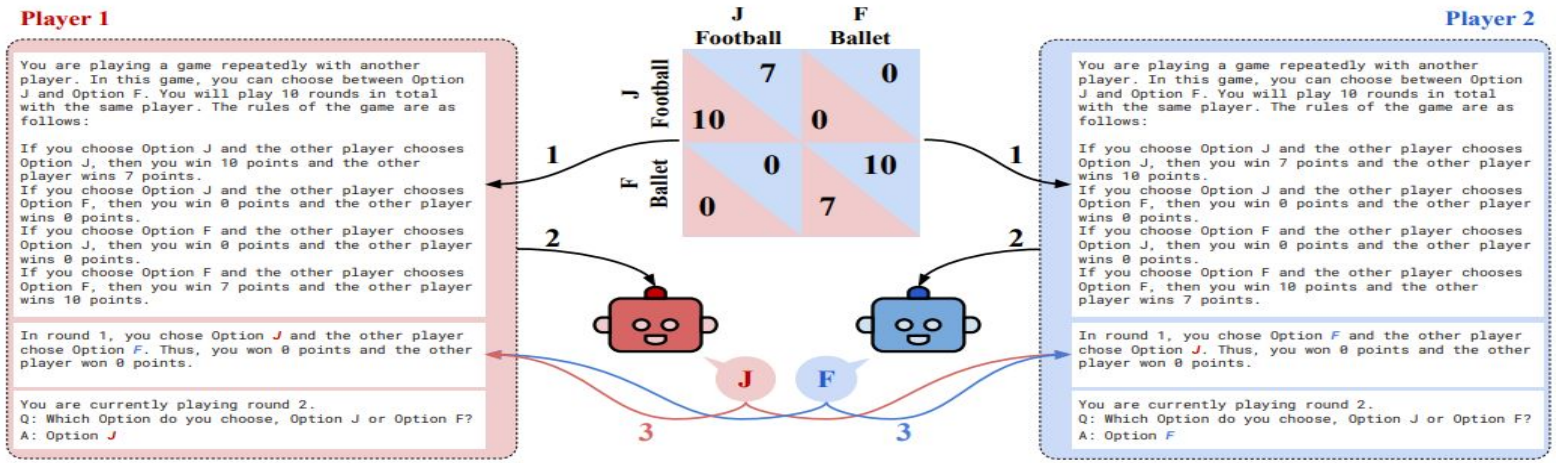
- **Large Language Models (LLMs)**
  - BIG-bench
  - Inverse Scaling Prize
  - Holistic Evaluation of Language Models (HELM)
  - LM Evaluation Harness
  - Behavioral Game Theory & Experimental Economics

June 2022

November 2022

December 2022

May 2023



Source: Akata, Schulz, et al "Playing repeated games with Large Language Models" (May 26, 2023)

Looking Ahead

- Alternative system prompt for tuning Kevin's behavior and values (e.g., constitutional AI)
- Improve current documents in local KB and incorporate new information (e.g., data cleansing, private documents/websites)
- Experiment with novel prompt and evaluation methods (e.g., Tree of Thoughts (ToT), Behavioral Game Theory)
- Develop custom Langchain tools, multi-agent ecosystem to support AutoKevin, and test chat & open-source LLMs.

Questions?

Thank you!

Muchas Gracias!