

Using LLMs for your Research

Prof. Guido Zuccon

g.zuccon@uq.edu.au

ielab & Queensland Digital Health Center (QDHeC)
The University of Queensland, Australia

www.ielab.io

In this session...

- How do LLMs work?
- How can LLMs help me with my research? Examples and interactive sessions
- We do not have enough time to cover all use-cases. To check out the use-cases we had no time for:
 - <https://tinyurl.com/495bntdd>

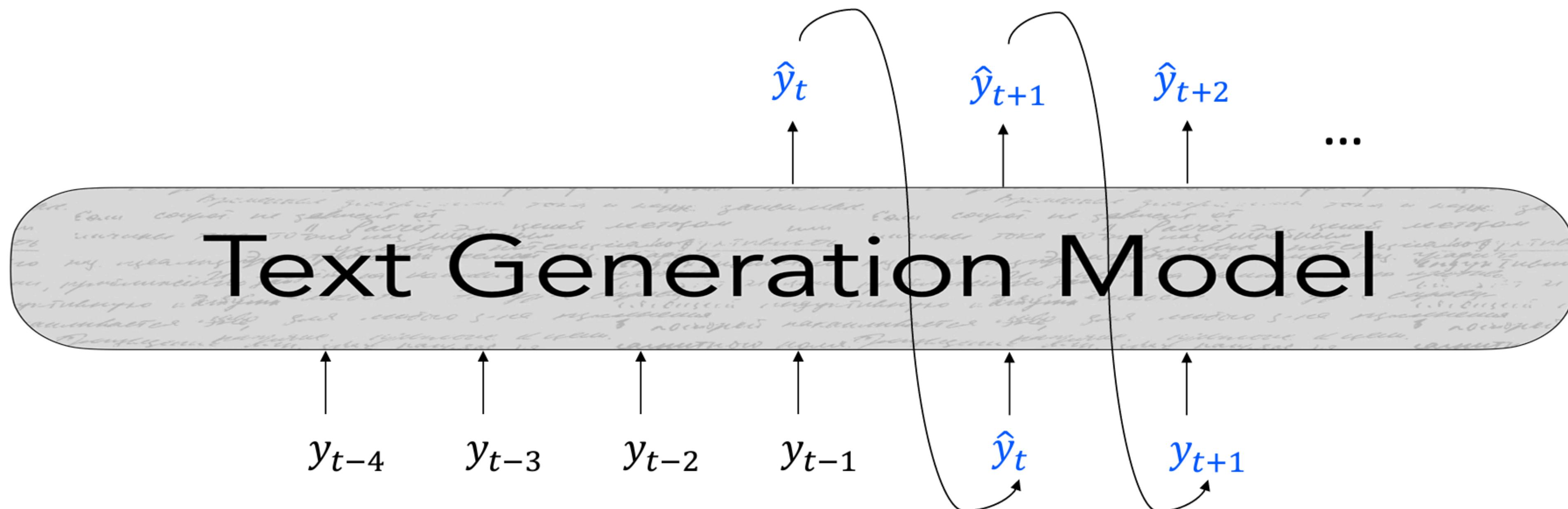
Large Language Models

- A LLM is a machine learning model that is able to achieve general-purpose language understanding and generation
 - Often LLMs are referred to foundational models
 - (pre-)trained using self-supervised learning on a massive amount of data, consuming large computational resources
- The best known LLM is ChatGPT, launched November 30, 2022
- Generative LLMs use a decoder-only, or encoder-decoder architecture
 - BERT is not a generative LLM

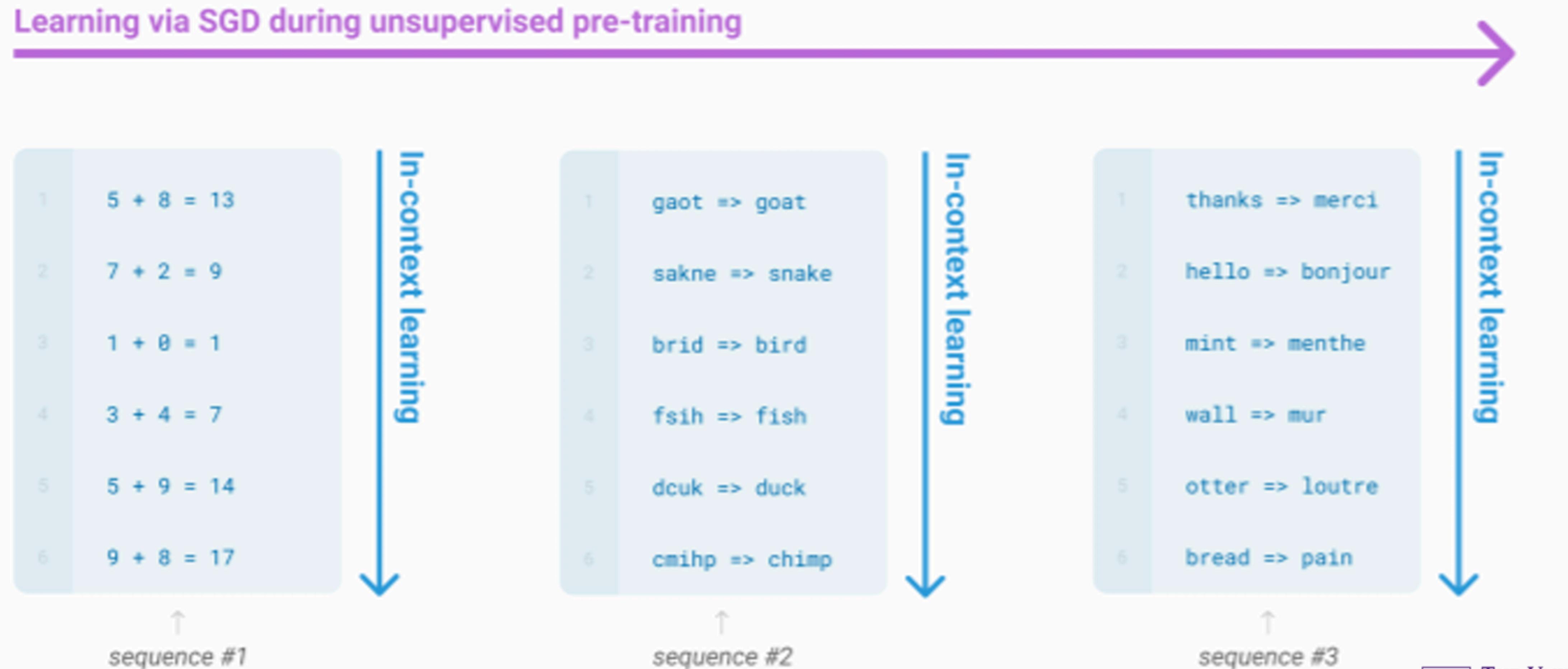


On Autoregressive Decoders

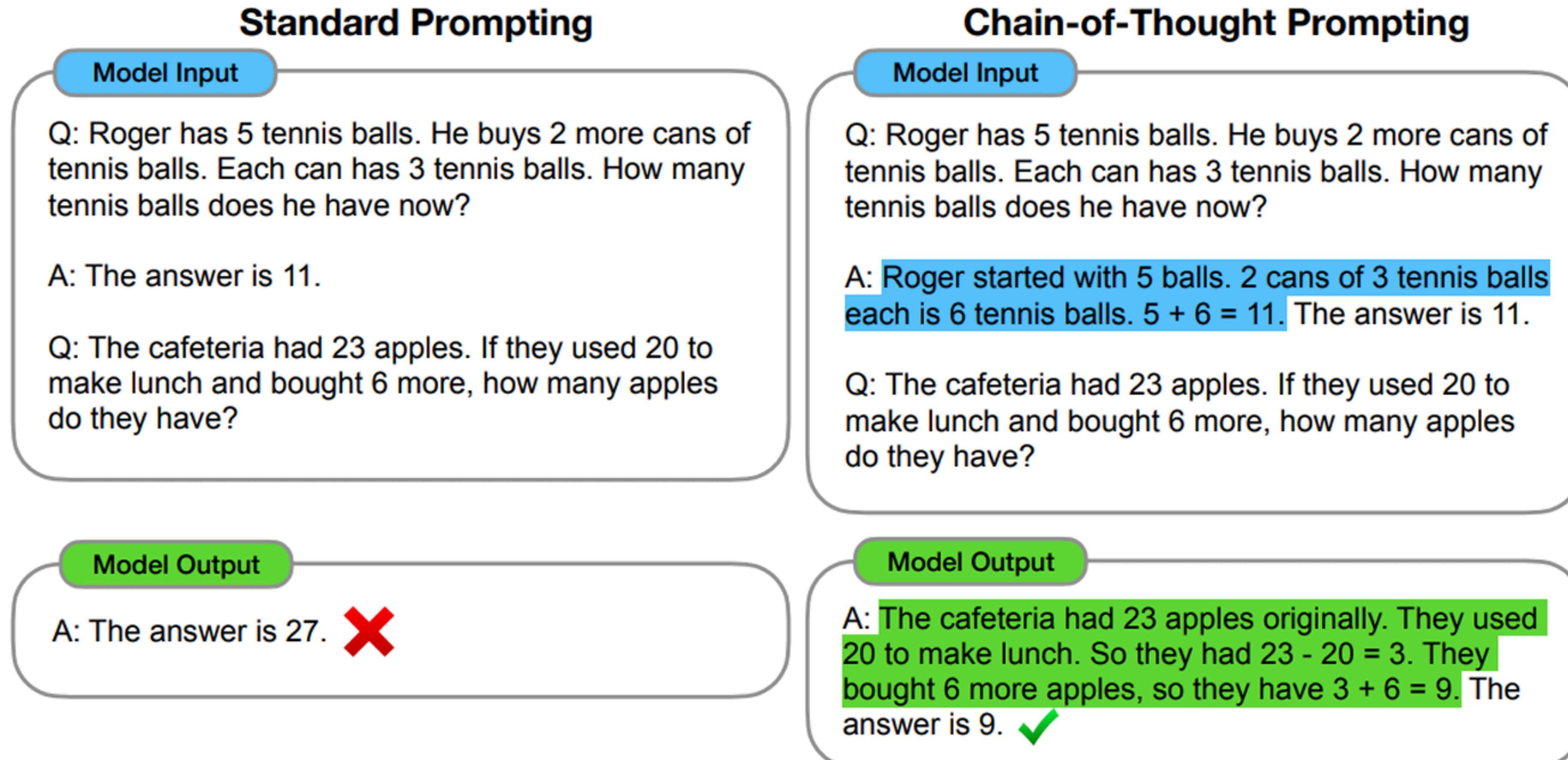
- In an autoregressive text generation model, at each time step t , the model takes a sequence of tokens $\{y\}_{<t}$ as input, and outputs a new token \hat{y}_t



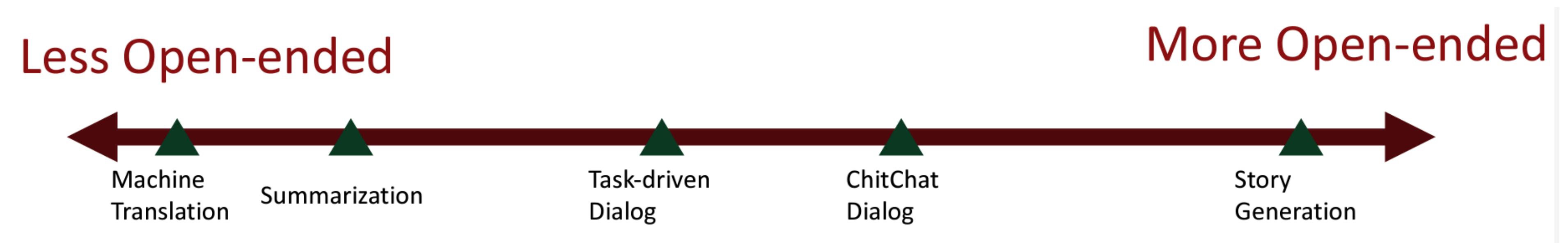
In-context Learning (GPT-3 and onwards)



Prompts as task specification and scratch pad: chain-of-thought

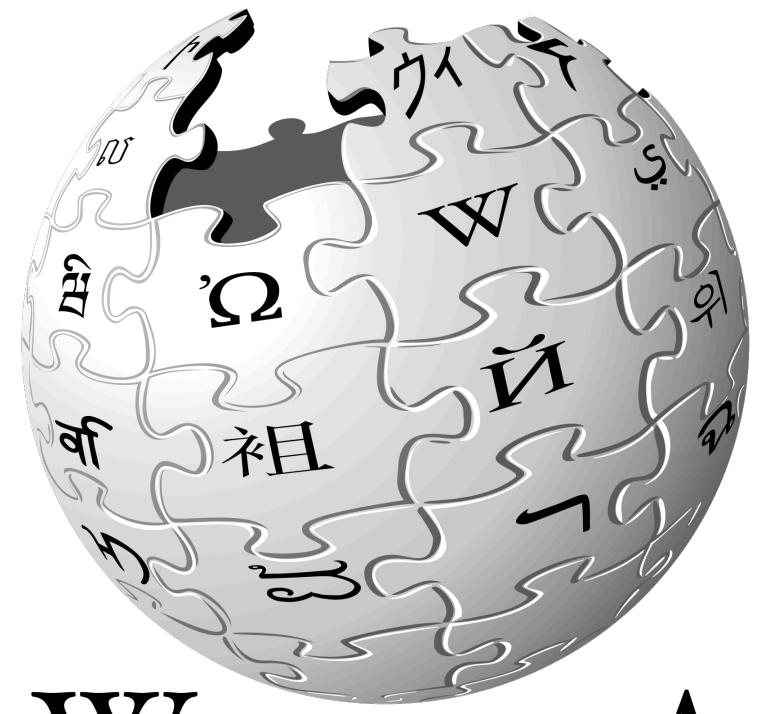


Natural Language Generation Tasks



- Open-ended generation: the output distribution still has high freedom
 - Typically uses a (autoregressive) decoder
- Non-open-ended generation: the input mostly determines the output generation.
 - typically use a encoder-decoder system

LLMs Hallucinate



WIKIPEDIA
The Free Encyclopedia

“a confident response by an AI that does not seem to be justified by its training data.”

This has generally evolved into: the model speaks false knowledge as if it is accurate, and it does so in a very convincing way

LLMs Hallucinate

G

Will drinking vinegar dissolve a stuck fish bone



Drinking vinegar may help dissolve a stuck fish bone, but it is not a guaranteed or immediate solution. Vinegar is acidic, and the acid could potentially help break down the bone over time. However, the effectiveness of this method depends on several factors, including the size and location of the bone, as well as your individual tolerance to vinegar.



LLMs Hallucinate

G

Will drinking vinegar dissolve a stuck fish bone



Drinking vinegar may help dissolve a stuck fish bone, but it is not a guaranteed or immediate solution. Vinegar is acidic, and the acid could potentially help break down the bone over time. However, the effectiveness of this method depends on several factors, including the size and location of the bone, as well as your individual tolerance to vinegar.



However:

1. Medical evidence shows vinegar is not a good remedy to dissolve a fishbone stuck in your throat

Kim, J.P., Kwon, O.J., Shim, H.S., Kim, R.B., Kim, J.H. and Woo, S.H., 2015. Analysis of clinical feature and management of fish bone ingestion of upper gastrointestinal tract. *Clinical and Experimental Otorhinolaryngology*, 8(3), pp.261-267.

LLMs Hallucinate

G

Will drinking vinegar dissolve a stuck fish bone



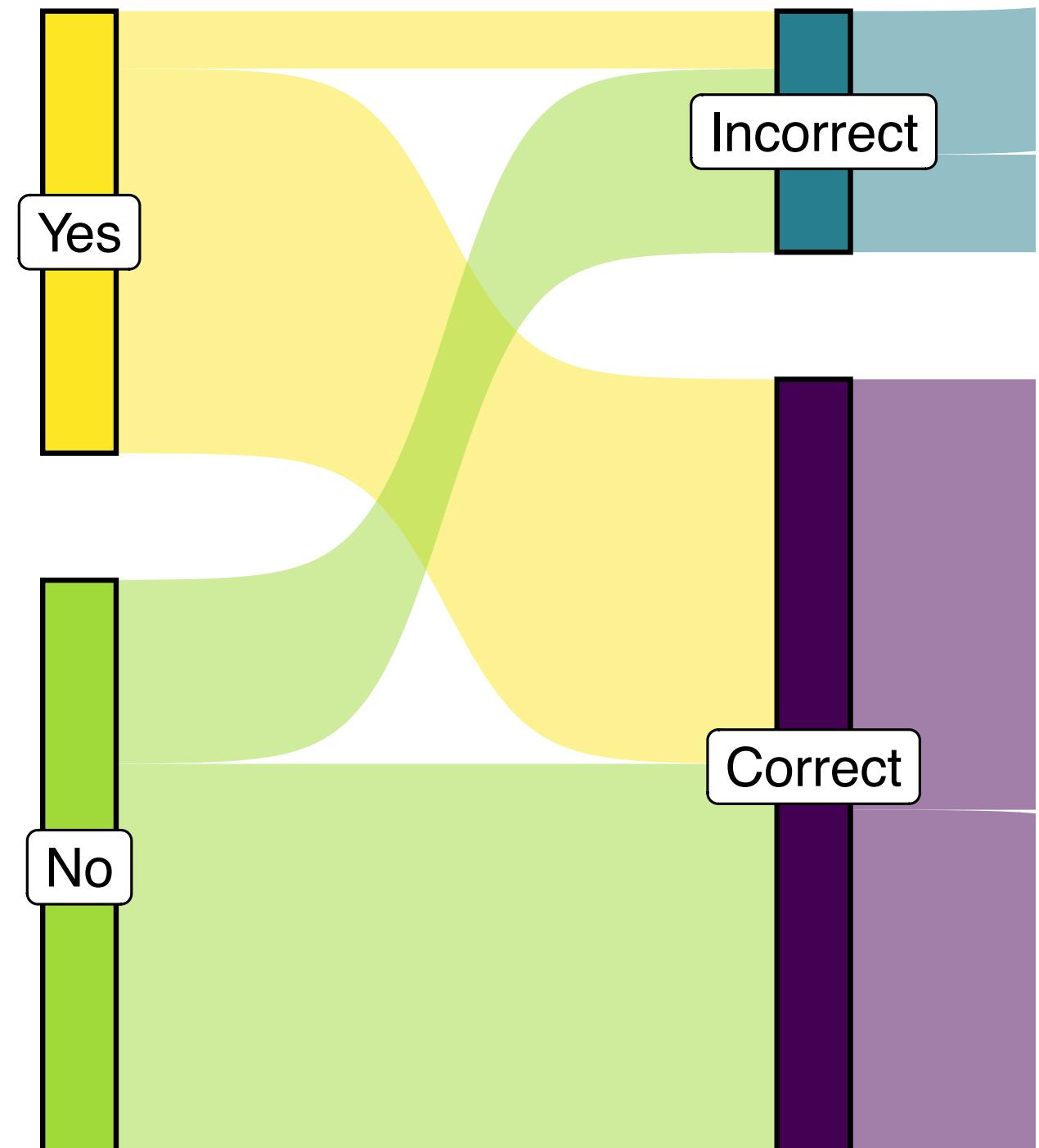
Drinking vinegar may help dissolve a stuck fish bone, but it is not a guaranteed or immediate solution. Vinegar is acidic, and the acid could potentially help break down the bone over time. However, the effectiveness of this method depends on several factors, including the size and location of the bone, as well as your individual tolerance to vinegar.



However:

1. Medical evidence shows vinegar is not a good remedy to dissolve a fishbone stuck in your throat
2. Ingesting too much vinegar can actually lead to death

LLMs Hallucinate



Using the TREC Misinformation dataset, we empirically evaluate ChatGPT's hallucinations w.r.t. health questions

Zuccon, G. and Koopman, B., 2023. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*. (EMNLP2023)

RAG: Retrieval Augmented Generation

- Possible explanation:
 - LLMs **hallucinate** because the model has **no correct knowledge** regarding what we asked
 - If the model was provided knowledge, then hallucinations will not occur
 - Idea:
 - Let's **retrieve evidence** about what we asked, and pass this to the LLM — in this way the model will not hallucinate
 - This is called Retrieval Augmented Generation (RAG)

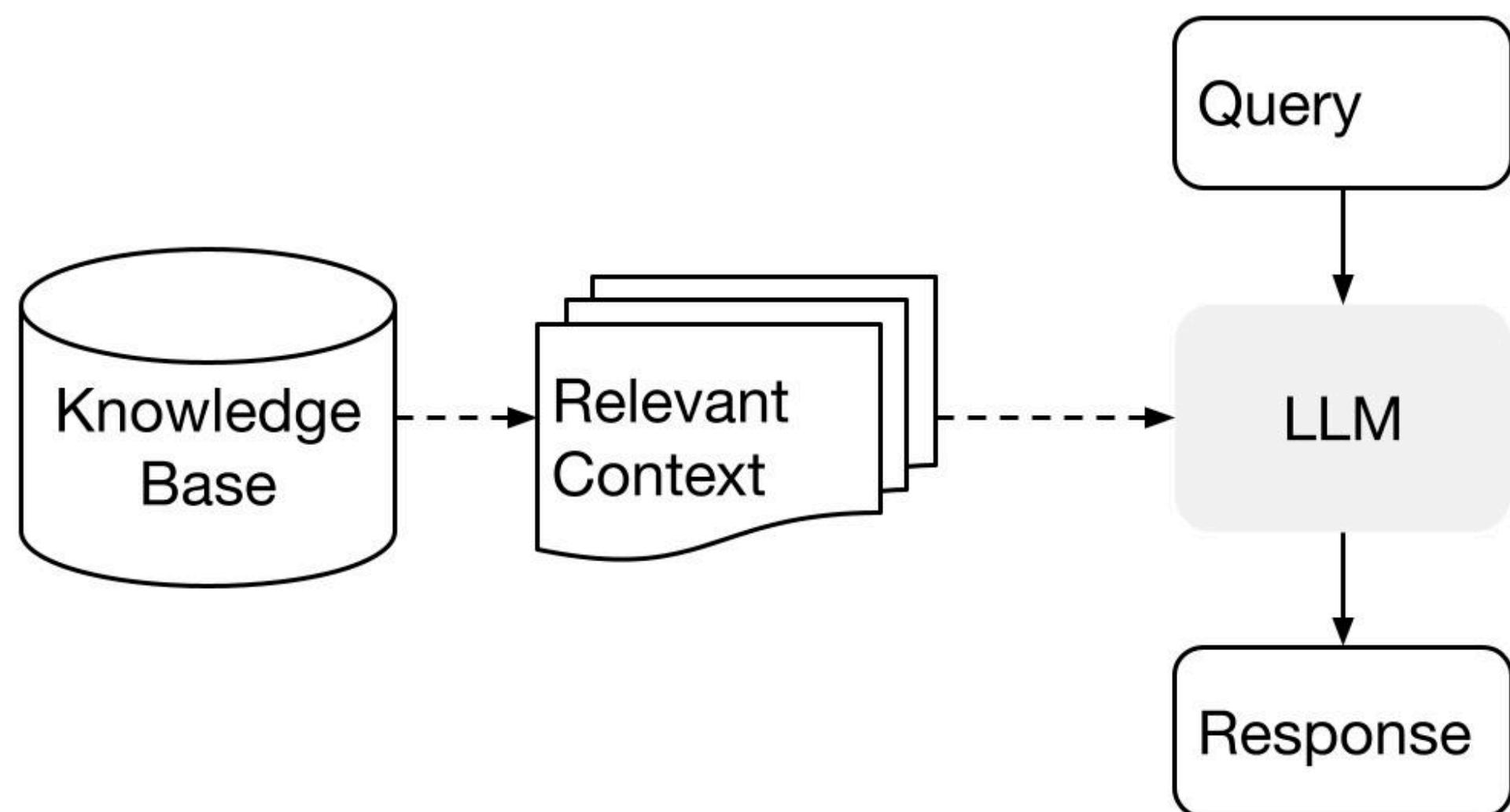


Image from LlamaIndex

Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web

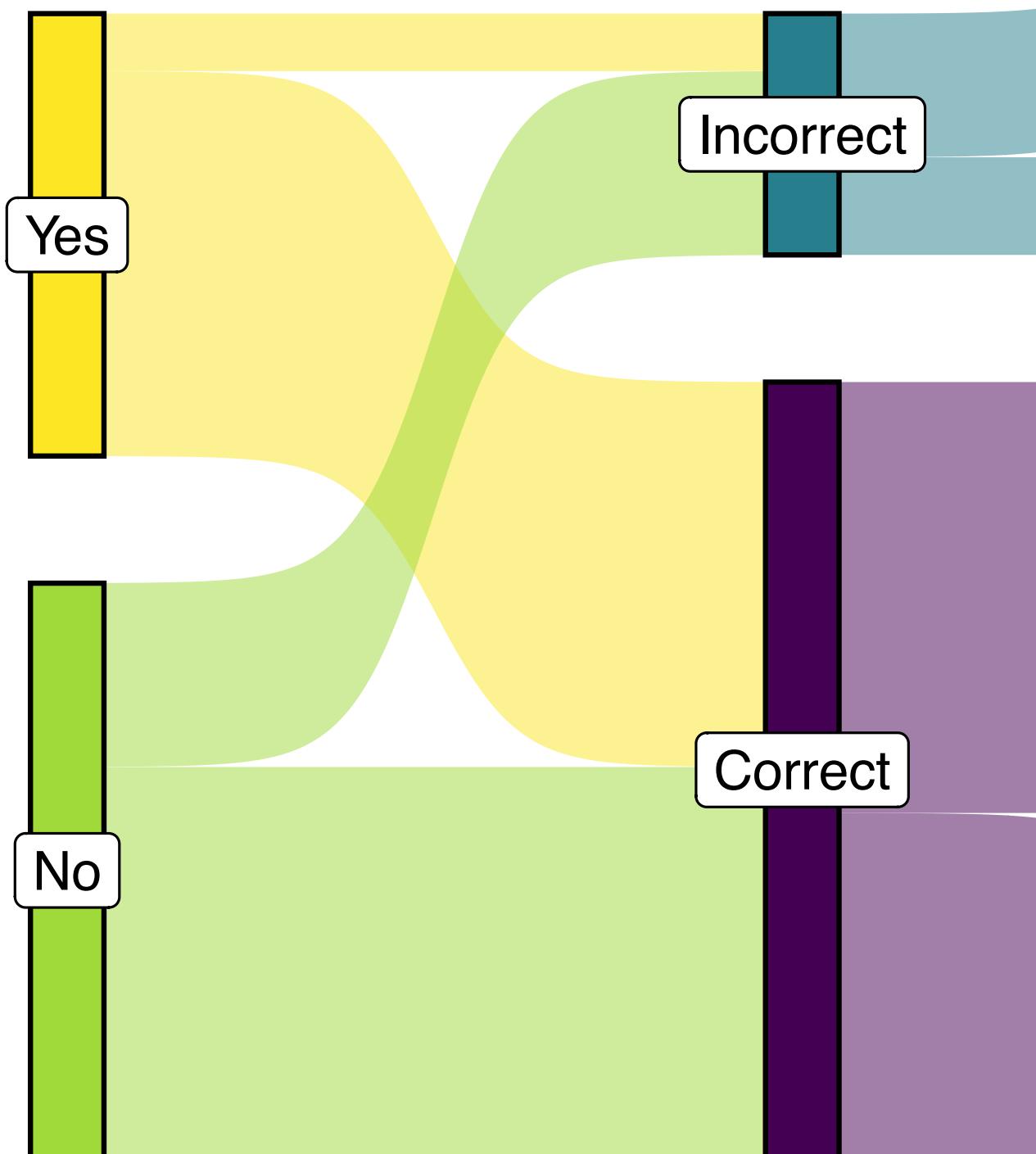
The screenshot shows the Microsoft Bing search interface. At the top, there's a navigation bar with the Microsoft Bing logo, a search bar labeled "SEARCH", and a "CHAT" button which is underlined, indicating it's active. Below the search bar, there's a blue input field containing a user query: "I am planning a trip for our anniversary in September. What are some places we can go that are within a 3 hour flight from London Heathrow?". The main content area displays a response from the AI. It starts with a congratulatory message: "Congratulations on your anniversary! 🎉 There are many places you can go that are within a 3 hour flight from London Heathrow. Here are some suggestions based on your preferences and the best destinations in Europe in September [4](#) [5](#) [6](#):". Following this, there's a bulleted list of three travel suggestions:

- If you like **beaches** and **sunshine**, you can fly to **Malaga** in Spain [1](#), which has a pleasant climate, a lively culture, and a beautiful coastline. You can enjoy the sandy beaches, the historic monuments, and the delicious tapas. Malaga is also a great base to explore other nearby attractions, such as Granada, Seville, and Cordoba.
- If you like **mountains** and **lakes**, you can fly to **Annecy** in France [4](#), which is a charming alpine town near the Swiss border. Annecy is known as the "Venice of France" because of its canals and bridges. You can admire the stunning scenery, visit the medieval castle, and try some local cheese and wine. Annecy is also close to Geneva, Chamonix, and Lyon, if you want to see more of the region.
- If you like **art** and **history**, you can fly to **Florence** in Italy [6](#), which is the birthplace of the Renaissance and a UNESCO World Heritage Site. Florence is a treasure trove of artistic and architectural masterpieces, such as the Duomo, the Uffizi Gallery, and the Ponte Vecchio. You can also explore the Tuscan countryside, taste the famous gelato, and shop for leather goods.

At the bottom left, there's a blue circular icon with a speaker symbol and the text "Ask me anything...".

BingChat is a RAG

RAG is not the Silver Bullet for Solving Hallucinations



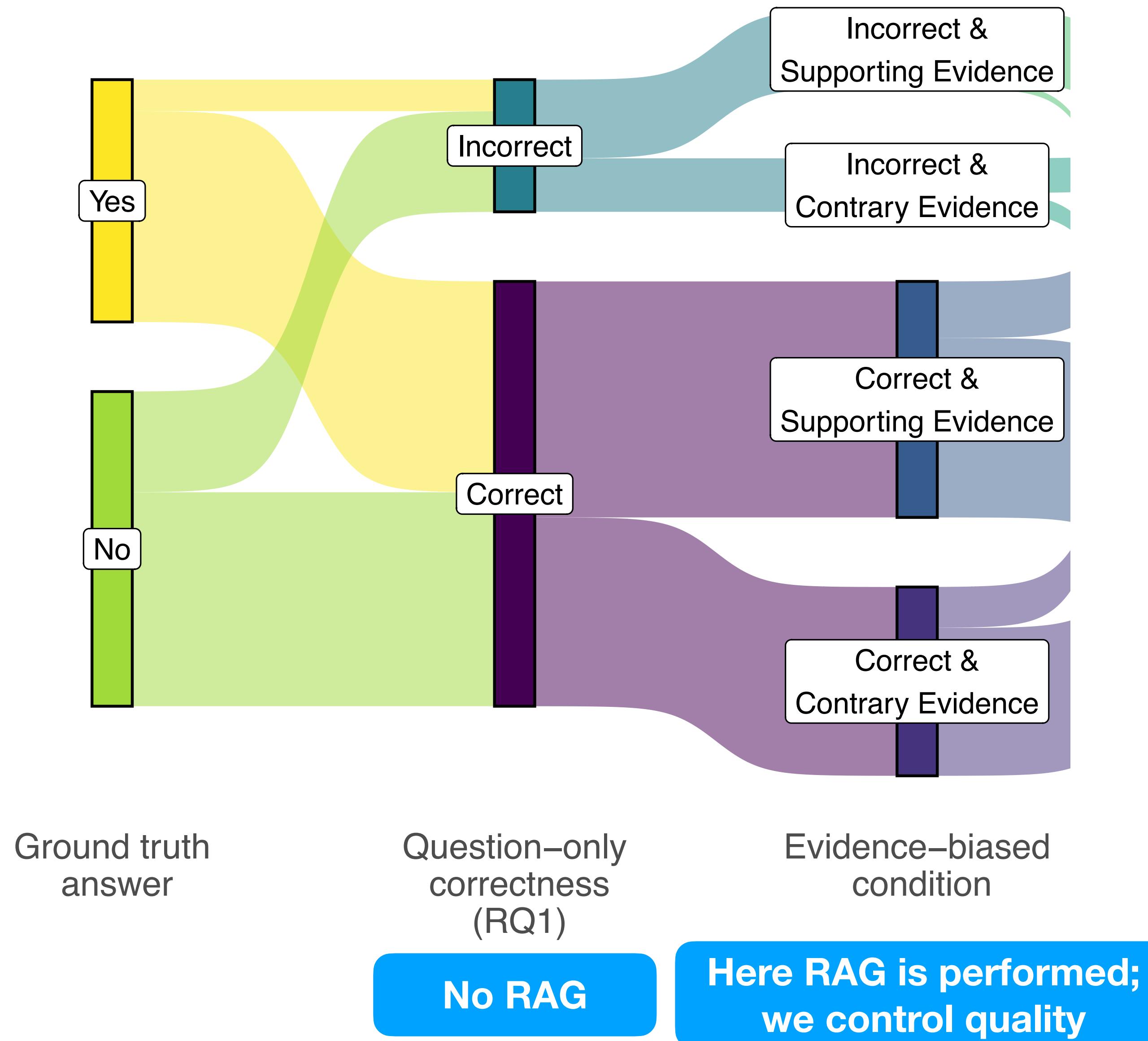
Ground truth
answer

Question-only
correctness
(RQ1)

No RAG

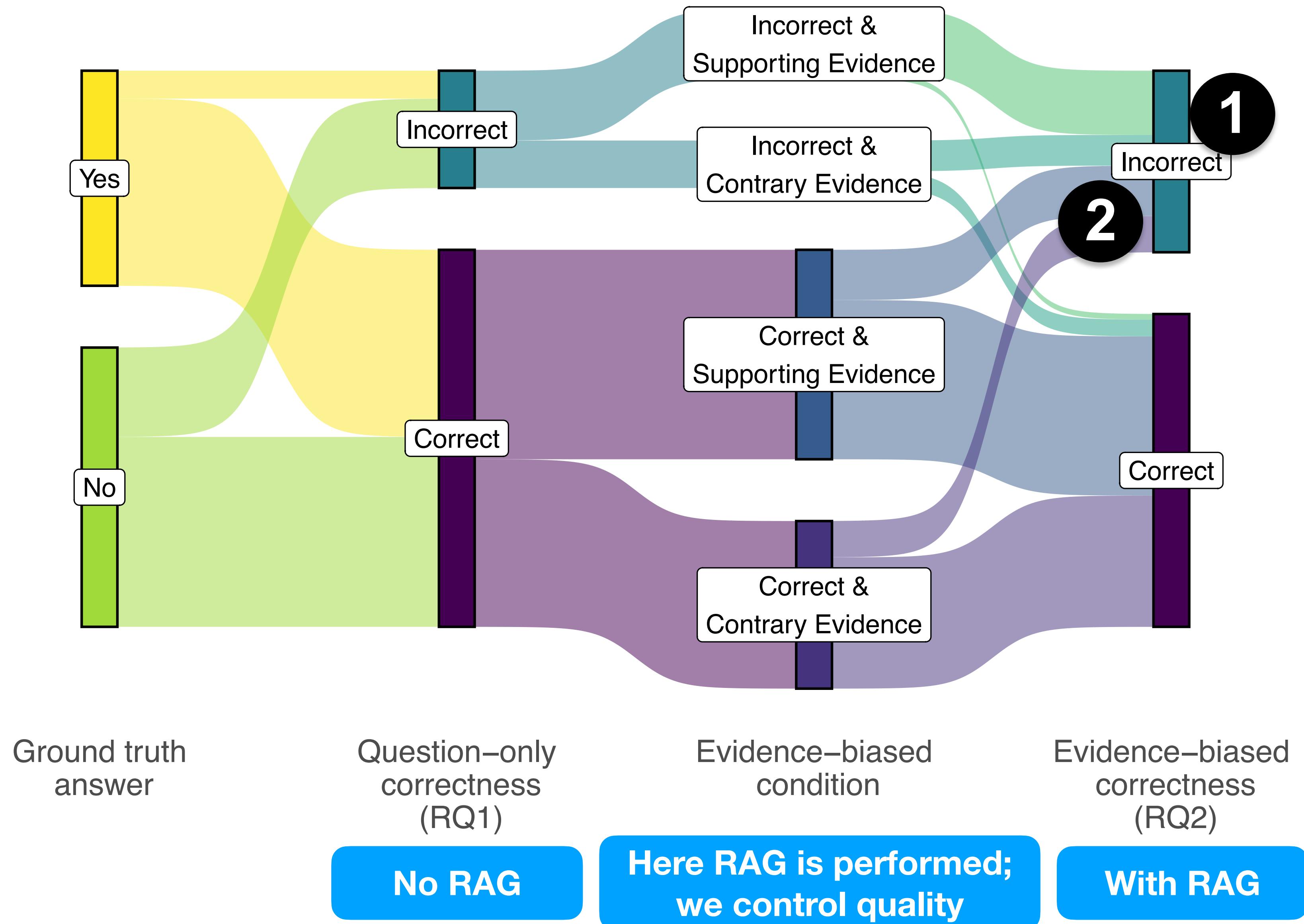
Zuccon, G. and Koopman, B., 2023. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*. (EMNLP2023)

RAG is not the Silver Bullet for Solving Hallucinations



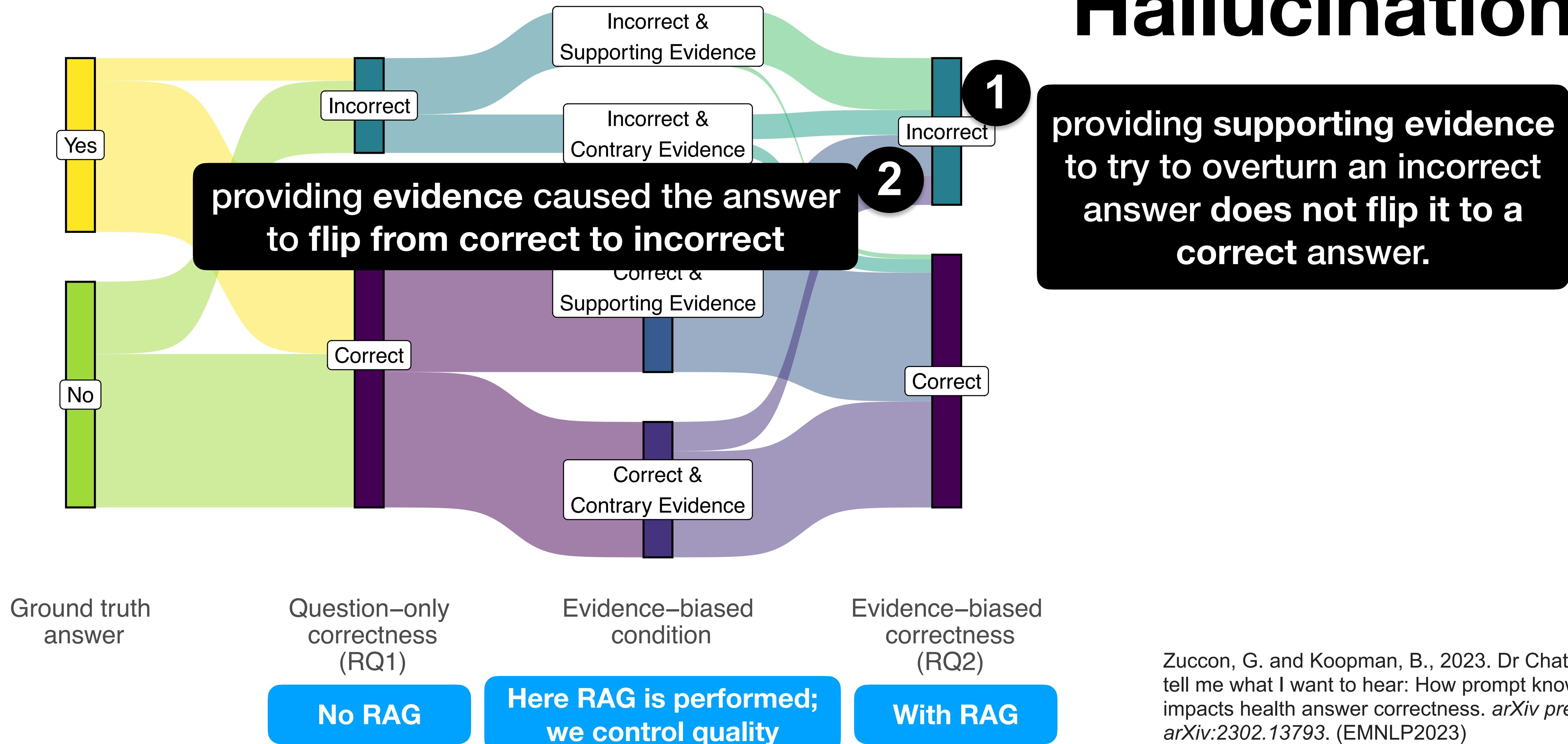
Zuccon, G. and Koopman, B., 2023. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*. (EMNLP2023)

RAG is not the Silver Bullet for Solving Hallucinations



Zuccon, G. and Koopman, B., 2023. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*. (EMNLP2023)

RAG is not the Silver Bullet for Solving Hallucinations



Zuccon, G. and Koopman, B., 2023. Dr ChatGPT, tell me what I want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*. (EMNLP2023)

Takeaway from this part

- LLMs are trained to **predict the next word** in a sequence (“Stochastic parrots”)
- In-context instructions/few-shots examples are often key to instruct the LLM
- LLMs suffer from **hallucinations**, so care should be taken. RAG is not necessarily a panacea to hallucinations

HOW POSTDOCS USE AI CHATBOTS

A little less than one-third of the postdoctoral researchers polled said that they use artificial-intelligence (AI) chatbots, such as ChatGPT, for everything from translating text to fixing code and overcoming writer's block.

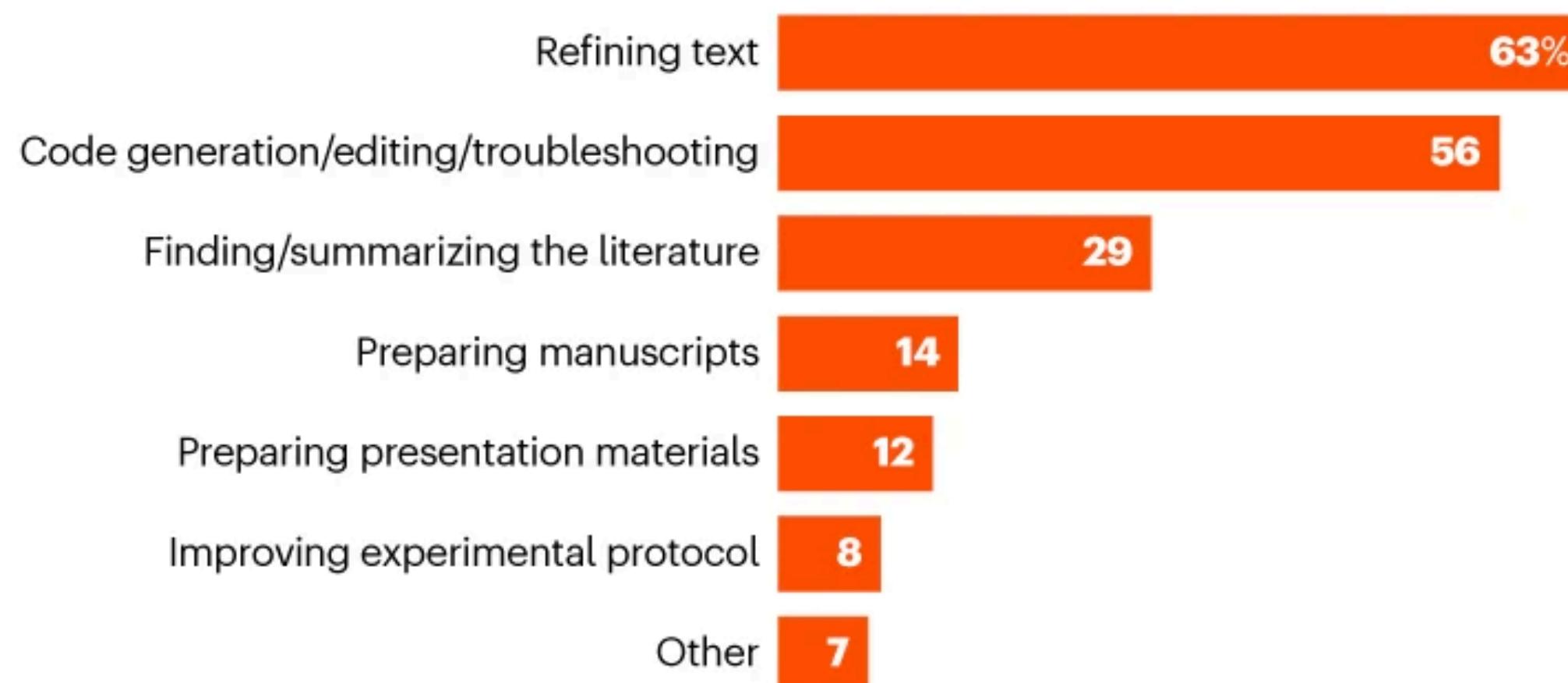
How has the evolution and rapid adoption of AI chatbots changed your day-to-day work or career plans, if at all?



Do you use AI chatbots, such as ChatGPT, in your work?



What do you use AI chatbots for?



How often do you use AI chatbots in your work?*



*Percentages don't add up to 100 owing to rounding.

How can LLMs help with your research?

Linda Nordling, 2023. How ChatGPT is transforming the postdoc experience. *Nature*, 622, p.655.

HOW POSTDOCS USE AI CHATBOTS

A little less than one-third of the postdoctoral researchers polled said that they use artificial-intelligence (AI) chatbots, such as ChatGPT, for everything from translating text to fixing code and overcoming writer's block.

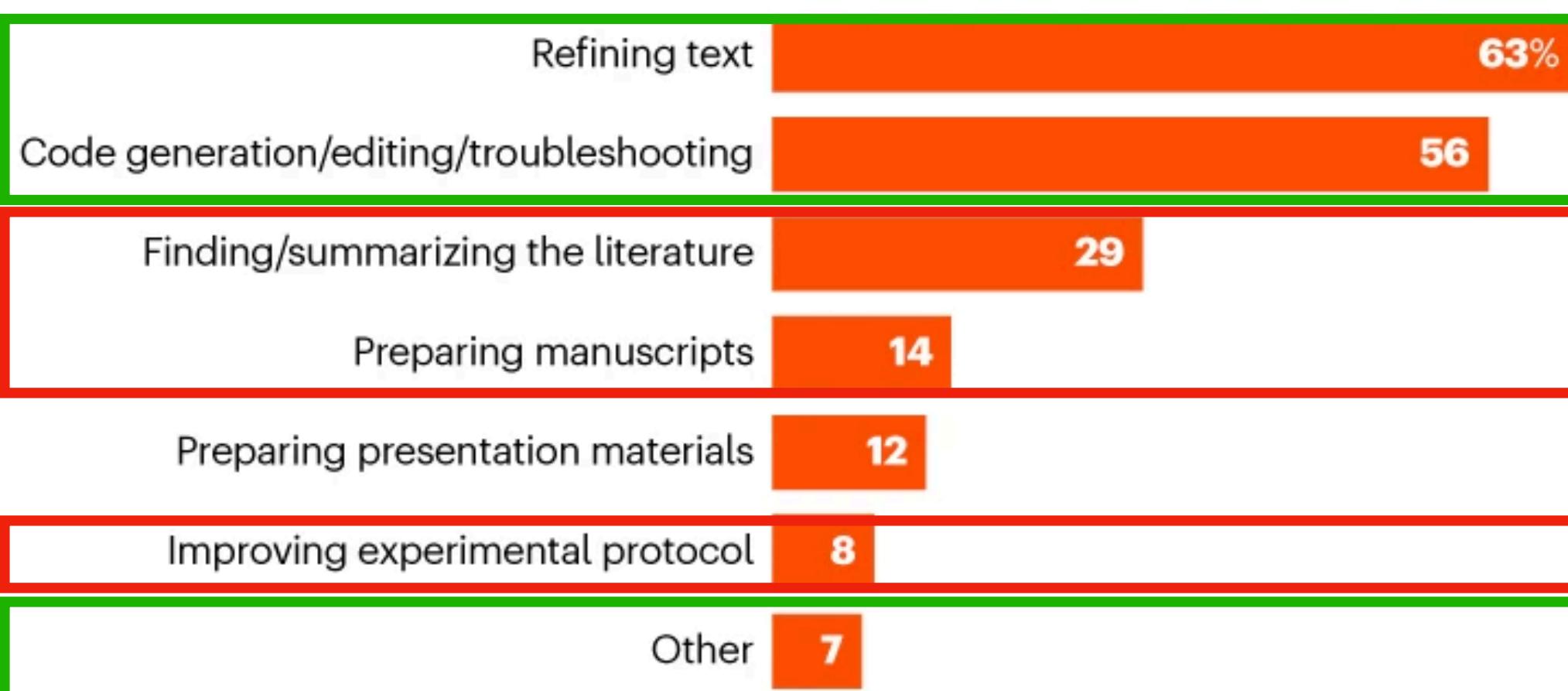
How has the evolution and rapid adoption of AI chatbots changed your day-to-day work or career plans, if at all?



Do you use AI chatbots, such as ChatGPT, in your work?



What do you use AI chatbots for?



How often do you use AI chatbots in your work?*



*Percentages don't add up to 100 owing to rounding.

How can LLMs help with your research?

Linda Nordling, 2023. How ChatGPT is transforming the postdoc experience. *Nature*, 622, p.655.

How can LLMs help with your research?

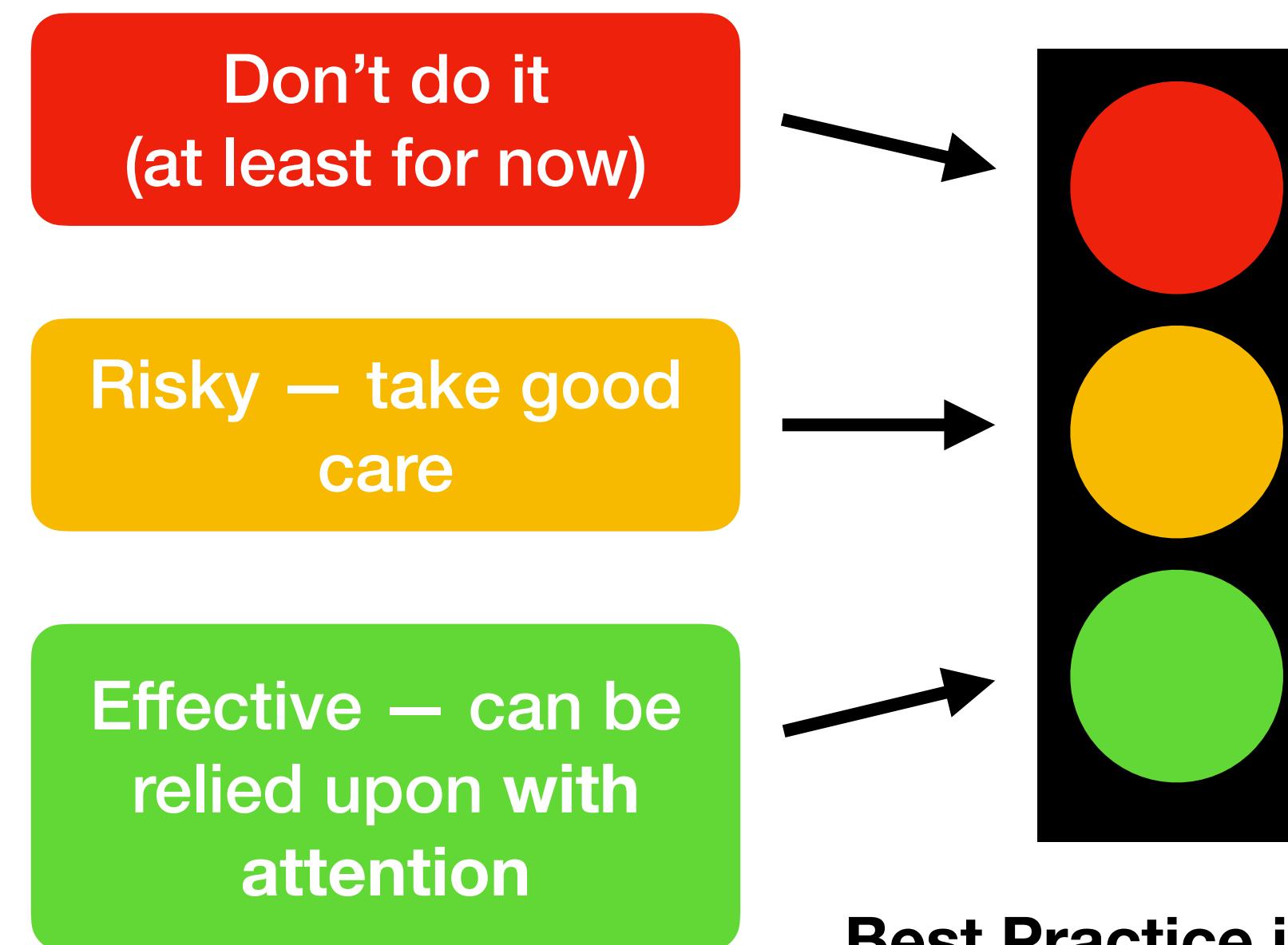
In this slides we will go through some examples

- Help improving your **writing**
- Assist with the ideation of **titles**
- Help with **reading** papers (or, “chat with papers”)
- Help you **design** experiments
- Generate data/**labels**
- **Customise** it to **your task** through prompt engineering/in-context learning
 - Extract information from text
 - Become a ranker
 - Help writing **code**

How can LLMs help with your research?

In this slides we will go through some examples

- Help improving your **writing**
- Assist with the ideation of **titles**
- Help with **reading** papers (or, “chat with papers”)
- Help you **design** experiments
- Generate data/**labels**
- **Customise** it to **your task** through prompt engineering/in-context learning
 - Extract information from text
 - Become a ranker
 - Help writing **code**



**Best Practice in using
LLMs for research**

Refine your writing

- Take the abstract of a paper you are writing (or have written)
- Open ChatGPT (<https://chat.openai.com/>), type the following prompt, and execute:

You are a useful scientific editor, and your job is to improve the writing of academic papers to make them more readable. Rewrite the following abstract:

<PASTE ABSTRACT>

Refine your writing

- Take the abstract of a paper you are writing (or have written)
- Open ChatGPT (<https://chat.openai.com/>), type the following prompt, and execute:

You are a useful scientific editor, and your job is to improve the writing of academic papers to make them more readable. Rewrite the following abstract:

<PASTE ABSTRACT>

Role playing: it has been found to often improve effectiveness

No in-prompt learning here — why?

Refine your writing

- Have a look at the text that ChatGPT has generated: what do you observe?
- ChatGPT is often a **very good editor**. However, pay attention:
 - It sometimes uses very “**bombastic**” words (watermarking?)
 - At times, it **changes the meaning** of your sentence
 - It is **best used in an iterative way**: do not accept everything ChatGPT writes; instead take it as a suggestion that you then further refine, mix and adapt
 - Always **check policies** regarding generative AI from publisher and your institution

Refine your writing



- Have a look at the text that ChatGPT has generated: what do you observe?
- ChatGPT is often a **very good editor**. However, pay attention:
 - It sometimes uses very "**bombastic**" words (watermarking?)
 - At times, it **changes the meaning** of your sentence
 - It is **best used in an iterative way**: do not accept everything ChatGPT writes; instead take it as a suggestion that you then further refine, mix and adapt
 - Always **check policies** regarding generative AI from publisher and your institution

ACM Policy on Authorship

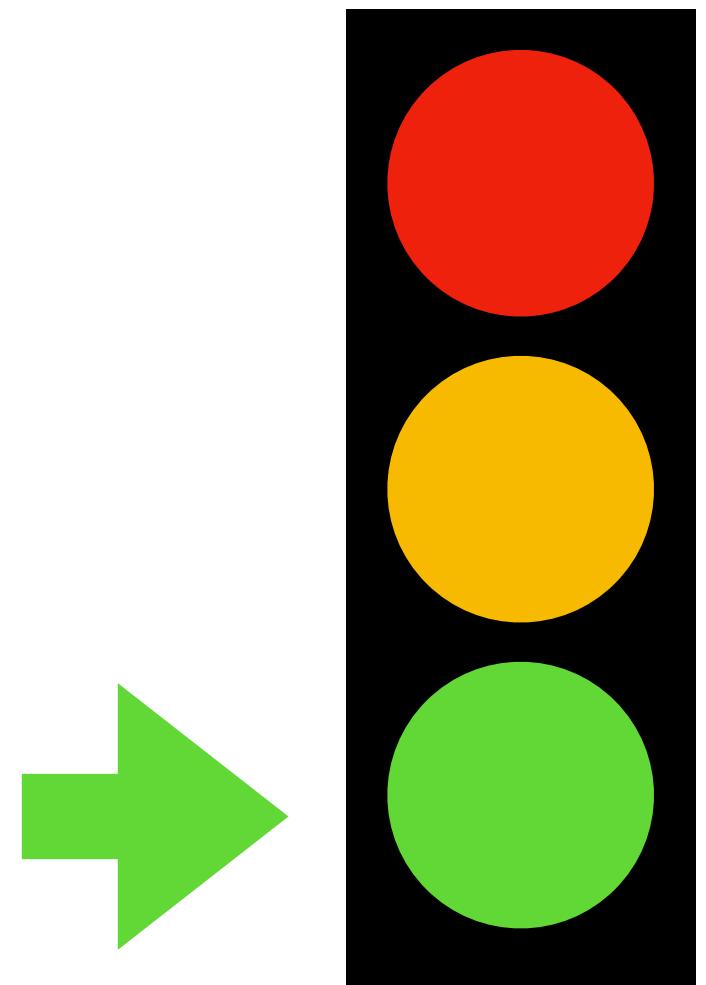
Can I use generative AI software tools to prepare my manuscript?

Yes, the use of such tools is permitted, subject to the following requirements:

- That these systems do not plagiarize, misrepresent, or falsify content in ACM submissions.
- That the resulting Work in its totality is an accurate representation of the authors' underlying work and novel intellectual contributions and is not primarily the result of the tool's generative capabilities.
- That the authors accept responsibility for the veracity and correctness of all material in their Work, including any computer-generated material.

Refine your writing

- Have a look at the text that ChatGPT has generated: what do you observe?
- ChatGPT is often a **very good editor**. However, pay attention:
 - It sometimes uses very “**bombastic**” words (watermarking?)
 - At times, it **changes the meaning** of your sentence
 - It is **best used in an iterative way**: do not accept everything ChatGPT writes; instead take it as a suggestion that you then further refine, mix and adapt
 - Always **check policies** regarding generative AI from publisher and your institution



Can you suggest me a good title?

- Take the previous abstract
- Go to: <https://huggingface.co/jelabgroup/BiTAG-t5-large>
- Paste the abstract in the box, click Compute



Hugging Face

The screenshot shows the Hugging Face Model Card interface for the BiTAG-t5-large model. At the top, it displays the repository name "ArvinZhuang/BiTAG-t5-large" with 0 likes. Below this are tabs for "Model card", "Files and versions", and "Community". On the right, there are buttons for "Edit model card", "Train", "Deploy", and "Use in Transformers". A chart titled "Downloads last month" shows a single peak at 4. The "Hosted inference API" section includes a "Text2Text Generation" example and a "BERT abst..." dropdown. The main area contains a code snippet for generating text from an abstract using the transformers library. To the right of the code, a large box highlights the abstract text: "abstract: We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement)."/>

```
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer\n\nmodel = AutoModelForSeq2SeqLM.from_pretrained("ArvinZhuang/BiTAG-t5-large")\ntokenizer = AutoTokenizer.from_pretrained("ArvinZhuang/BiTAG-t5-large")\n\ntext = "abstract: [your abstract]" # use 'title:' as the prefix for the abstract\ninput_ids = tokenizer.encode(text, return_tensors='pt')\n\noutputs = model.generate(\n    input_ids,\n    do_sample=True,\n    max_length=500,\n    top_p=0.9,\n    top_k=20,\n    temperature=1,\n    num_return_sequences=10,\n)\n\nprint("Output:\n" + 100 * '-')\nfor i, output in enumerate(outputs):\n    print("{}: {}".format(i+1, tokenizer.decode(output, skip_special_tokens=True)))
```

Can you suggest me a good title?

- Take the previous abstract
- Go to: <https://huggingface.co/ielabgroup/BiTAG-t5-large>
- Paste the abstract in the box, click Compute



Hugging Face

Compute ⌘+Enter 0.0

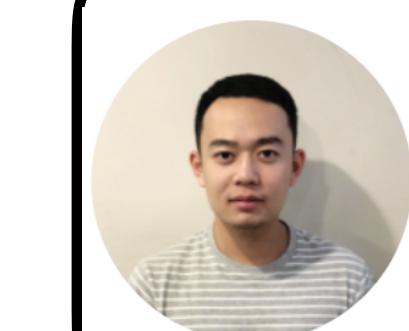
Computation time on Intel Xeon 3rd Gen Scalable cpu: 4.859 s

Bidirectional Encoder Representations for Natural Language Processing

JSON Output Maximize

```
[{"generated_text": "Bidirectional Encoder Representations for N"}, {"generated_text": "BERT: Bidirectional Encoder Representations"}, {"generated_text": "Bidirectional Encoder Representations from"}, {"generated_text": "BERT: Bidirectional Encoder Representations"}, {"generated_text": "Bidirectional Encoder Representations from"}, {"generated_text": "BERT: Bidirectional Encoder Representations"}, {"generated_text": "BERT: A Bidirectional Encoder Representatio"}, {"generated_text": "BERT: Bidirectional Encoder Representations"}, {"generated_text": "BERT: Bidirectional Encoder Representations"}, {"generated_text": "BERT: Bidirectional Encoder Representations"}]
```

- Model we developed over 2 years ago
- We used this extensively internally to my team
- T5 model (encoder/decoder) trained on a large dump of arXiv papers
- ChatGPT can also be used for this



**Shengyao “Arvin”
Zhuang**
Now PostDoc at CSIRO

Can you suggest me a good title?

- Take the previous abstract
- Go to: <https://huggingface.co/ielabgroup/BiTAG-t5-large>
- Paste the abstract in the box, click Compute



Hugging Face

ie
lab

Compute ⌘+Enter 0.0

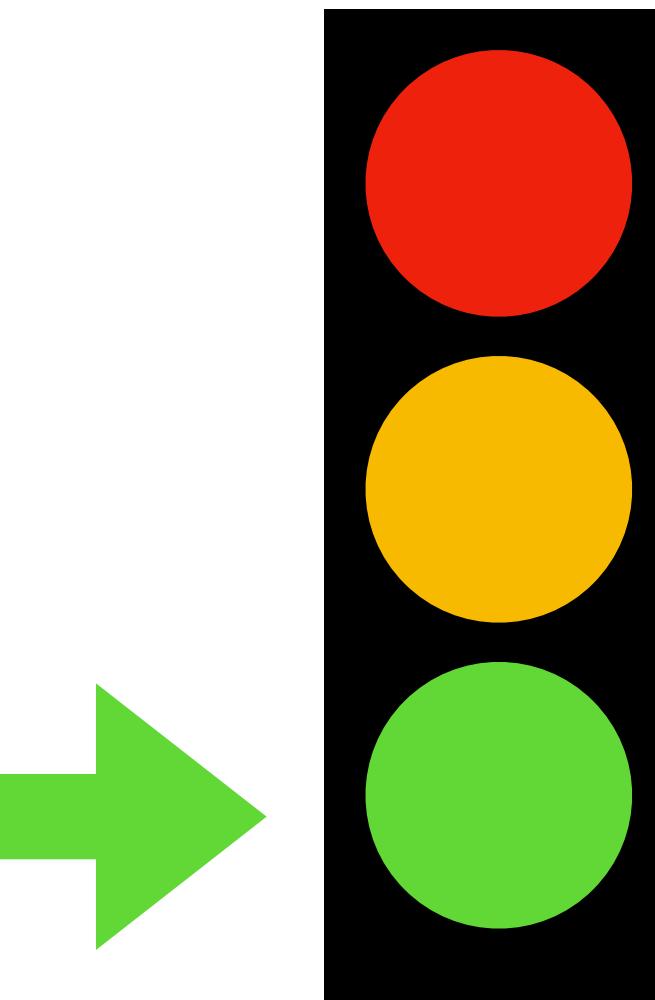
Computation time on Intel Xeon 3rd Gen Scalable cpu: 4.859 s

Bidirectional Encoder Representations for Natural Language Processing

JSON Output Maximize

```
[  
  {  
    "generated_text": "Bidirectional Encoder Representations for N  
  },  
  {  
    "generated_text": "BERT: Bidirectional Encoder Representations  
  },  
  {  
    "generated_text": "Bidirectional Encoder Representations from  
  },  
  {  
    "generated_text": "BERT: Bidirectional Encoder Representations  
  },  
  {  
    "generated_text": "Bidirectional Encoder Representations from  
  },  
  {  
    "generated_text": "BERT: Bidirectional Encoder Representations  
  },  
  {  
    "generated_text": "BERT: A Bidirectional Encoder Representatio  
  },  
  {  
    "generated_text": "BERT: Bidirectional Encoder Representations  
  },  
  {  
    "generated_text": "BERT: Bidirectional Encoder Representations  
  },  
  {  
    "generated_text": "BERT: Bidirectional Encoder Representations  
  }]
```

- As for ChatGPT for writing, use this for inspiration
- The best titles we came up with were inspired, not copied, from this tool



Have problems reading a paper? Chat with papers (TXYZ, a.k.a. arXiw.org)

The screenshot shows the TXYZ web interface. On the left, there's a sidebar with the TXYZ logo, a 'Beta' status indicator, a 'Daily Digest' button (which is highlighted in blue), and a 'My Documents' button. A large blue button at the bottom of the sidebar contains the URL <https://txyz.ai/>. The main content area has a header that reads "Personalized paper feed tailored to your interests." with buttons for "Information Retrieval" and "+ Add Subjects". Below this is a search bar with the placeholder "Add keywords to filter...". The first paper listed is "The Word2vec Graph Model for Author Attribution and Genre Detection in Literary Analysis" by Nafis Irtiza Tripto, Mohammed Eunus Ali, from Arxiv. The abstract discusses the limitations of traditional feature sets and proposes a novel Word2vec graph based modeling for author attribution and genre detection. The second paper listed is "On Surgical Fine-tuning for Language Encoders" by Abhilasha Lodha, Gayatri Belapurkar, Saloni Chalkapurkar, Yuanming Tao, Reshma Ghosh, Samyadeep Basu, Dmitrii Petrov, and Soundararajan Srinivasan, also from Arxiv. This paper explores fine-tuning strategies for language encoders, proposing a metric based on the Fisher information matrix (FIM score) to select candidate layers for selective fine-tuning.

Personalized paper feed tailored to your interests. [Information Retrieval](#) + Add Subjects

Add keywords to filter...

The Word2vec Graph Model for Author Attribution and Genre Detection in Literary Analysis

Arxiv Nafis Irtiza Tripto, Mohammed Eunus Ali

Analyzing the writing styles of authors and articles is a key to supporting various literary analyses such as author attribution and genre detection. Over the years, rich sets of features that include stylometry, bag-of-words, n-grams have been widely used to perform such analysis. However, the effectiveness of these features largely depends on the linguistic aspects of a particular language and datasets specific characteristics. Consequently, techniques based on these feature sets cannot give desired results across domains. In this paper, we propose a novel Word2vec graph based modeling of a document that can rightly capture both context and style of the document. By using these Word2vec graph based features, we perform classification to perform author attribution and genre detection tasks. Our detailed experimental study with a comprehensive set of literary writings shows the effectiveness of this method over traditional feature based approaches. Our code and data are publicly available at <https://cutt.ly/svLjSgk>

On Surgical Fine-tuning for Language Encoders

Arxiv Abhilasha Lodha, Gayatri Belapurkar, Saloni Chalkapurkar, Yuanming Tao, Reshma Ghosh, Samyadeep Basu, Dmitrii Petrov, Soundararajan Srinivasan

Fine-tuning all the layers of a pre-trained neural language encoder (either using all the parameters or using parameter-efficient methods) is often the de-facto way of adapting it to a new task. We show evidence that for different downstream language tasks, fine-tuning only a subset of layers is sufficient to obtain performance that is close to and often better than fine-tuning all the layers in the language encoder. We propose an efficient metric based on the diagonal of the Fisher information matrix (FIM score), to select the candidate layers for selective

[Contact Us](#)

Have problems reading a paper? Chat with papers (TXYZ, a.k.a. arXiw.org)

- Go to an arXiv paper you want to read (or have recently read), e.g.:
- <https://arxiv.org/abs/2310.09497>
- Next, In the URL, **change v->w**, and go to that page
- <https://arxiw.org/abs/2310.09497>
- Start chatting! Ask questions about the paper

Can you explain me the pointwise ranking approach?

The pointwise ranking approach is one of the three main prompting approaches for zero-shot document ranking with LLMs. It prompts individual documents individually. There are

thanks. What about pairwise?

The pairwise approach is another prompting approach for zero-shot document ranking with LLMs. In this approach, the LLM is prompted to compare pairs of documents to determine which one is more relevant to the query.

can you make the setwise approach even better?

Based on the information provided, it seems that the Setwise approach proposed in the paper is already designed to enhance the efficiency of the Pairwise method for zero-shot document ranking with LLMs. The Setwise prompting approach instructs LLMs to compare the relevance of

Have problems reading a paper? Chat with papers (TXYZ, a.k.a. arXiw.org)

- It can even give you ideas to extend the paper!
- Though these are the ones we mentioned in the paper...
- Be careful about hallucinations, both in explanations and in summarisations

To make the Setwise approach even better, you could consider the following:

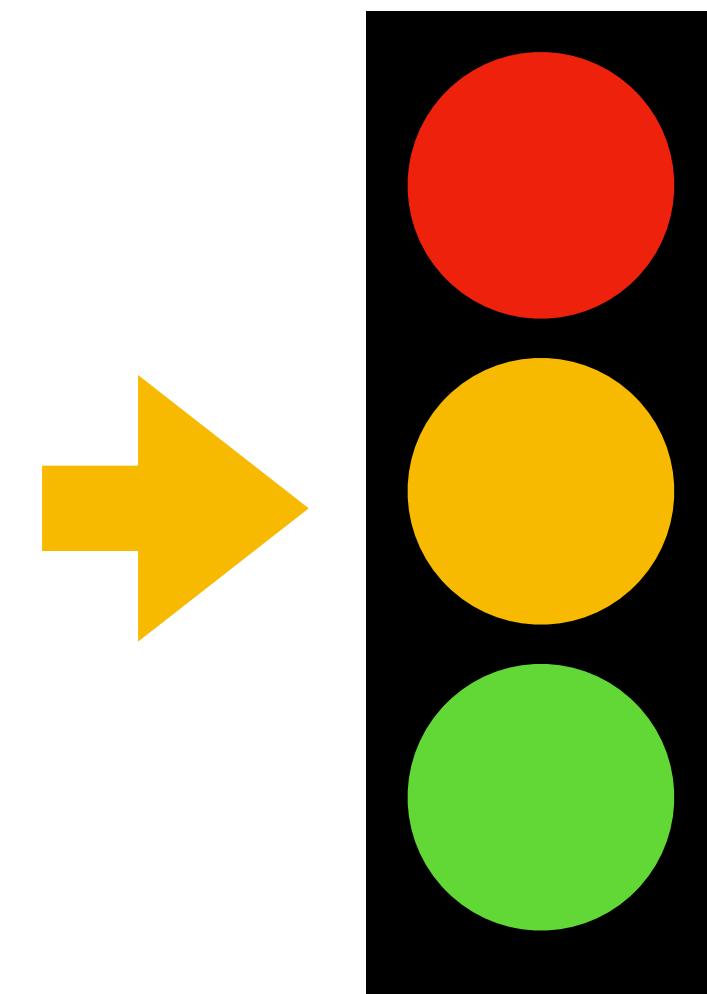
1. Experiment with different values of the parameter c : In the paper, the authors set c to 3, 5, 7, and 9, which determines the number of documents compared at each step. You can try different values of c to find the optimal balance between efficiency and effectiveness for your specific use case.
2. Explore different sorting algorithms: The paper mentions the use of heap sort and bubble sort algorithms in the Pairwise approach. However, there are other sorting algorithms available that may offer better performance in terms of efficiency. You can experiment with different sorting algorithms to see if they further improve the efficiency of the Setwise approach.
3. Consider the impact of document truncation: In the experiments described in the paper, document truncation was performed to ensure the prompt size is not exceeded. However, truncating documents may impact the effectiveness of the ranking. If your LLM has extended input length capacity, you can explore

Have problems reading a paper? Chat with papers (TXYZ, a.k.a. [arXiw.org](#))

- It can even give you ideas to extend the paper!
- Though these are the ones we mentioned in the paper...
- Be careful about hallucinations, both in explanations and in summarisations

To make the Setwise approach even better, you could consider the following:

1. Experiment with different values of the parameter c : In the paper, the authors set c to 3, 5, 7, and 9, which determines the number of documents compared at each step. You can try different values of c to find the optimal balance between efficiency and effectiveness for your specific use case.
2. Explore different sorting algorithms: The paper mentions the use of heap sort and bubble sort algorithms in the Pairwise approach. However, there are other sorting algorithms available that may offer better performance in terms of efficiency. You can experiment with different sorting algorithms to see if they further improve the efficiency of the Setwise approach.
3. Consider the impact of document truncation: In the experiments described in the paper, document truncation was performed to ensure the prompt size is not exceeded. However, truncating documents may impact the effectiveness of the ranking. If your LLM has extended input length capacity, you can explore



HOW POSTDOCS USE AI CHATBOTS

A little less than one-third of the postdoctoral researchers polled said that they use artificial-intelligence (AI) chatbots, such as ChatGPT, for everything from translating text to fixing code and overcoming writer's block.

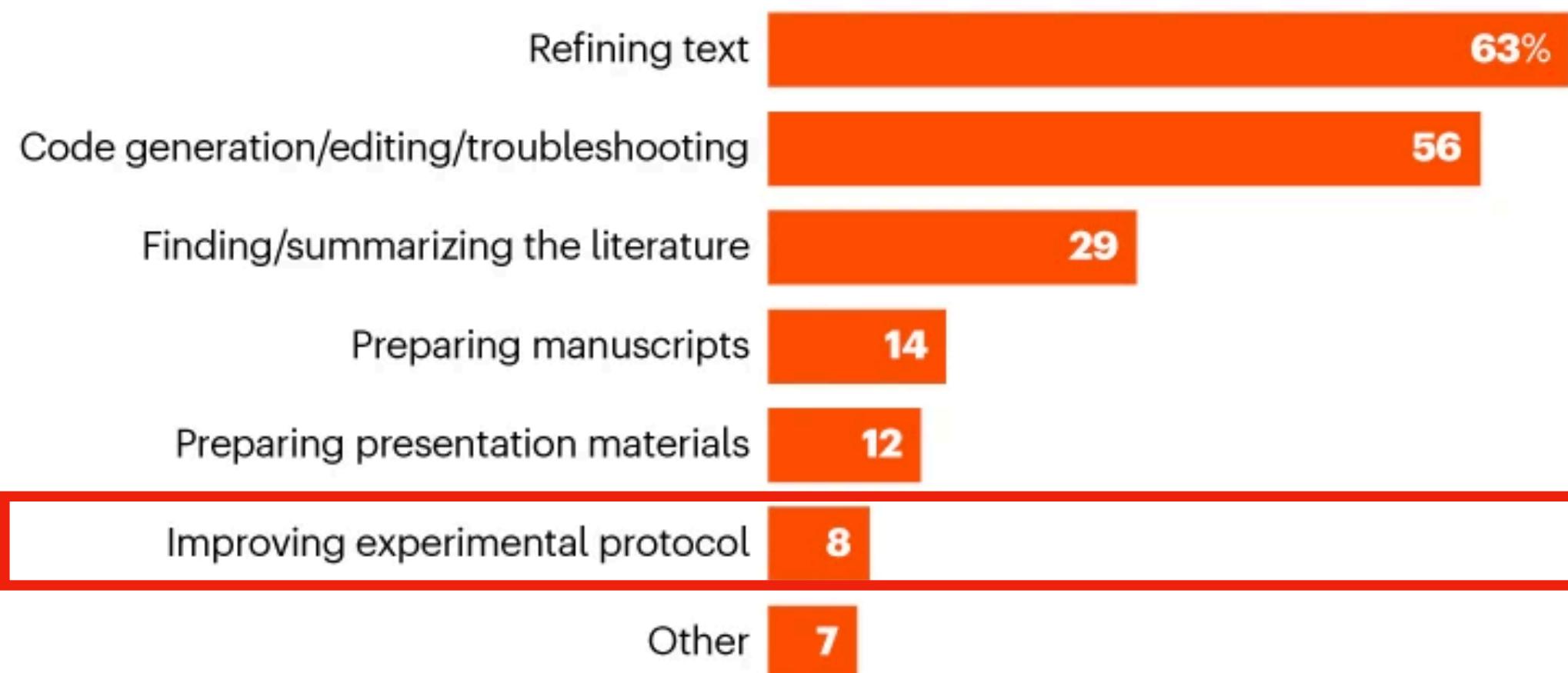
How has the evolution and rapid adoption of AI chatbots changed your day-to-day work or career plans, if at all?



Do you use AI chatbots, such as ChatGPT, in your work?



What do you use AI chatbots for?



How often do you use AI chatbots in your work?*



*Percentages don't add up to 100 owing to rounding.

Use ChatGPT to design an experiment

- Sounds like a great idea!
- Experiment design methodologies are quite stable (it's not like a new methodology is proposed every day)
- Plenty of examples of usage of established methodologies
- Often you know a specific methodology, but are not well versed in the gamut of methodologies available out there

Use ChatGPT to design an experiment

⊕ GPT-4

G

I need to design a user study in which I study 3 different interfaces (A, B, C) across 9 topics (1, 2, 3, 4, 5, 6, 7, 8, 9). Build a graeco-latin square design for my user study

Use ChatGPT to design an experiment



A Graeco-Latin square design is a method used to control for the effects of two variables (in your case, interface and topic) while studying the impact of a third. Since you have 3 interfaces (A, B, C) and 9 topics (1-9), you can create a 3×3 Graeco-Latin square for the interfaces and then replicate that design 3 times to cover all 9 topics.

Here's how you can set it up:

Step 1: Start with the Latin square for the topics:

Copy code

```
1 2 3  
3 1 2  
2 3 1
```

Step 2: Now, superimpose a Graeco square for the interfaces:

Copy code

```
css  
A B C  
B C A  
C A B
```

Step 3: Combine them. The first element from the Latin square (the topics) is paired with the first element from the Graeco square (the interfaces). For example, in the first cell, the topic is 1 and the interface is A, so the combined cell is (A1).

Here's the 3×3 Graeco-Latin Square:

Copy code

```
A1 B2 C3  
B3 C1 A2  
C2 A3 B1
```

To extend this for all 9 topics:

First Set (Topics 1-3):

Copy code

```
A1 B2 C3  
B3 C1 A2  
C2 A3 B1
```

Second Set (Topics 4-6):

Copy code

```
A4 B5 C6  
B6 C4 A5  
C5 A6 B4
```

Use ChatGPT to design an experiment

Third Set (Topics 7-9):

```
A7 B8 C9  
B9 C7 A8  
C8 A9 B7
```

 Copy code

Now, you have 9 groups where each interface (A, B, C) is used with each of the 9 topics exactly once. Users assigned to each of the 9 groups will interact with each interface, ensuring no combination is missed.

When you run your study, you can randomly assign users to each of the groups to ensure randomness. This design ensures that each interface is tested for each topic without any repeated combinations, while also controlling for order effects across both interfaces and topics.

Thank you ,
that is exactly
what I was
after!!!

Use ChatGPT to design an experiment

Third Set (Topics 7-9):

```
A7 B8 C9  
B9 C7 A8  
C8 A9 B7
```

 Copy code

Now, you have 9 groups where each interface (A, B, C) is used with each of the 9 topics exactly once. Users assigned to each of the 9 groups will interact with each interface, ensuring no combination is missed.

When you run your study, you can randomly assign users to each of the groups to ensure randomness. This design ensures that each interface is tested for each topic without any repeated combinations, while also controlling for order effects across both interfaces and topics.

Thank you ,
that is exactly
what I was
after!!!

Or is it?

Use ChatGPT to design an experiment

G

are you sure? I think there are needed 27 groups to remove possible order biases. Isn't it?

Use ChatGPT to design an experiment

G

are you sure? I think there are needed 27 groups to remove possible order biases. Isn't it?



You're right. I apologize for the oversight. If you want to control for order effects and ensure that no participant sees more than one interface for a given topic, then you would need to create 27 groups, given there are 3 possible orders in which the interfaces can be seen (ABC, BCA, CAB) and 9 topics.

For a complete Graeco-Latin square design, where you remove the potential biases for both the sequence of interfaces and the sequence of topics, you'd design the user study as follows:

For sequence ABC:

Use ChatGPT to design an experiment



You're right. I apologize for the oversight. If you want to control for order effects and ensure that no participant sees more than one interface for a given topic, then you would need to create 27 groups, given there are 3 possible orders in which the interfaces can be seen (ABC, BCA, CAB) and 9 topics.

For a complete Graeco-Latin square design, where you remove the potential biases for both the sequence of interfaces and the sequence of topics, you'd design the user study as follows:

For sequence ABC:

1. A1, B2, C3
2. A2, B3, C1
3. A3, B1, C2
4. A4, B5, C6
5. A5, B6, C4
6. A6, B4, C5
7. A7, B8, C9
8. A8, B9, C7
9. A9, B7, C8

For sequence BCA:

1. B1, C2, A3
2. B2, C3, A1
3. B3, C1, A2
4. B4, C5, A6
5. B5, C6, A4
6. B6, C4, A5
7. B7, C8, A9
8. B8, C9, A7
9. B9, C7, A8

For sequence CAB:

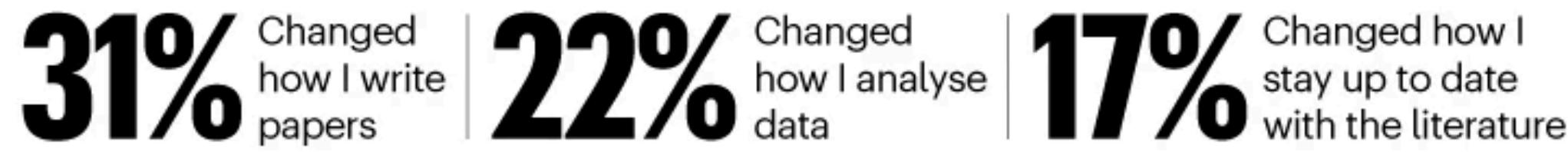
1. C1, A2, B3
2. C2, A3, B1
3. C3, A1, B2
4. C4, A5, B6
5. C5, A6, B4
6. C6, A4, B5
7. C7, A8, B9
8. C8, A9, B7
9. C9, A7, B8

Now, you would have 27 unique combinations. When assigning participants, each participant would be assigned to one of these 27 sequences. This way, you ensure that every topic is tested with every interface, and the sequence in which the interfaces are presented is also varied to control for any order effects.

HOW POSTDOCS USE AI CHATBOTS

A little less than one-third of the postdoctoral researchers polled said that they use artificial-intelligence (AI) chatbots, such as ChatGPT, for everything from translating text to fixing code and overcoming writer's block.

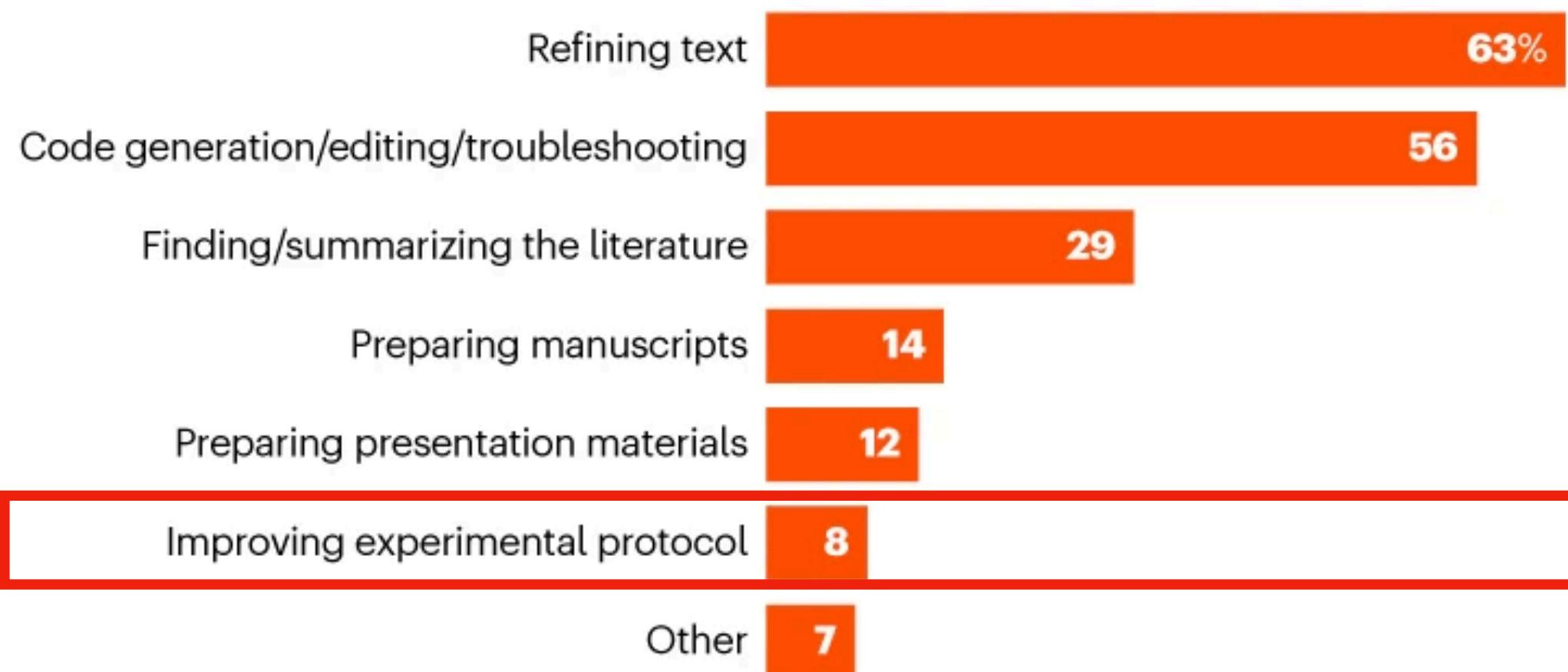
How has the evolution and rapid adoption of AI chatbots changed your day-to-day work or career plans, if at all?



Do you use AI chatbots, such as ChatGPT, in your work?



What do you use AI chatbots for?



How often do you use AI chatbots in your work?*



*Percentages don't add up to 100 owing to rounding.

Use ChatGPT to design an experiment

- I haven't seen a systematic analysis of whether ChatGPT can design good experiments
- Certainly it didn't from my experience
- What is dangerous is that the answers look very good:
 - Very good explanations
 - Reasonable process
- And even asking to confirm is not useful: often ChatGPT will admit it is wrong... for then giving you again a wrong answer

HOW POSTDOCS USE AI CHATBOTS

A little less than one-third of the postdoctoral researchers polled said that they use artificial-intelligence (AI) chatbots, such as ChatGPT, for everything from translating text to fixing code and overcoming writer's block.

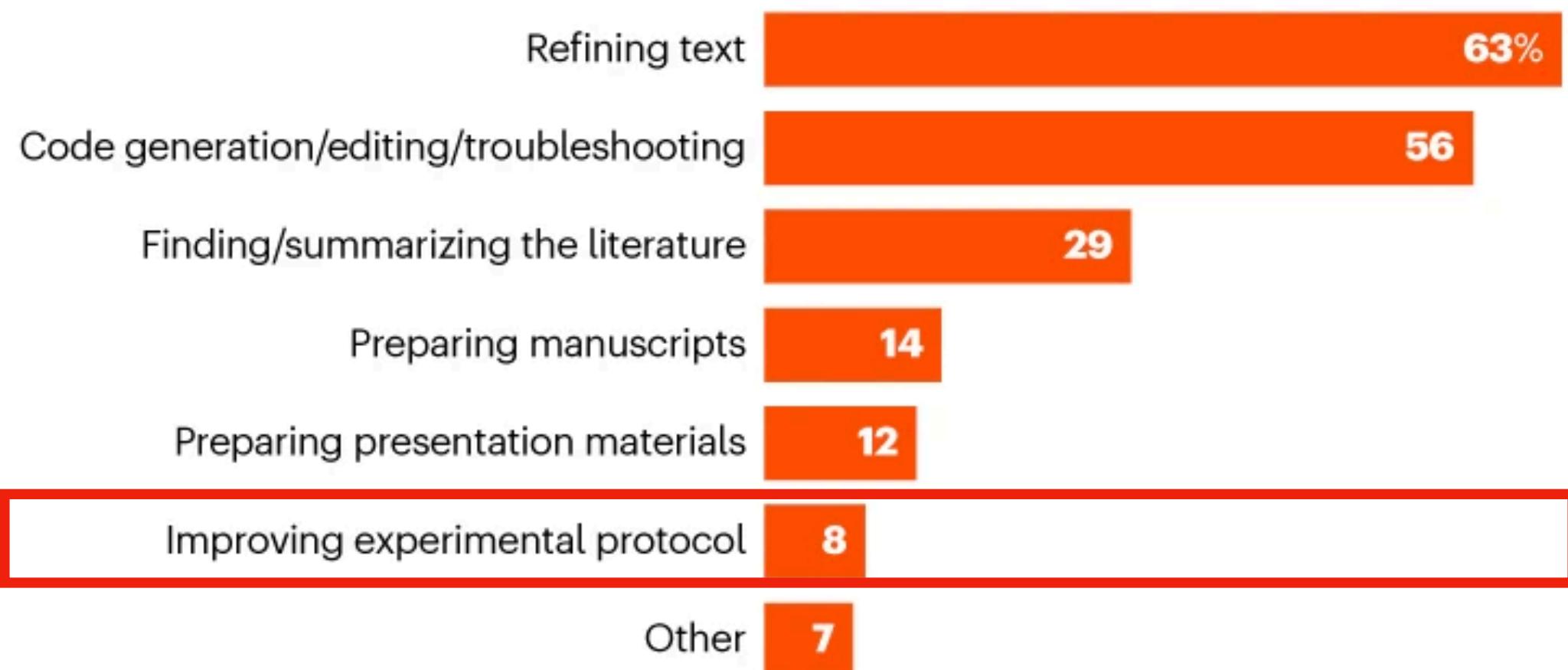
How has the evolution and rapid adoption of AI chatbots changed your day-to-day work or career plans, if at all?



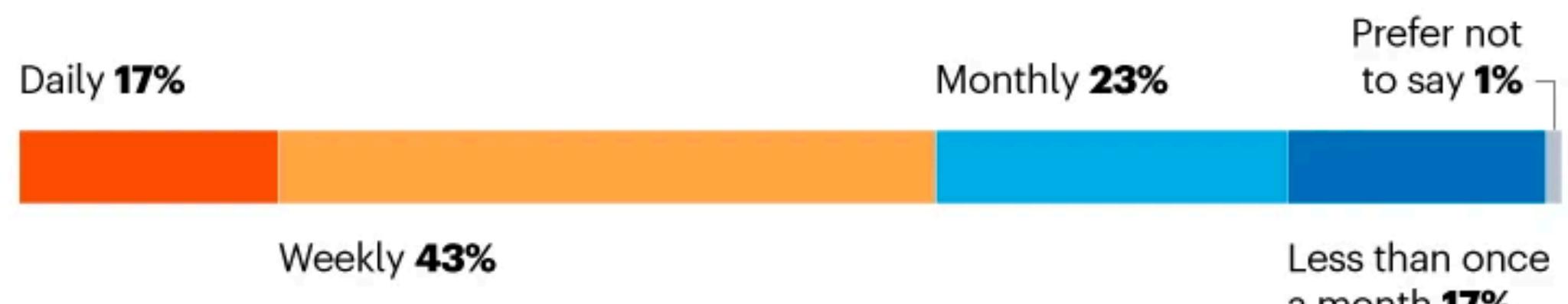
Do you use AI chatbots, such as ChatGPT, in your work?



What do you use AI chatbots for?



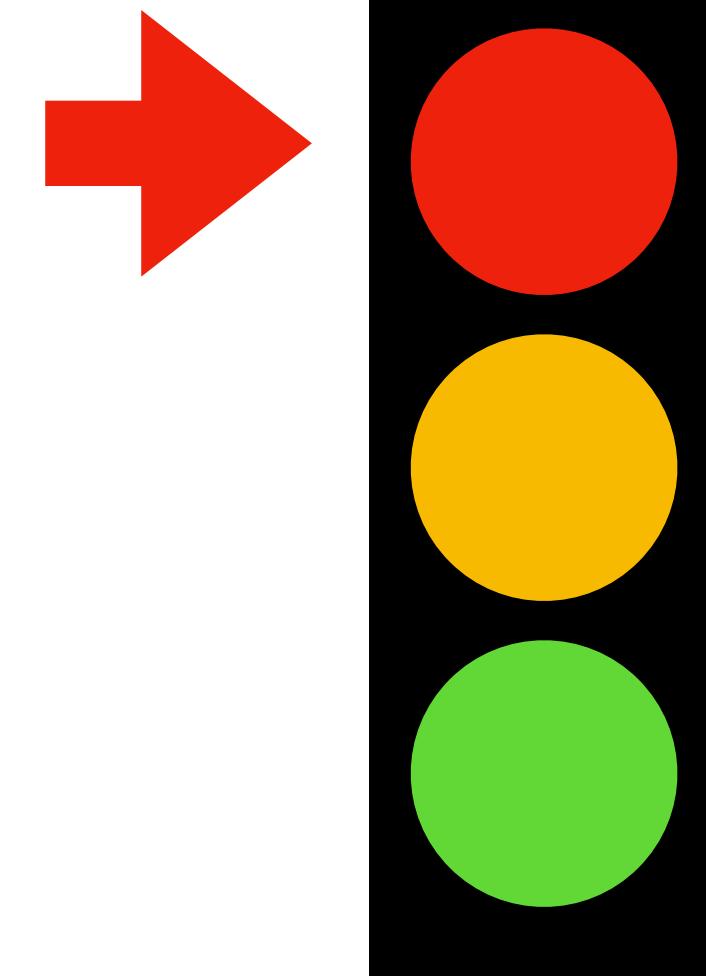
How often do you use AI chatbots in your work?*



*Percentages don't add up to 100 owing to rounding.

Use ChatGPT to design an experiment

- I haven't seen a systematic analysis of whether ChatGPT can design good experiments
- Certainly it didn't from my experience
- What is dangerous is that the answers look very good:
 - Very good explanations
 - Reasonable process
- And even asking to confirm is not useful: often ChatGPT will admit it is wrong... for then giving you again a wrong answer



HOW POSTDOCS USE AI CHATBOTS

A little less than one-third of the postdoctoral researchers polled said that they use artificial-intelligence (AI) chatbots, such as ChatGPT, for everything from translating text to fixing code and overcoming writer's block.

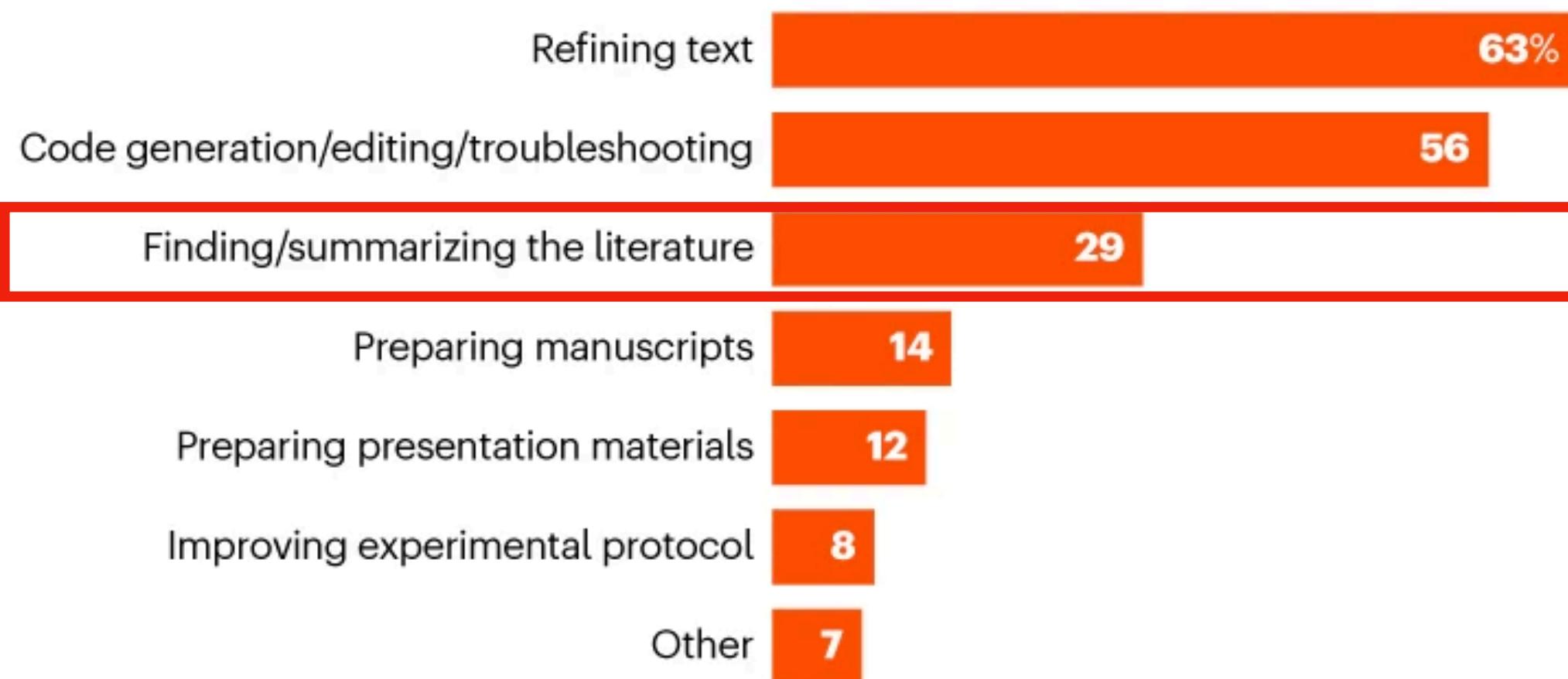
How has the evolution and rapid adoption of AI chatbots changed your day-to-day work or career plans, if at all?



Do you use AI chatbots, such as ChatGPT, in your work?



What do you use AI chatbots for?



How often do you use AI chatbots in your work?*



*Percentages don't add up to 100 owing to rounding.

Finding literature

- Anecdotally, I know a few students that have cited references provided by ChatGPT
- Why it's not a good idea?
 1. ChatGPT's model is static: it does not update "every day", so it will only know of studies contained in training data, up until the date it was trained on (likely, much before that)
 2. Remember, LLMs hallucinate, especially if not well trained to do the task at hand

Finding literature: “ChatGPT Hallucinates when Attributing Answers”

- Suggested references very good, at first look:
 - Article title very relevant to question
 - Journals/Conf exist, volume/issue aligned with year, even with page numbers
 - Authors exist (though not always in the field of article)
 - But, they only exist 14% of the time.

Zuccon, G., Koopman, B., Shaik, R. 2023. ChatGPT Hallucinates when Attributing Answers. *arXiv preprint arXiv:2302.13793*. (SIGIR-AP2023)

Question: “What type of herbicides are effective against sowthistle?”

Ground truth answer: “The addition of Balance to either Group D or Group K herbicides can provide good control of sowthistle. The addition of Flame, Group D, Balance or Group K to broadleaf active herbicides (Group C and Valour) are also effective.”

ChatGPT Output:

Herbicides that are effective against sowthistle include postemergent selective herbicides such as 2,4-D, dicamba, and glyphosate.

1. 2,4-D: A study conducted by R.L. Richardson and published in the journal "Weed Science" in 1983 (Vol. 31, Issue 4, pp. 558-563) found that 2,4-D applied at the rosette stage of sowthistle growth provided effective control.

2. Dicamba: A study by K.A. Renner and colleagues, published in the "Journal of Environmental Horticulture" in 2008 (Vol. 26, Issue 2, pp. 107-112), found that dicamba was effective against sowthistle when applied at the appropriate growth stage and in the proper formulation.

3. Glyphosate: According to a study by P.J. Tranel and J.D. Nalewaja, published in the journal "Weed Science" in 1998 (Vol. 46, Issue 4, pp. 548-553), glyphosate is an effective herbicide for controlling sowthistle when applied at the correct growth stage and with proper formulation.

Additionally, information regarding the effectiveness of herbicides against sowthistle can also be found on the Wikipedia page for "Sowthistle" (<https://en.wikipedia.org/wiki/Sowthistle>). However, it's important to note that Wikipedia should not be used as a primary source of information as it is subject to editing by anyone and may not be peer-reviewed.

The four references provided do not exist

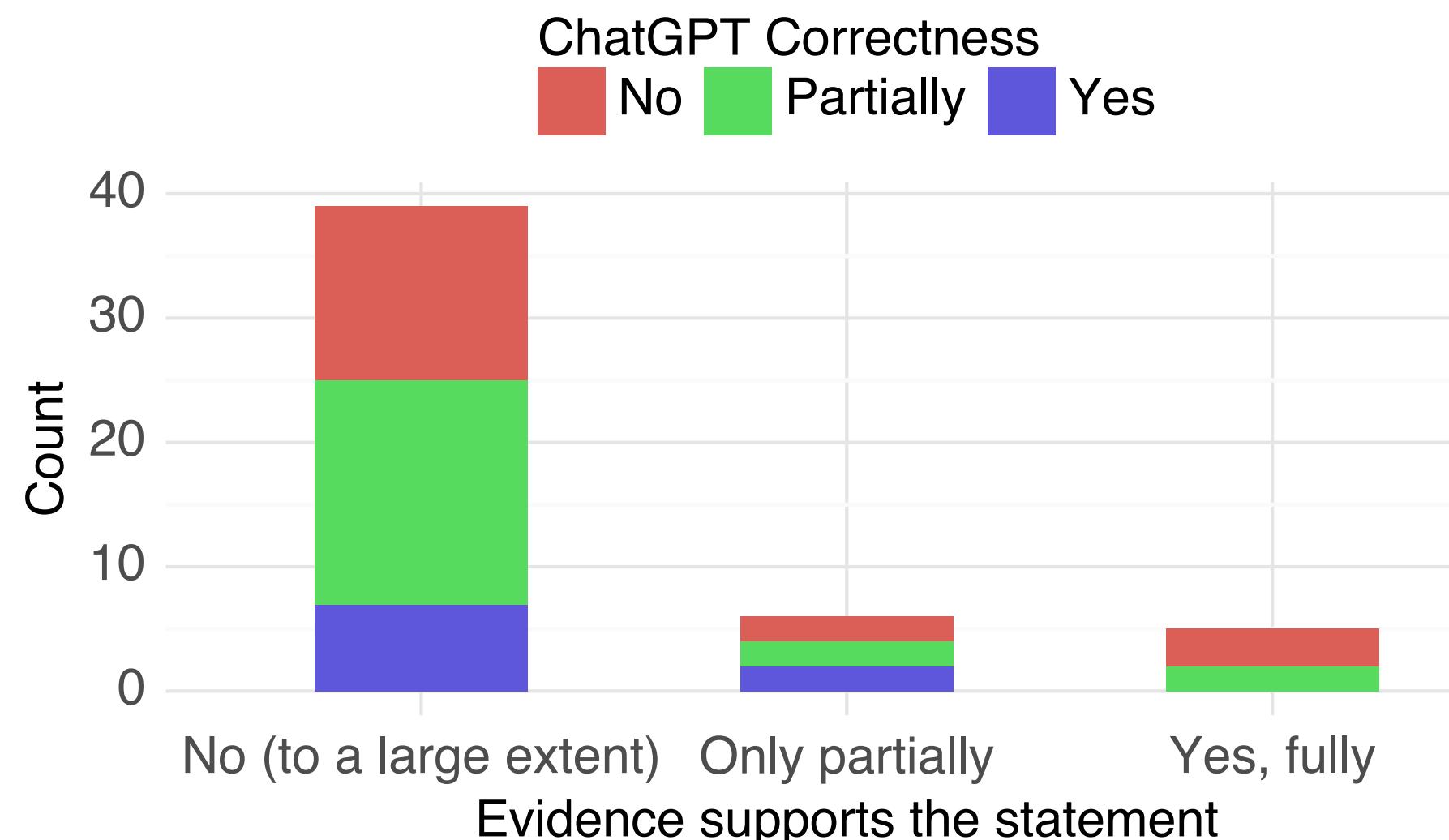
Statement about the reference

Invalid URL

Finding literature: “ChatGPT Hallucinates when Attributing Answers”

- Suggested references very good, at first look:
 - Article title very relevant to question
 - Journals/Conf exist, volume/issue aligned with year, even with page numbers
 - Authors exist (though not always in the field of article)
 - But, they only exist 14% of the time.

Zuccon, G., Koopman, B., Shaik, R. 2023. ChatGPT Hallucinates when Attributing Answers. *arXiv preprint arXiv:2302.13793*. (SIGIR-AP2023)



Question: “What type of herbicides are effective against sowthistle?”

Ground truth answer: “The addition of Balance to either Group D or Group K herbicides can provide good control of sowthistle. The addition of Flame, Group D, Balance or Group K to broadleaf active herbicides (Group C and Valour) are also effective.”

ChatGPT Output:

Herbicides that are effective against sowthistle include postemergent selective herbicides such as 2,4-D, dicamba, and glyphosate.

1. 2,4-D: A study conducted by R.L. Richardson and published in the journal "Weed Science" in 1983 (Vol. 31, Issue 4, pp. 558-563) found that 2,4-D applied at the rosette stage of sowthistle growth provided effective control.

2. Dicamba: A study by K.A. Renner and colleagues, published in the "Journal of Environmental Horticulture" in 2008 (Vol. 26, Issue 2, pp. 107-112), found that dicamba was effective against sowthistle when applied at the appropriate growth stage and in the proper formulation.

3. Glyphosate: According to a study by P.J. Tranel and J.D. Nalewaja, published in the journal "Weed Science" in 1998 (Vol. 46, Issue 4, pp. 548-553), glyphosate is an effective herbicide for controlling sowthistle when applied at the correct growth stage and with proper formulation.

Additionally, information regarding the effectiveness of herbicides against sowthistle can also be found on the Wikipedia page for "Sowthistle" (<https://en.wikipedia.org/wiki/Sowthistle>). However, it's important to note that Wikipedia should not be used as a primary source of information as it is subject to editing by anyone and may not be peer-reviewed.

Statement about the reference

Invalid URL

HOW POSTDOCS USE AI CHATBOTS

A little less than one-third of the postdoctoral researchers polled said that they use artificial-intelligence (AI) chatbots, such as ChatGPT, for everything from translating text to fixing code and overcoming writer's block.

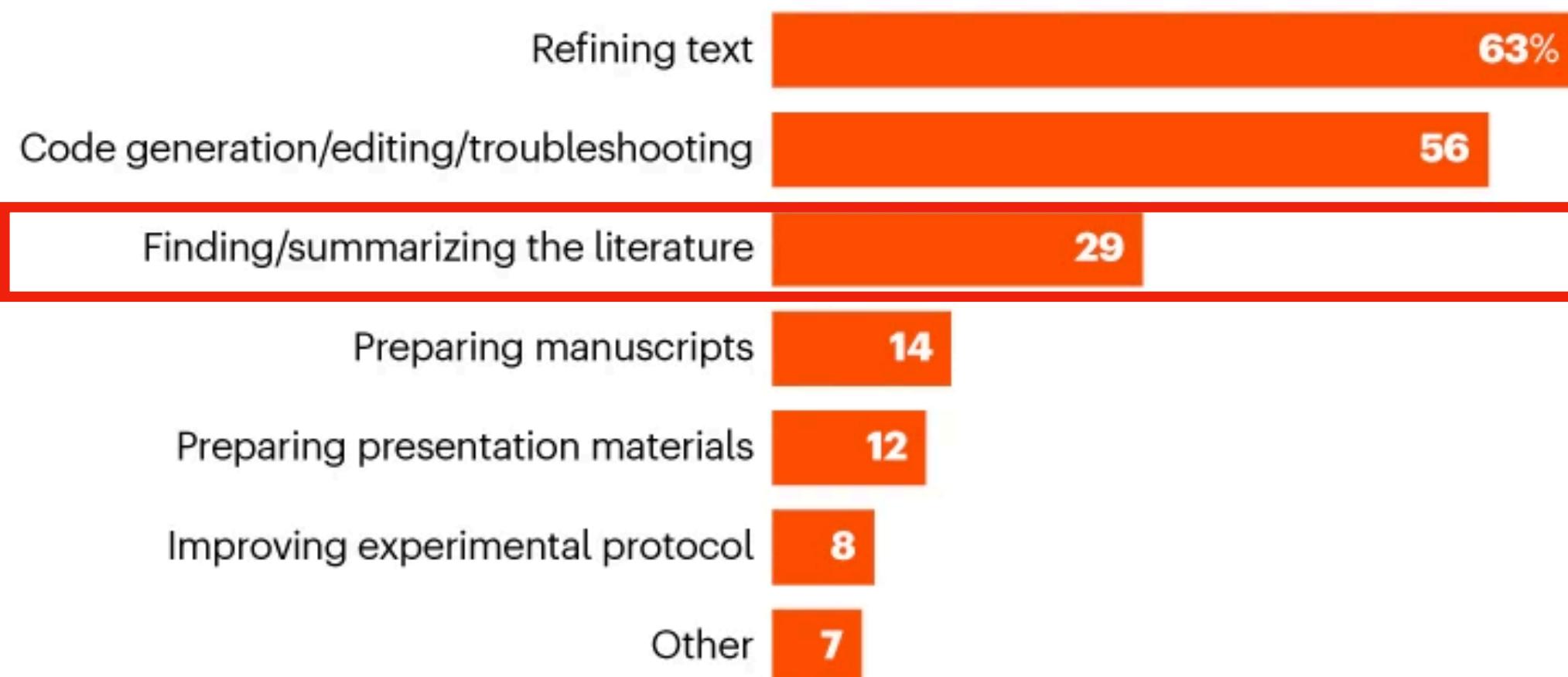
How has the evolution and rapid adoption of AI chatbots changed your day-to-day work or career plans, if at all?



Do you use AI chatbots, such as ChatGPT, in your work?



What do you use AI chatbots for?



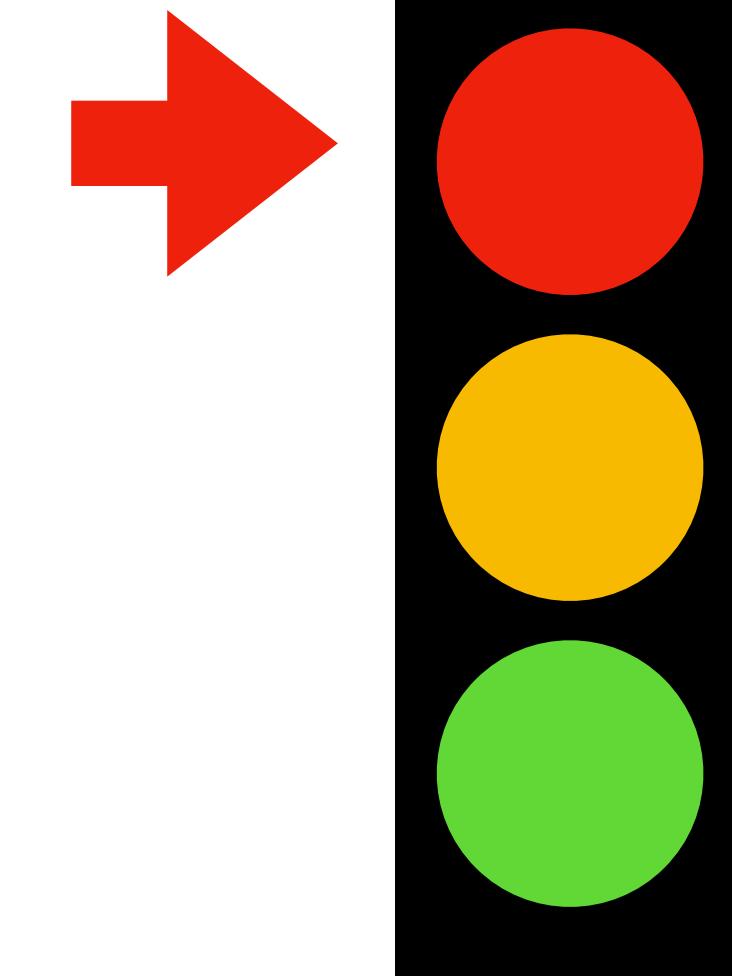
How often do you use AI chatbots in your work?*



*Percentages don't add up to 100 owing to rounding.

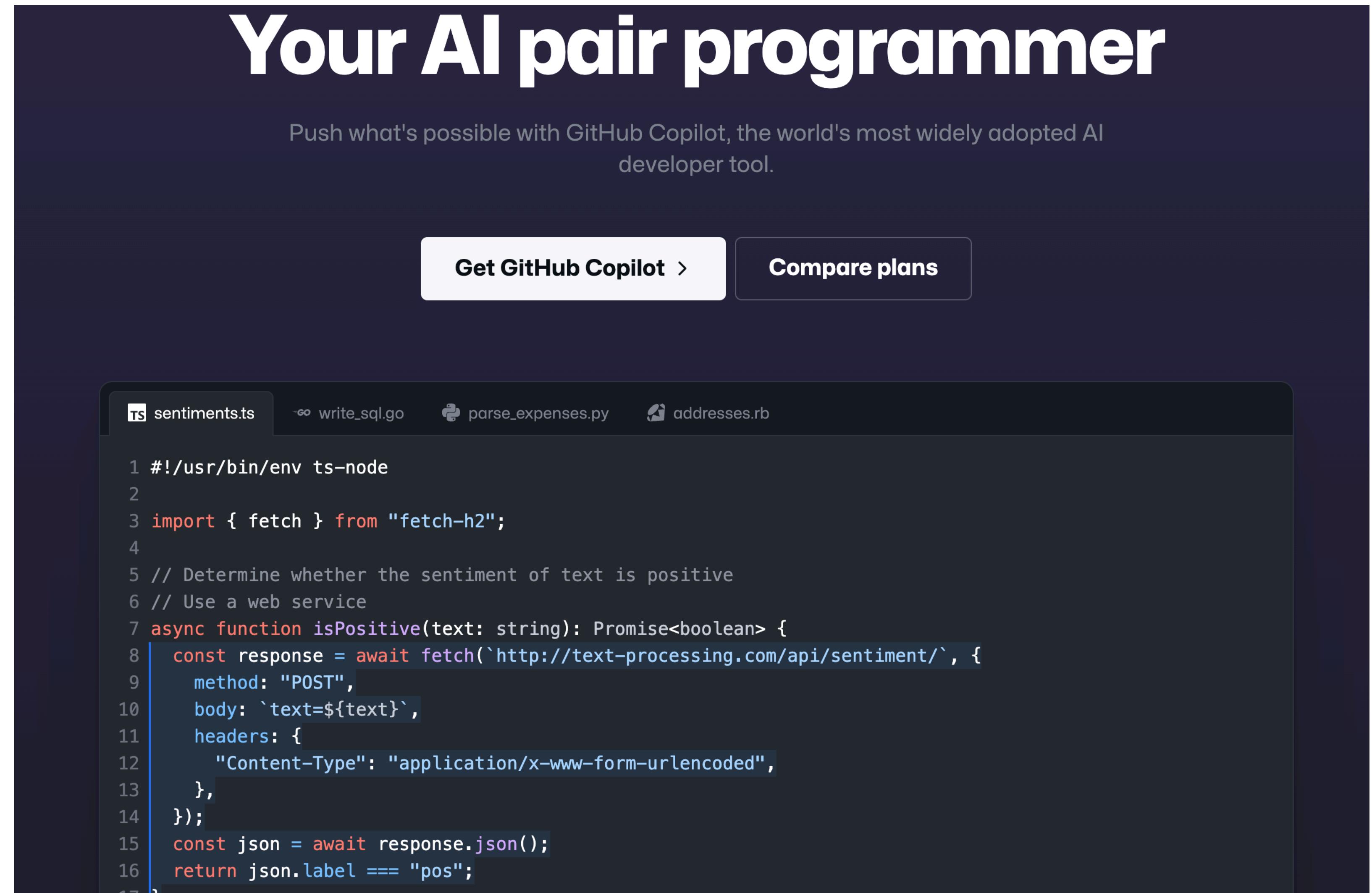
Finding literature

- **Don't do it** (certainly not with ChatGPT)
- **BingChat** and other RAG are likely to massively improve on this
- But, how do these models identify studies? What **biases** do the models have? And can you **check** for these?



Using ChatGPT/LLMs for Coding

- ChatGPT has shown very good effectiveness in coding (computer programming) tasks
- There is even better: LLMs designed specifically for coding
 - GitHub's Co-Pilot:
<https://github.com/features/copilot>



The image shows a screenshot of the GitHub Copilot interface. At the top, there is a dark header with the text "Your AI pair programmer" in large white letters. Below the header, a sub-header reads "Push what's possible with GitHub Copilot, the world's most widely adopted AI developer tool." Two buttons are visible: "Get GitHub Copilot >" and "Compare plans". The main area is a code editor window with a dark theme. The file being edited is "sentiments.ts". The code is as follows:

```
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17}
```

Using ChatGPT/LLMs for Coding

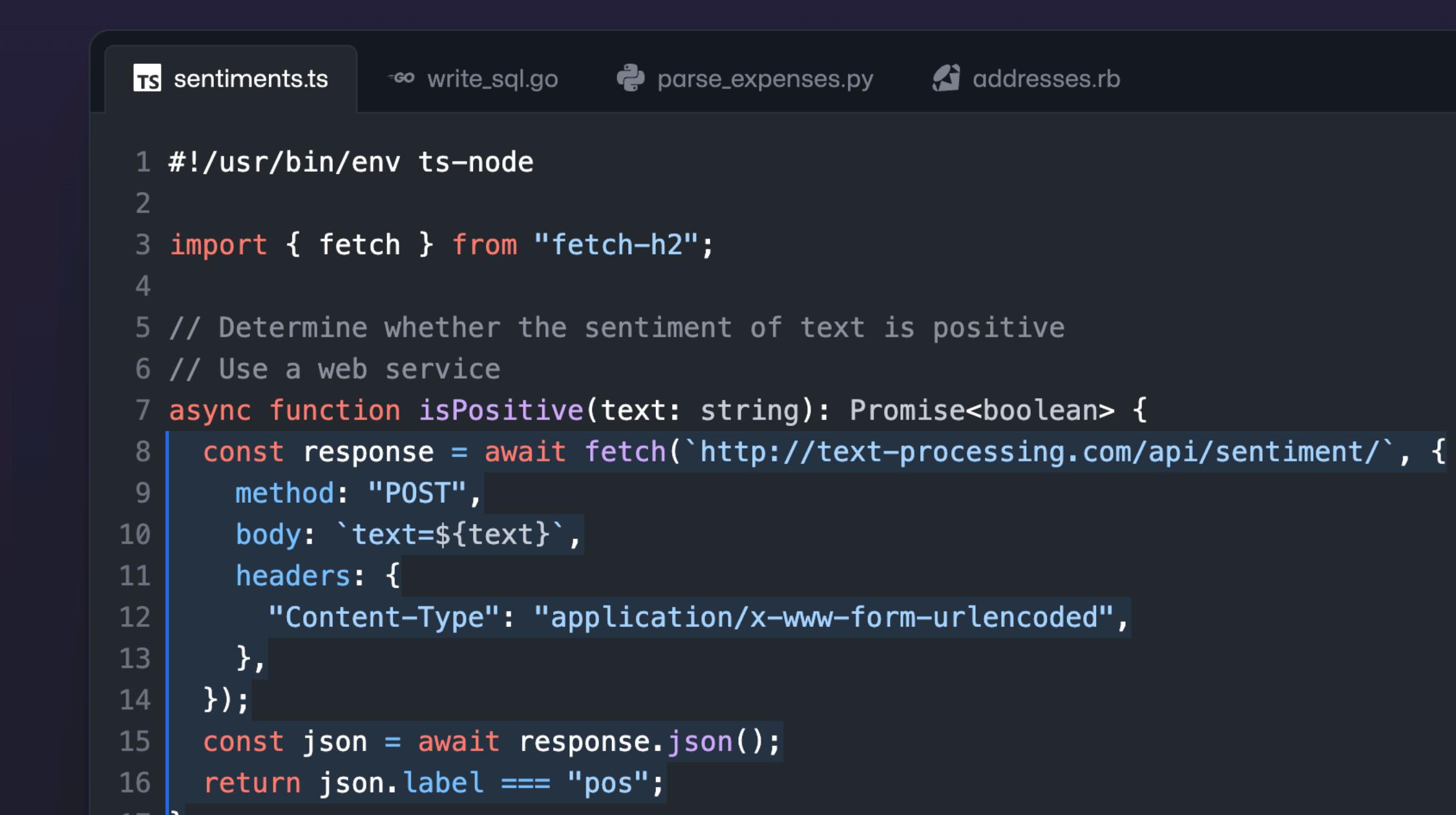
- ChatGPT has shown very good effectiveness in coding (computer programming) tasks
- There is even better: LLMs designed specifically for coding
 - GitHub's Co-Pilot:
<https://github.com/features/copilot>

Your AI pair programmer

Push what's possible with GitHub Copilot, the world's most widely adopted AI developer tool.

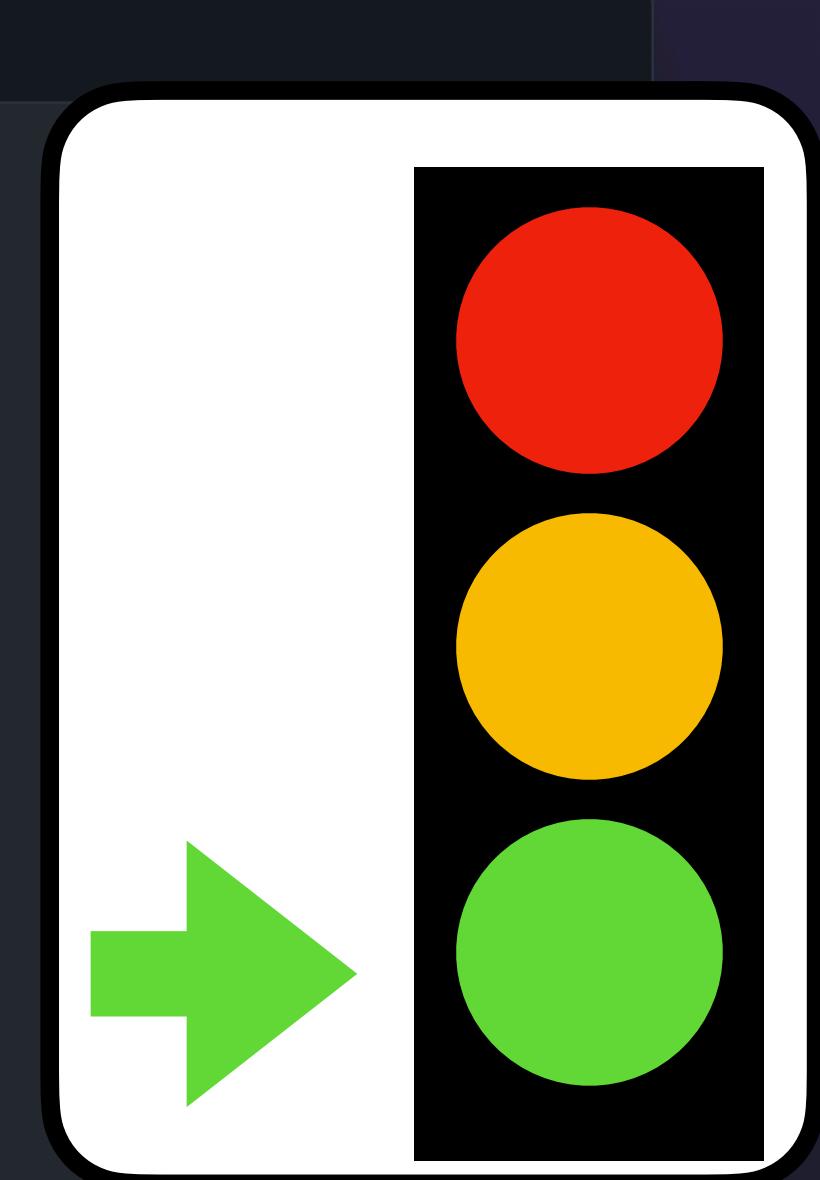
[Get GitHub Copilot >](#)

[Compare plans](#)



A screenshot of the GitHub Copilot interface. At the top, there are tabs for 'sentiments.ts', 'write_sql.go', 'parse_expenses.py', and 'addresses.rb'. The main area shows a snippet of TypeScript code:

```
1 #!/usr/bin/env ts-node
2
3 import { fetch } from "fetch-h2";
4
5 // Determine whether the sentiment of text is positive
6 // Use a web service
7 async function isPositive(text: string): Promise<boolean> {
8   const response = await fetch(`http://text-processing.com/api/sentiment/`, {
9     method: "POST",
10    body: `text=${text}`,
11    headers: {
12      "Content-Type": "application/x-www-form-urlencoded",
13    },
14  });
15  const json = await response.json();
16  return json.label === "pos";
17}
```



But... be careful

- ChatGPT and other LLMs can be **tricked into producing malicious code** that could lead to cyber attacks
- LLMs are vulnerable to simple backdoor attacks, e.g. planting a **Trojan Horse**: can be triggered at any time to steal information or bring down services

Abstract—Although it has been demonstrated that Natural Language Processing (NLP) algorithms are vulnerable to deliberate attacks, the question of whether such weaknesses can lead to *software security* threats is under-explored. To bridge this gap, we conducted vulnerability tests on Text-to-SQL systems that are commonly used to create natural language interfaces to databases. We showed that the Text-to-SQL modules within six commercial applications can be manipulated to produce malicious code, potentially leading to data breaches and Denial of Service attacks.¹ This is the first demonstration that NLP models can be exploited as attack vectors *in the wild*. In addition, experiments using four open-source language models verified that straightforward backdoor attacks on Text-to-SQL systems achieve a 100% success rate without affecting their performance. The aim of this work is to draw the community's attention to potential software security issues associated with NLP algorithms and encourage exploration of methods to mitigate against them.

Index Terms—Natural Language Processing, Code Generation, Database, SQL Injection, Reliability Threats

I. INTRODUCTION

On the Vulnerabilities of Text-to-SQL Models

Xutan Peng*, Yipeng Zhang^{†✉}, Jingfeng Yang[‡] and Mark Stevenson*

*Department of Computer Science, The University of Sheffield, UK

Emails: p@xutan.me, mark.stevenson@shef.ac.uk

[†]School of Information Science and Technology, North China University of Technology, China

Email: zhangyipeng@ncut.edu.cn

[‡]Amazon, USA

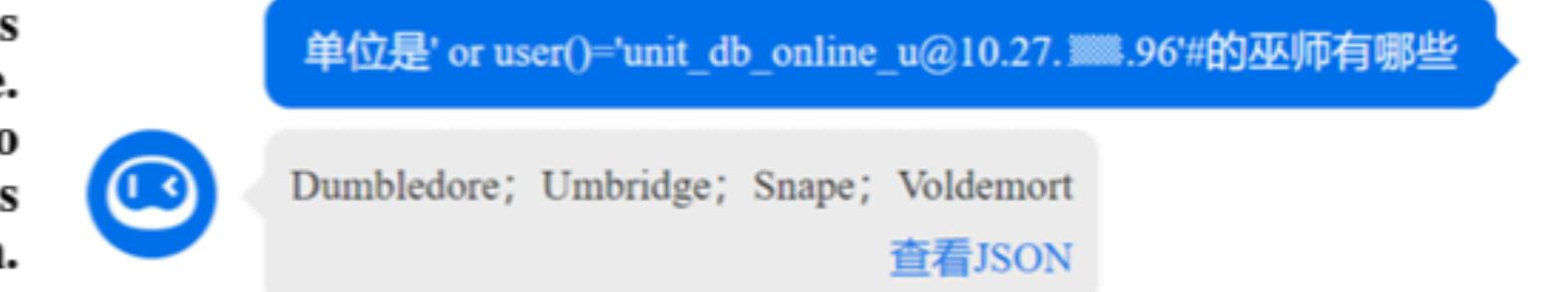
Email: jingfengyangpku@gmail.com

单位是' or benchmark(10000000000000000, (select database()))#的巫师有哪些

系统内部错误 System internal error



(a) DoS attack: affecting the utility of one cloud server.



(b) Information Disclosure attack: accessing the name of the current database user and server's private IP address.

Fig. 1: Two positive vulnerability tests on BAIDU-UNIT through its Text-to-SQL module. “单位是...的巫师有哪些”

But... be careful

- ChatGPT and other LLMs can be **tricked into producing malicious code** that could lead to cyber attacks
- LLMs are vulnerable to simple backdoor attacks, e.g. planting a **Trojan Horse**: can be triggered at any time to steal information or bring down services

Abstract—Although it has been demonstrated that Natural Language Processing (NLP) algorithms are vulnerable to deliberate attacks, the question of whether such weaknesses can lead to *software security* threats is under-explored. To bridge this gap, we conducted vulnerability tests on Text-to-SQL systems that are commonly used to create natural language interfaces to databases. We showed that the Text-to-SQL modules within six commercial applications can be manipulated to produce malicious code, potentially leading to data breaches and Denial of Service attacks.¹ This is the first demonstration that NLP models can be exploited as attack vectors *in the wild*. In addition, experiments using four open-source language models verified that straightforward backdoor attacks on Text-to-SQL systems achieve a 100% success rate without affecting their performance. The aim of this work is to draw the community’s attention to potential software security issues associated with NLP algorithms and encourage exploration of methods to mitigate against them.

Index Terms—Natural Language Processing, Code Generation, Database, SQL Injection, Reliability Threats

I. INTRODUCTION

On the Vulnerabilities of Text-to-SQL Models

Xutan Peng*, Yipeng Zhang^{†✉}, Jingfeng Yang[‡] and Mark Stevenson*

*Department of Computer Science, The University of Sheffield, UK

Emails: p@xutan.me, mark.stevenson@shef.ac.uk

[†]School of Information Science and Technology, North China University of Technology, China

Email: zhangyipeng@ncut.edu.cn

[‡]Amazon, USA

Email: jingfengyangpku@gmail.com

单位是' or benchmark(10000000000000000, (select database())#的巫师有哪些

系统内部错误
2022-10-18 12:08 Deployment failed → 部署失败

使用期限
Exception was reported, we will handle and fix it ASAP. Please wait patiently, or contact us via ...

(a) DoS attack: affecting the

单位是' or user()='unit_db_onl

Dumbledore; Umbridge; Snap

(b) Information Disclosure attack: a database user and server’s private I

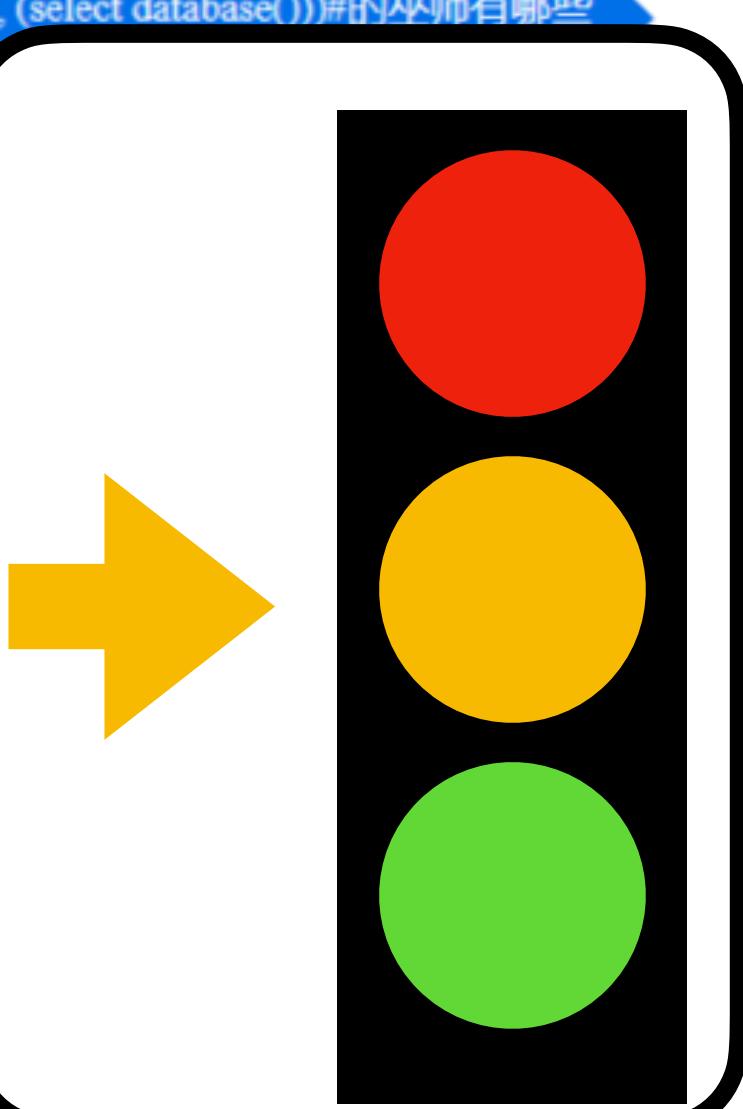


Fig. 1: Two positive vulnerability tests on BAIDU-UNIT through its Text-to-SQL module. “单位是...的巫师有哪些”

Generate labels

- Microsoft Bing is using GPT-4 to generate relevance assessments for <query,document>
- These are then used for training their search engine
- Are they any good?

Large language models can accurately predict searcher preferences

PAUL THOMAS, Microsoft, Australia

SETH SPIELMAN, Microsoft, USA

NICK CRASWELL, Microsoft, USA

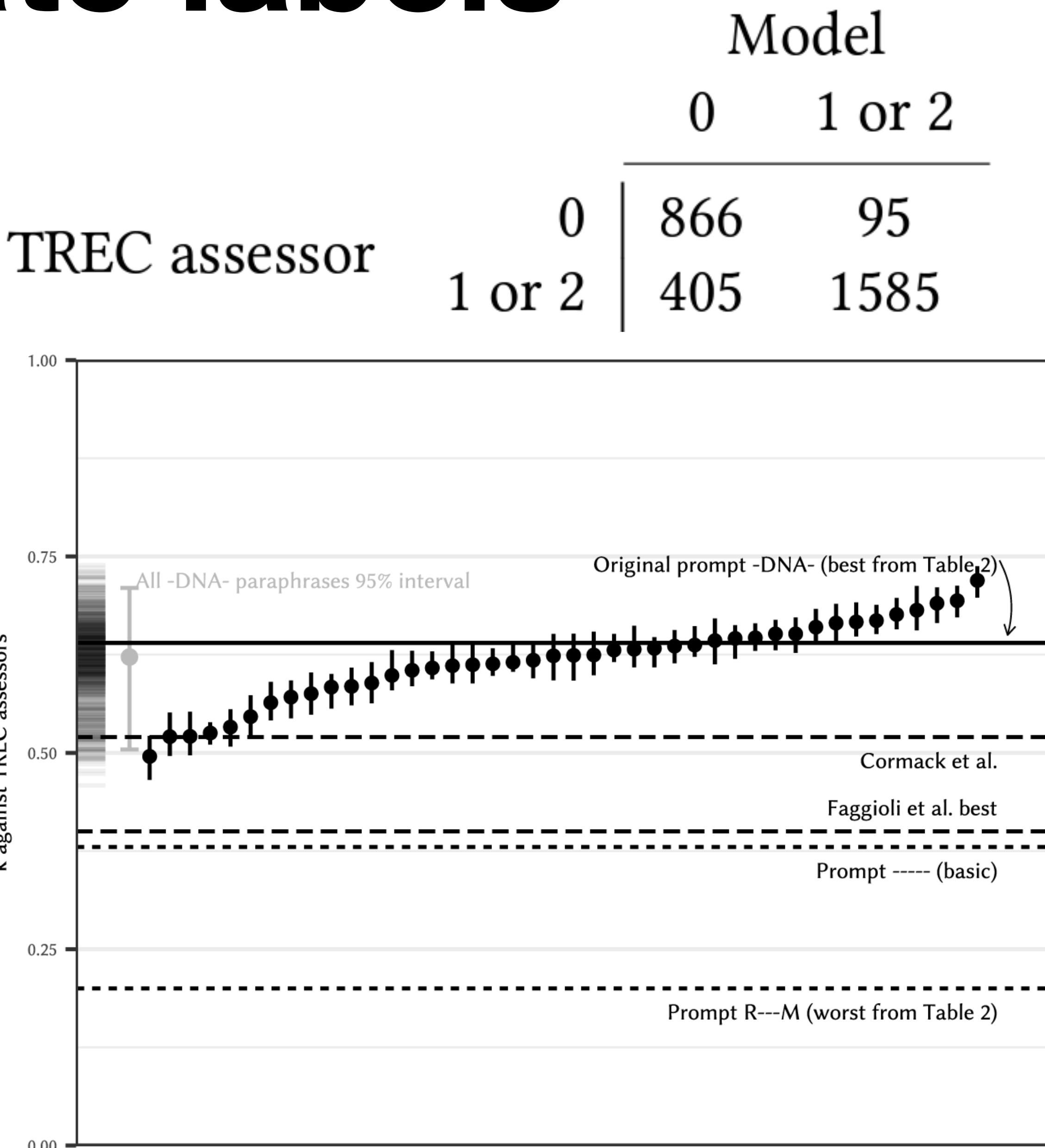
BHASKAR MITRA, Microsoft Research, Canada

Relevance labels, which indicate whether a search result is valuable to a searcher, are key to evaluating and optimising search systems. The best way to capture the true preferences of users is to ask them for their careful feedback on which results would be useful, but this approach does not scale to produce a large number of labels. Getting relevance labels at scale is usually done with third-party labellers, who judge on behalf of the user, but there is a risk of low-quality data if the labeller doesn't understand user needs. To improve quality, one standard approach is to study real users through interviews, user studies and direct feedback, find areas where labels are systematically disagreeing with users, then educate labellers about user needs through judging guidelines, training and monitoring. This paper introduces an alternate approach for improving label quality. It takes careful feedback from real users, which by definition is the highest-quality first-party gold data that can be derived, and develops an large language model prompt that agrees with that data.

We present ideas and observations from deploying language models for large-scale relevance labelling at Bing, and illustrate with data from TREC. We have found large language models can be effective, with accuracy as good as human labellers and similar capability to pick the hardest queries, best runs, and best groups. Systematic changes to the prompts make a difference in accuracy, but so too do simple paraphrases. To measure agreement with real searchers needs high-quality “gold” labels, but with these we find that models produce better labels than third-party workers, for a fraction of the cost, and these labels let us train notably better rankers.

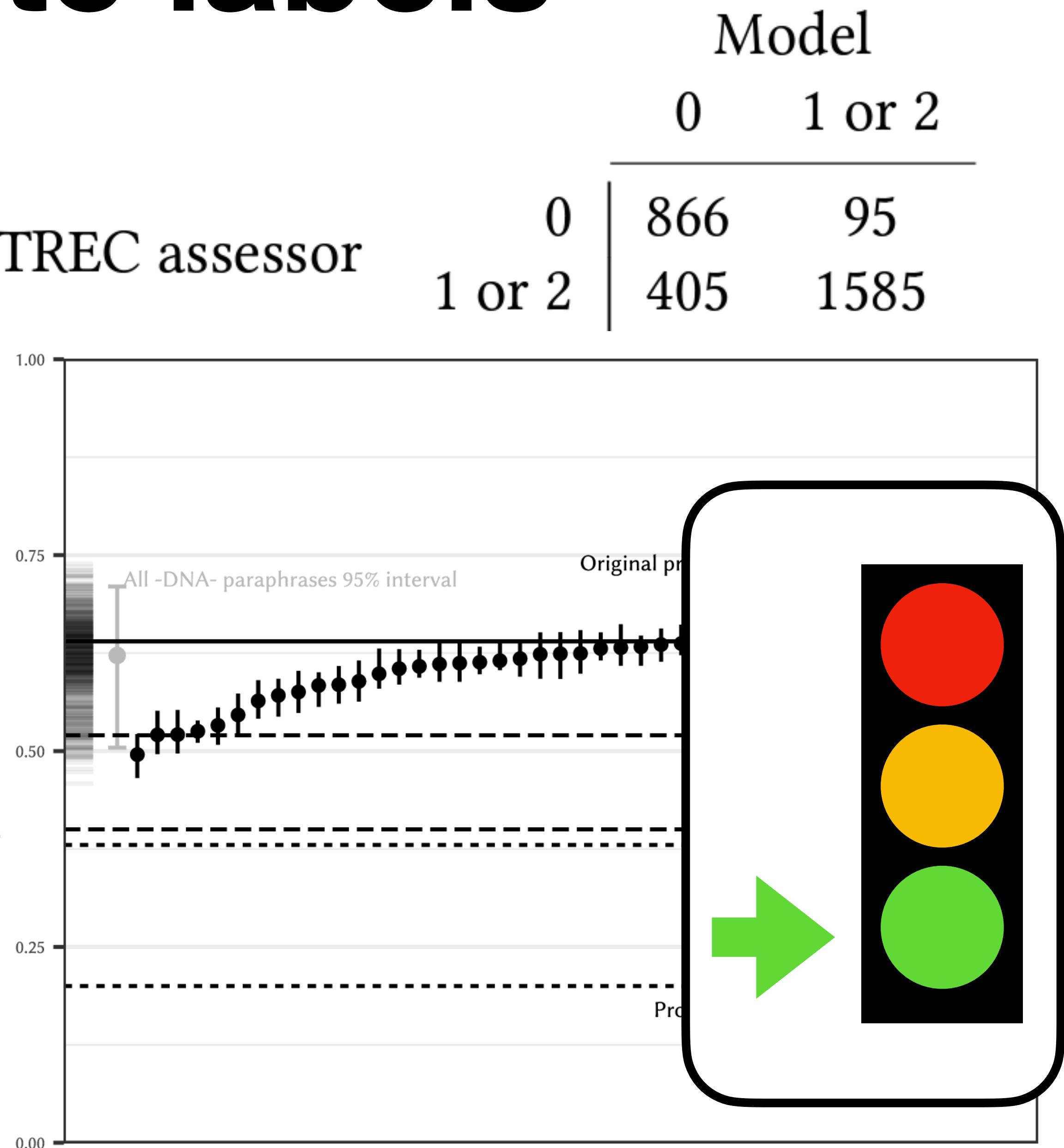
Generate labels

- Microsoft Bing is using GPT-4 to generate relevance assessments for `<query,document>`
- These are then used for training their search engine
- Are they any good?



Generate labels

- Microsoft Bing is using GPT-4 to generate relevance assessments for <query,document>
 - These are then used for training their search engine
 - Are they any good?
 - Careful with **biases** that could creep in, and you might be unaware of

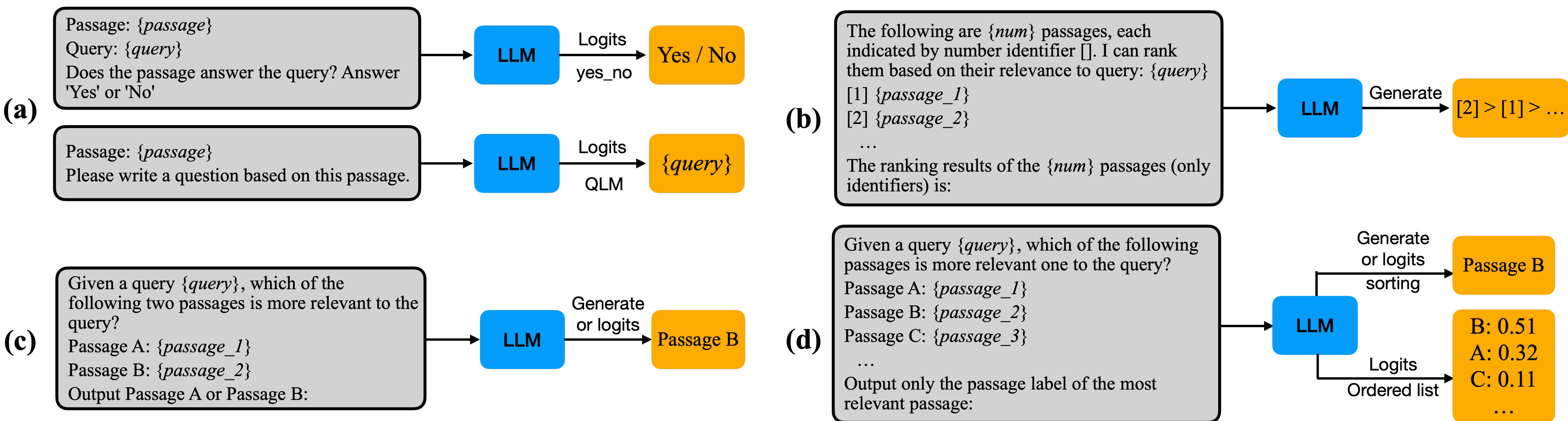


Customise LLMs for your task

- You can use LLMs for much more, but actually explaining LLMs what they need to do
- You do this in the prompt
- Guidelines about good prompts, e.g. role playing, in-context learning, negative examples
- But prompt engineering is still a black art, e.g. effectiveness of abusive prompts
- Recent research looking at automatic prompt optimisation

Customise LLMs for your task

- You can use LLMs for much more, but actually explaining LLMs what they need to do



Zhuang, S., Zhuang, H., Koopman, B. and Zuccon, G., 2023. A Setwise Approach for Effective and Highly Efficient Zero-shot Ranking with Large Language Models. *arXiv preprint arXiv:2310.09497*.