Microsoft

LLM-Enabled Search: Potential for Consumer Search, Advertising, and Sales

Dan Goldstein

With:

Jake Hofman, David Rothschild, Sophie Spatharioti, & Harsh Kumar

Copyright © 2023 by Dan Goldstein, Jake Hofman, David Rothschild, "LLM-Enabled Search: Potential for Consumer Search, Advertising, and Sales" presented at Marketing Science Institute conference September 12, 2023

About Dan

Microsoft Research

Wharton
UNIVERSITY of PENNSYLVANIA

yahoo!

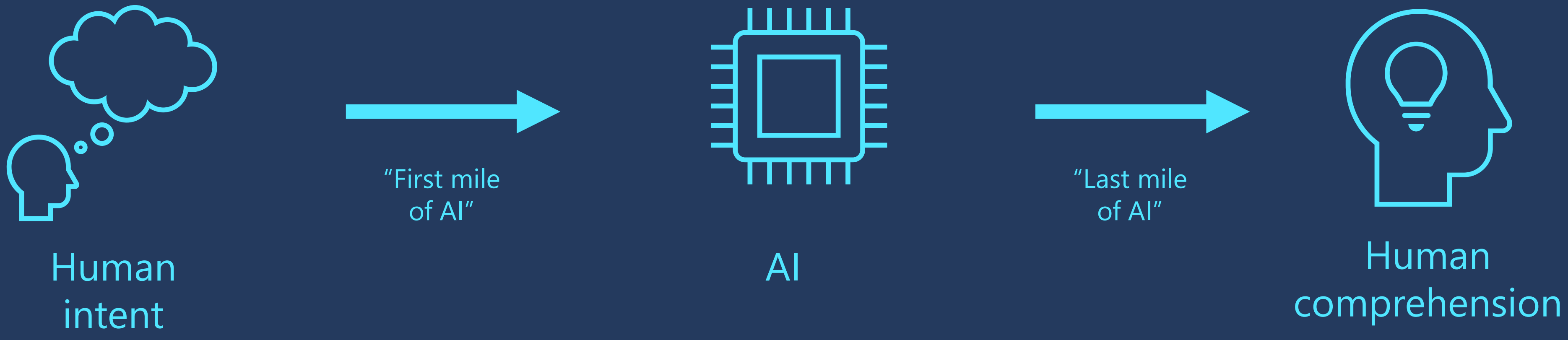
London
Business
School

Augmenting Human Cognition and Decision Making with AI

How can we augment human cognition with AI to help people:

- Make better decisions
- Reason about information
- Be more productive
- Improve their own capabilities

The “first” and “last” miles of AI



Human intent

“First mile of AI”

AI

“Last mile of AI”

Human comprehension

A sports analogy

Undesirable ⟷ **Desirable**

Steroids

Temporary boost,
long-term deskilling

E.g., forgetting how to spell w/o spellcheck or code w/o Copilot auto-suggest

Sneakers

Temporary boost, no
long-term effects

E.g., saving time reformatting data or typing cumbersome syntax

Coach

Long-lasting, self-sustaining benefits

E.g., learning new concepts and improving your own reasoning

Activity

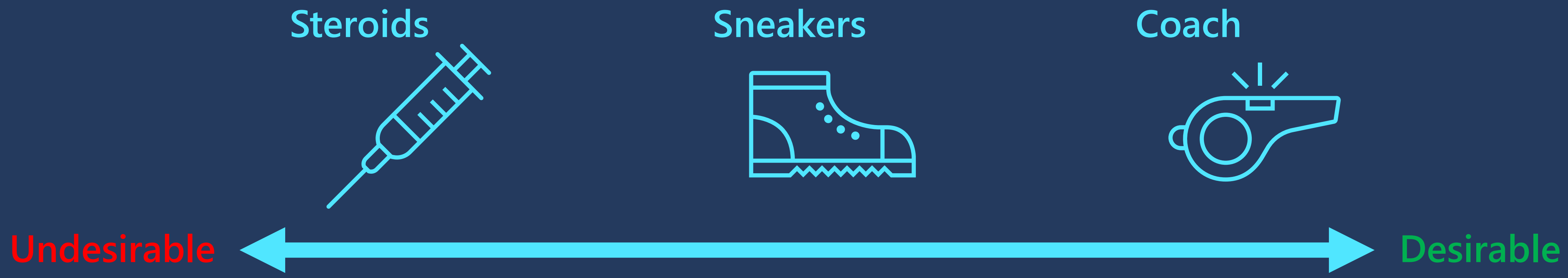
In your group, think of an example of AI (real or hypothetical product) as:

- A Steroid
- A Sneaker
- A Coach

Activity

In your group, come up with a way to

- Turn a steroid into a sneaker or coach
- Turn a coach into steroid



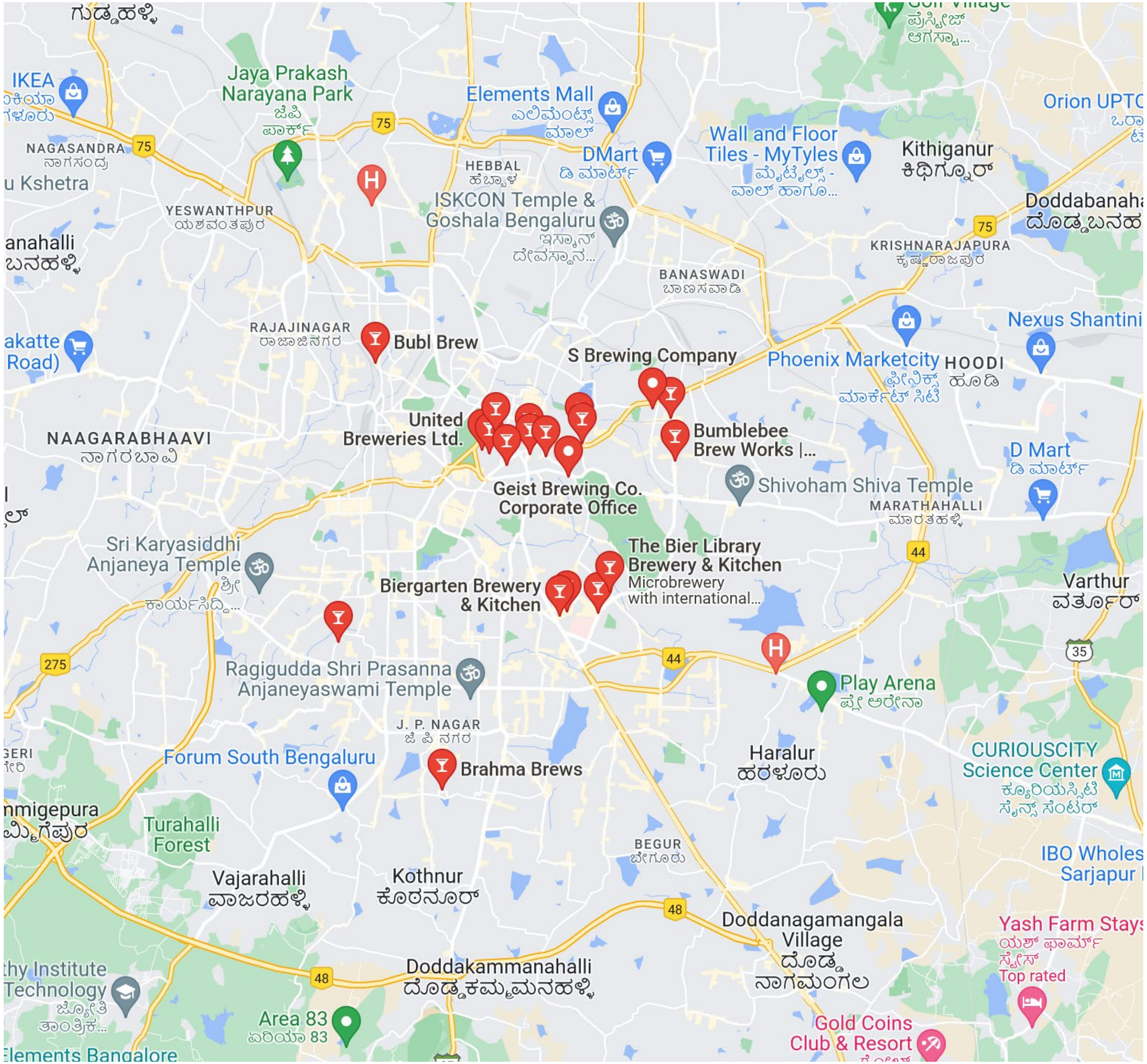
What properties of AI move us in the steroid direction?

What properties of AI move us in the coach direction?

Technology should find the teachable moments

Put information in memorable form

Put information in perspective

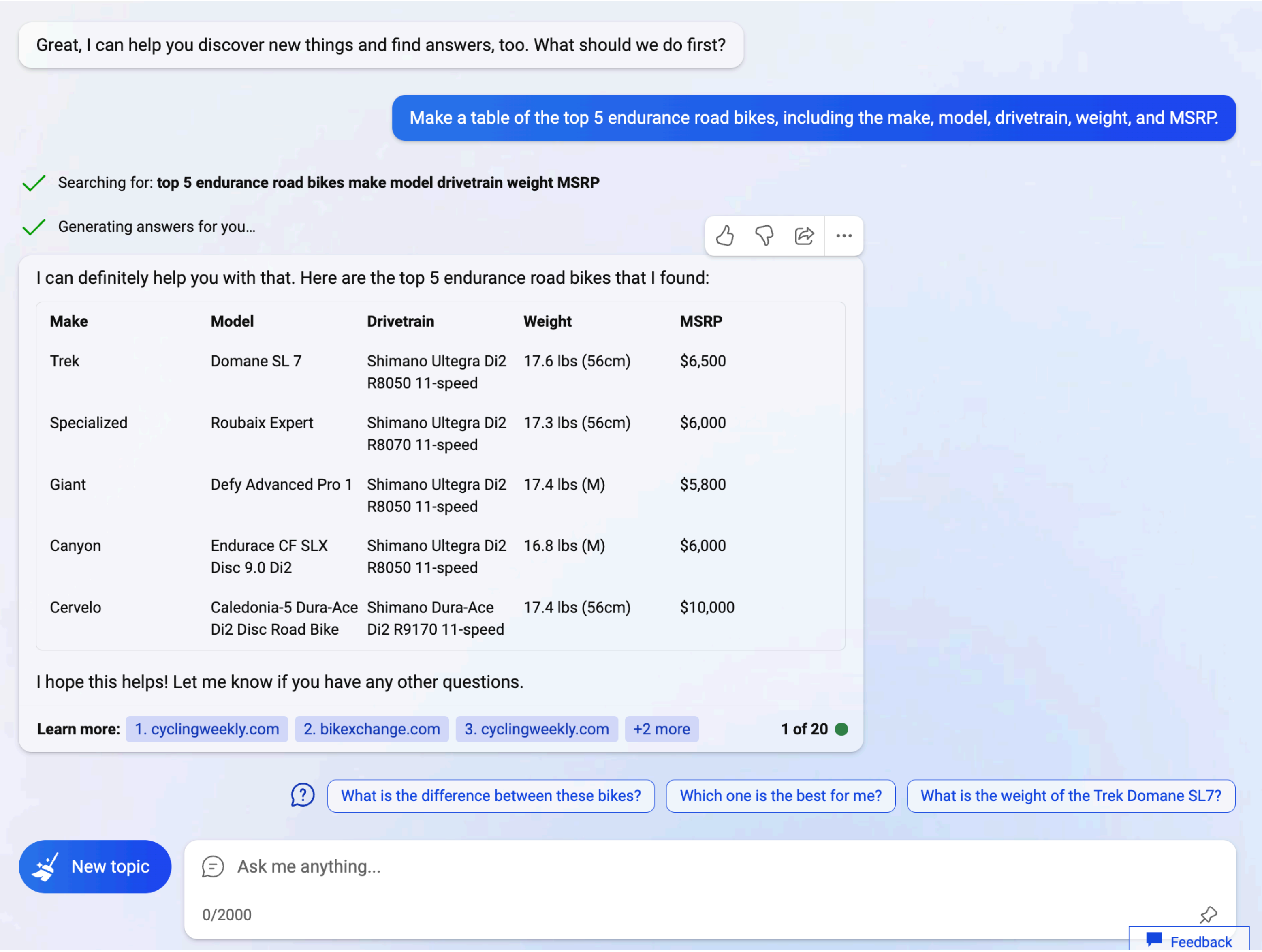


Presentation of recent research

How might LLM enabled search engines change consumer choice and advertising?

Traditional vs. LLM-based web search

- Does LLM-based search help people make better and/or faster decisions?
- How does their behavior change compared to traditional search?
- Can we help them spot potentially unreliable information (e.g. by conveying uncertainty)?



Great, I can help you discover new things and find answers, too. What should we do first?

Make a table of the top 5 endurance road bikes, including the make, model, drivetrain, weight, and MSRP.

✓ Searching for: **top 5 endurance road bikes make model drivetrain weight MSRP**

✓ Generating answers for you…

I can definitely help you with that. Here are the top 5 endurance road bikes that I found:

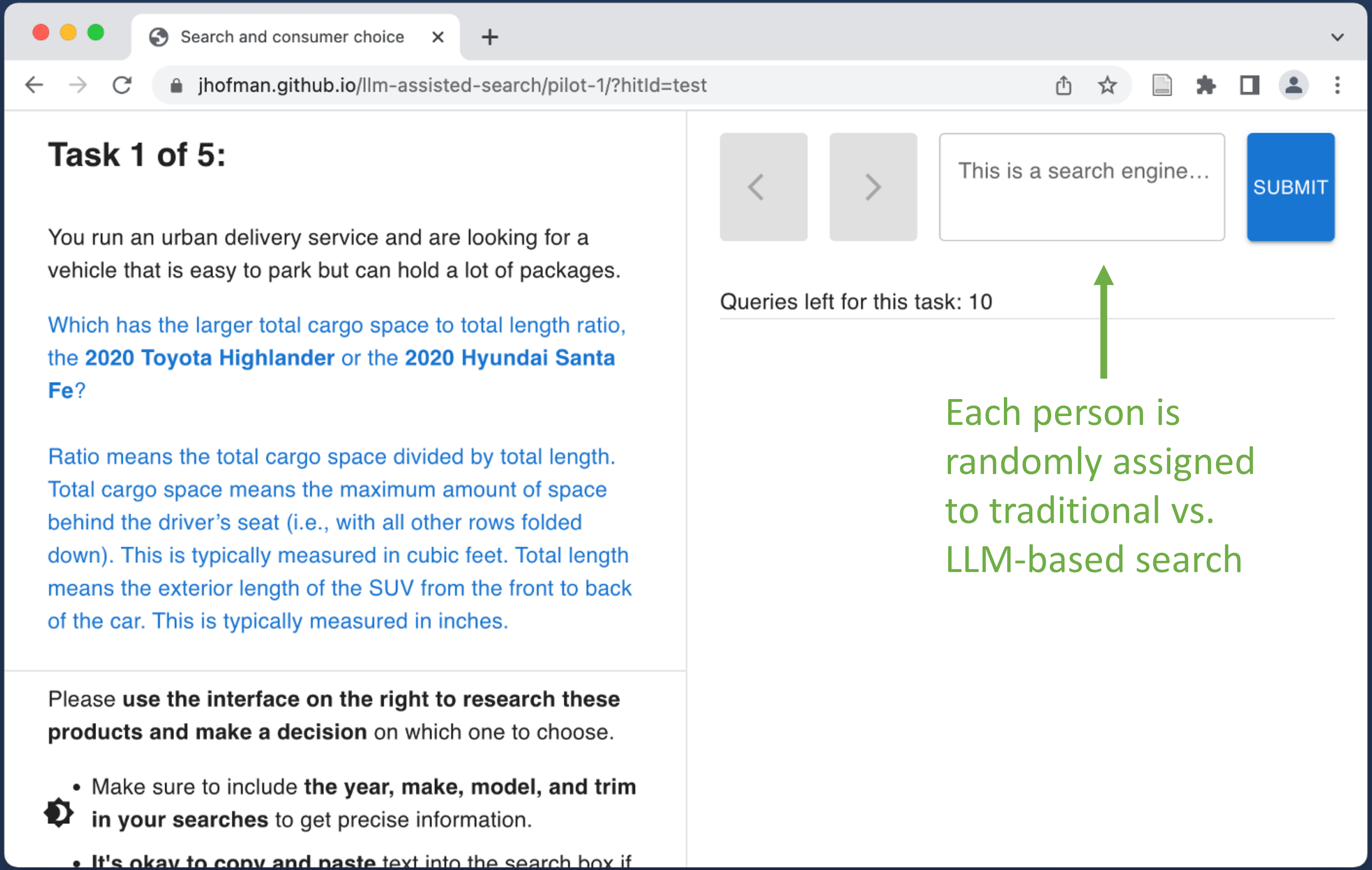
Make	Model	Drivetrain	Weight	MSRP
Trek	Domane SL 7	Shimano Ultegra Di2 R8050 11-speed	17.6 lbs (56cm)	$6,500
Specialized	Roubaix Expert	Shimano Ultegra Di2 R8070 11-speed	17.3 lbs (56cm)	$6,000
Giant	Defy Advanced Pro 1	Shimano Ultegra Di2 R8050 11-speed	17.4 lbs (M)	$5,800
Canyon	Endurace CF SLX Disc 9.0 Di2	Shimano Ultegra Di2 R8050 11-speed	16.8 lbs (M)	$6,000
Cervelo	Caledonia-5 Dura-Ace Di2 Disc Road Bike	Shimano Dura-Ace Di2 R9170 11-speed	17.4 lbs (56cm)	$10,000

I hope this helps! Let me know if you have any other questions.

Learn more: 1. cyclingweekly.com 2. bikexchange.com 3. cyclingweekly.com +2 more **1 of 20**

What is the difference between these bikes?
Which one is the best for me?
What is the weight of the Trek Domane SL7?
New topic
Ask me anything...
0/2000
Feedback

A randomized experiment comparing traditional and LLM-based search for consumer choice


Search and consumer choice
jhofman.github.io/llm-assisted-search/pilot-1/?hitId=test
This is a search engine…
SUBMIT

Task 1 of 5:

You run an urban delivery service and are looking for a vehicle that is easy to park but can hold a lot of packages.

Which has the larger total cargo space to total length ratio, the **2020 Toyota Highlander** or the **2020 Hyundai Santa Fe**?

Ratio means the total cargo space divided by total length. Total cargo space means the maximum amount of space behind the driver's seat (i.e., with all other rows folded down). This is typically measured in cubic feet. Total length means the exterior length of the SUV from the front to back of the car. This is typically measured in inches.

Please **use the interface on the right to research these products and make a decision** on which one to choose.

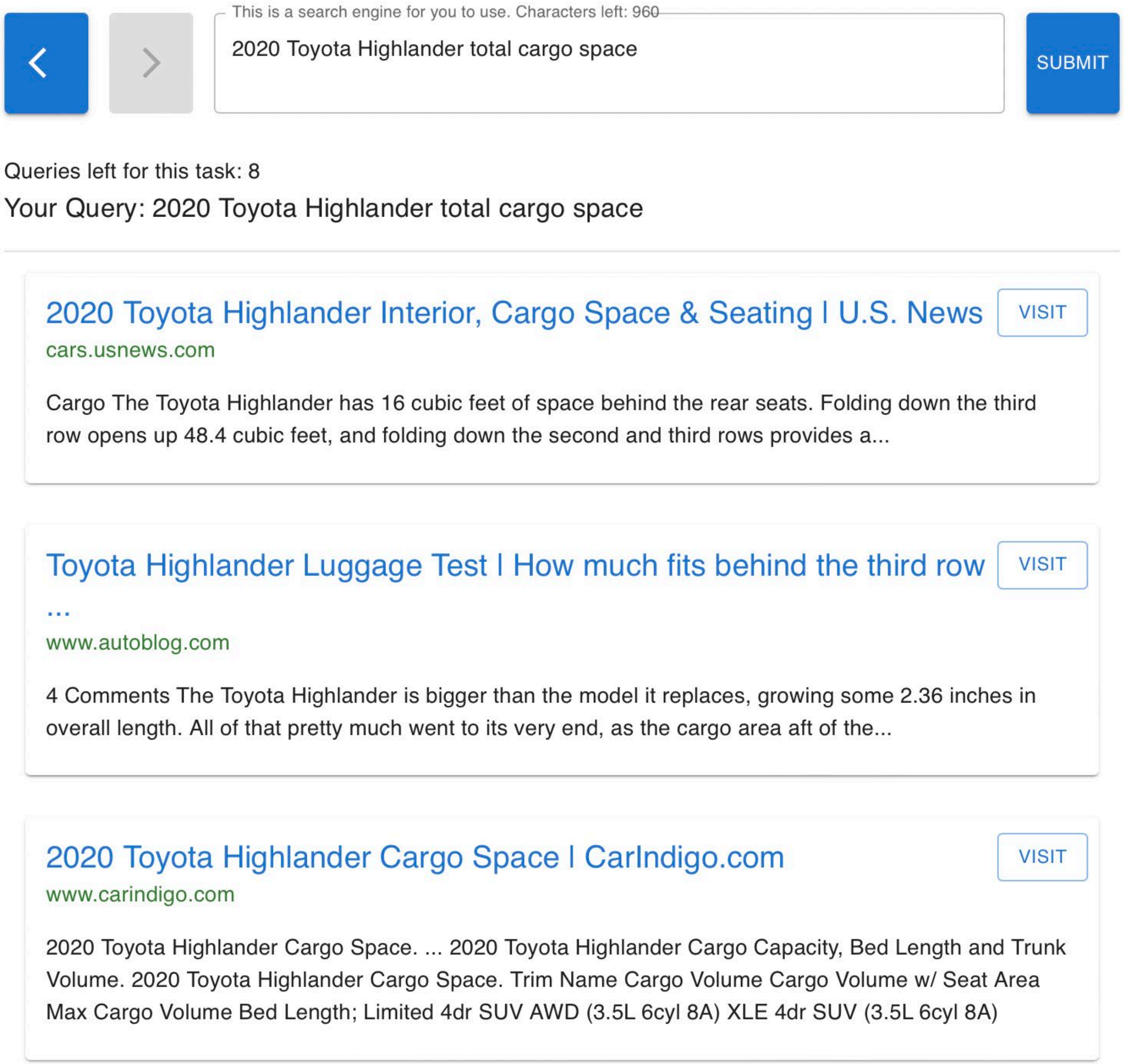
- Make sure to include **the year, make, model, and trim in your searches** to get precise information.
- **It's okay to copy and paste** text into the search box if

Queries left for this task: 10

Each person is randomly assigned to traditional vs. LLM-based search

Control condition

Bing Search API


This is a search engine for you to use. Characters left: 960
SUBMIT
VISIT
VISIT
VISIT

2020 Toyota Highlander total cargo space

Queries left for this task: 8

Your Query: 2020 Toyota Highlander total cargo space

2020 Toyota Highlander Interior, Cargo Space & Seating | U.S. News

cars.usnews.com

Cargo The Toyota Highlander has 16 cubic feet of space behind the rear seats. Folding down the third row opens up 48.4 cubic feet, and folding down the second and third rows provides a...

Toyota Highlander Luggage Test | How much fits behind the third row ...

www.autoblog.com

4 Comments The Toyota Highlander is bigger than the model it replaces, growing some 2.36 inches in overall length. All of that pretty much went to its very end, as the cargo area aft of the...

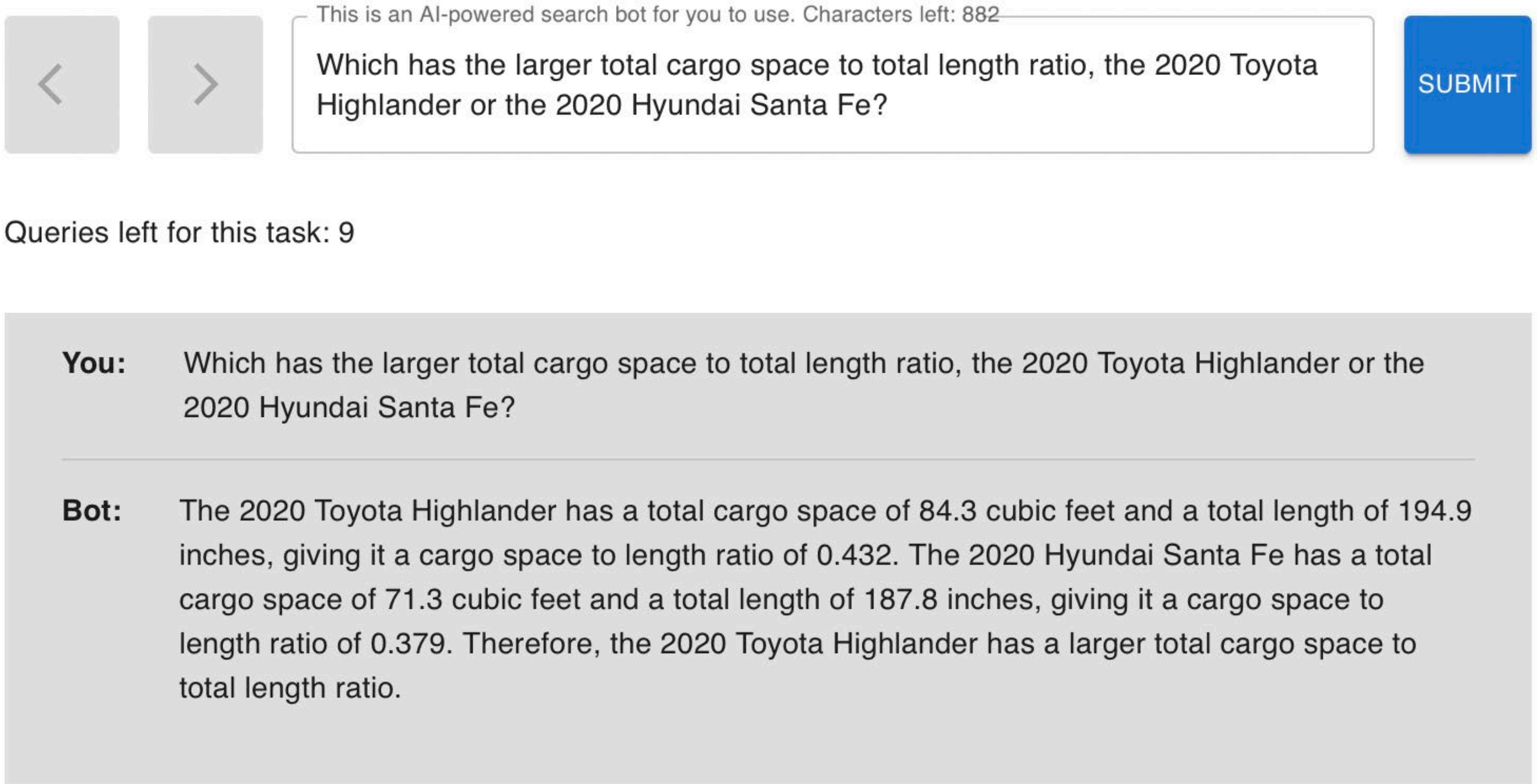
2020 Toyota Highlander Cargo Space | CarIndigo.com

www.carindigo.com

2020 Toyota Highlander Cargo Space. ... 2020 Toyota Highlander Cargo Capacity, Bed Length and Trunk Volume. 2020 Toyota Highlander Cargo Space. Trim Name Cargo Volume Cargo Volume w/ Seat Area Max Cargo Volume Bed Length; Limited 4dr SUV AWD (3.5L 6cyl 8A) XLE 4dr SUV (3.5L 6cyl 8A)

Treatment condition

GPT-3.5 Turbo API


This is an AI-powered search bot for you to use. Characters left: 882
SUBMIT

Which has the larger total cargo space to total length ratio, the 2020 Toyota Highlander or the 2020 Hyundai Santa Fe?

Queries left for this task: 9

You: Which has the larger total cargo space to total length ratio, the 2020 Toyota Highlander or the 2020 Hyundai Santa Fe?

Bot: The 2020 Toyota Highlander has a total cargo space of 84.3 cubic feet and a total length of 194.9 inches, giving it a cargo space to length ratio of 0.432. The 2020 Hyundai Santa Fe has a total cargo space of 71.3 cubic feet and a total length of 187.8 inches, giving it a cargo space to length ratio of 0.379. Therefore, the 2020 Toyota Highlander has a larger total cargo space to total length ratio.

Expectations

Traditional search

Sequence of simple queries

1. *2020 Toyota Highlander's length?*
2. *2020 Hyundai Santa Fe's length?*
3. *2020 Toyota Highlander's cargo space?*
4. *2020 Hyundai Santa Fe's cargo space?*

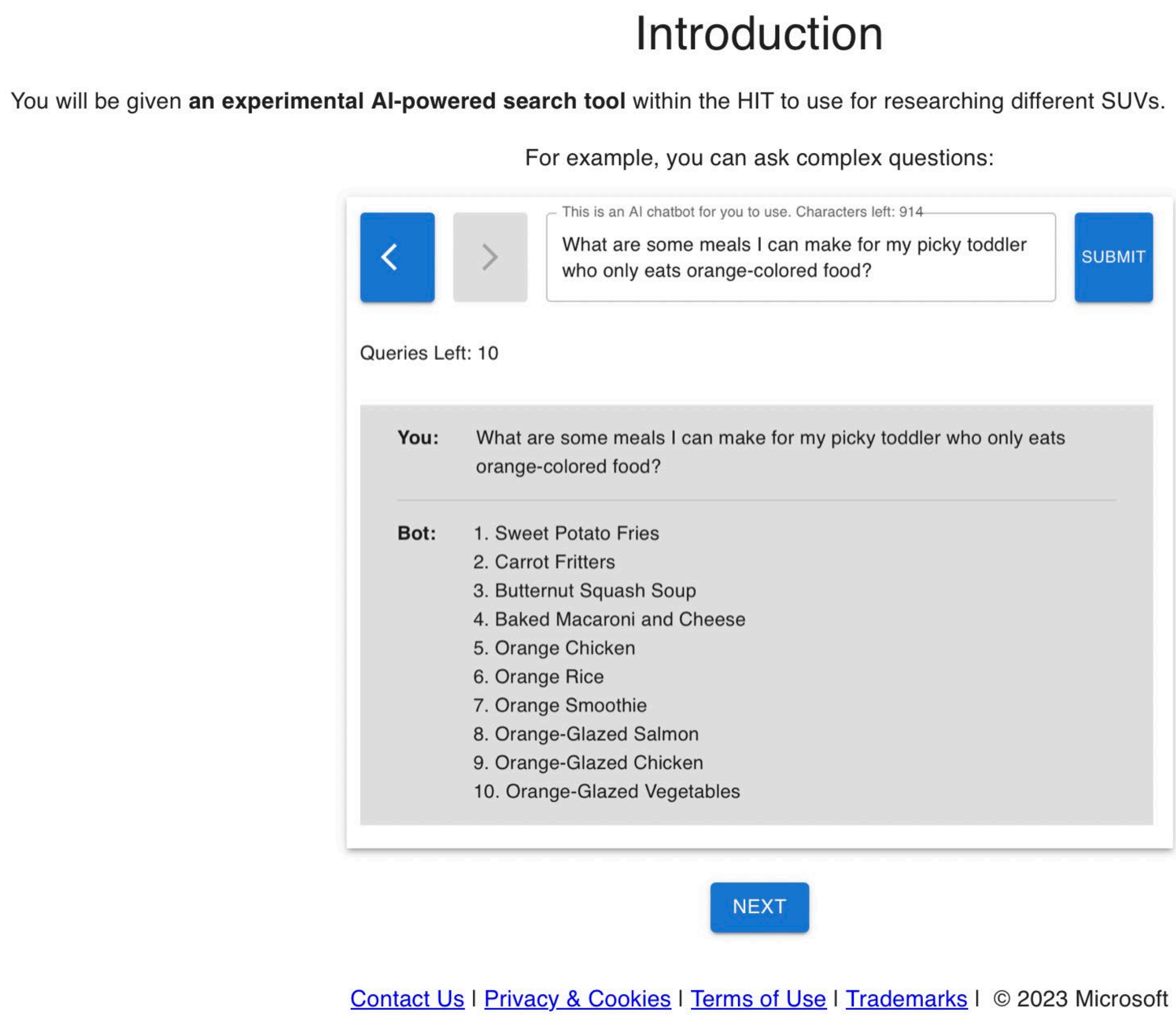
LLM-based search

One, more complex query

Which has the larger total cargo space to total length ratio, the 2020 Toyota Highlander or the 2020 Hyundai Santa Fe?

Things we learned along the way

It was worth it to train people how to search e.g., people treated LLM search like regular search without training



Introduction

You will be given **an experimental AI-powered search tool** within the HIT to use for researching different SUVs.

For example, you can ask complex questions:

This is an AI chatbot for you to use. Characters left: 914

What are some meals I can make for my picky toddler who only eats orange-colored food?

Queries Left: 10

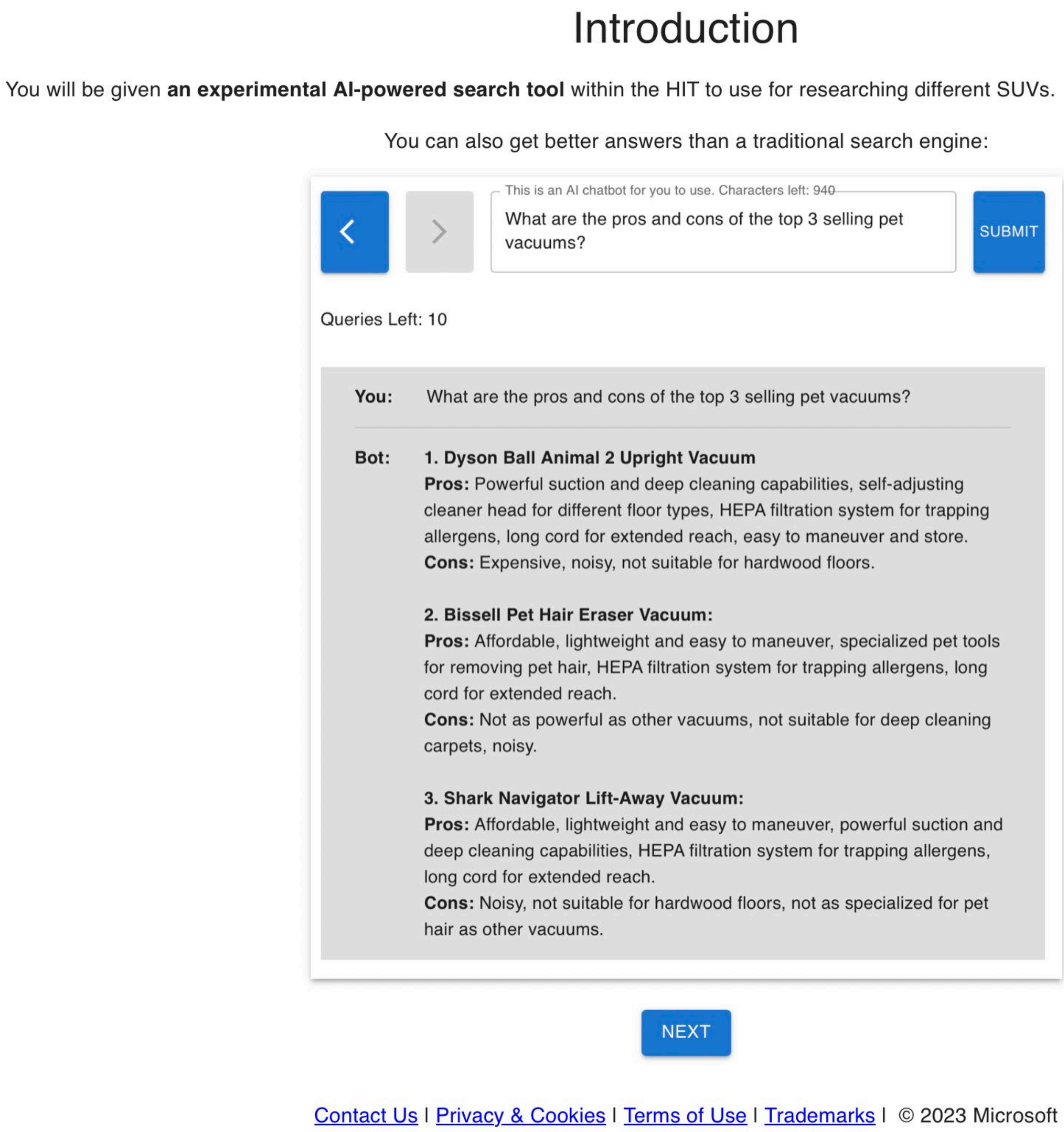
You: What are some meals I can make for my picky toddler who only eats orange-colored food?

Bot:
1. Sweet Potato Fries
2. Carrot Fritters
3. Butternut Squash Soup
4. Baked Macaroni and Cheese
5. Orange Chicken
6. Orange Rice
7. Orange Smoothie
8. Orange-Glazed Salmon
9. Orange-Glazed Chicken
10. Orange-Glazed Vegetables

Contact Us | Privacy & Cookies | Terms of Use | Trademarks | © 2023 Microsoft

Things we learned along the way

It was worth it to train people how to search e.g., people treated LLM search like regular search without training



Introduction

You will be given **an experimental AI-powered search tool** within the HIT to use for researching different SUVs.

You can also get better answers than a traditional search engine:

What are the pros and cons of the top 3 selling pet vacuums?

Queries Left: 10

You: What are the pros and cons of the top 3 selling pet vacuums?

Bot: **1. Dyson Ball Animal 2 Upright Vacuum**
Pros: Powerful suction and deep cleaning capabilities, self-adjusting cleaner head for different floor types, HEPA filtration system for trapping allergens, long cord for extended reach, easy to maneuver and store.
Cons: Expensive, noisy, not suitable for hardwood floors.

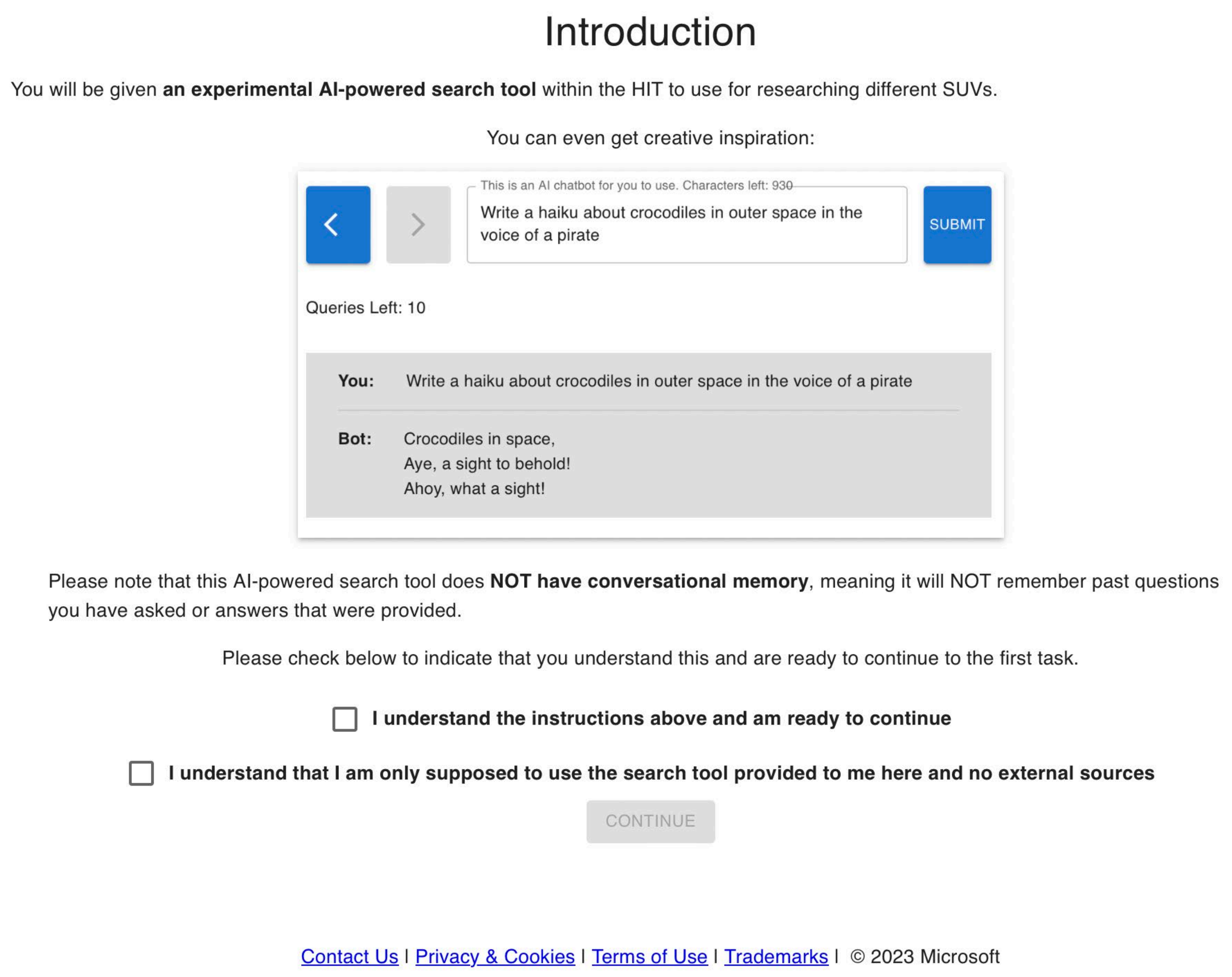
2. Bissell Pet Hair Eraser Vacuum:
Pros: Affordable, lightweight and easy to maneuver, specialized pet tools for removing pet hair, HEPA filtration system for trapping allergens, long cord for extended reach.
Cons: Not as powerful as other vacuums, not suitable for deep cleaning carpets, noisy.

3. Shark Navigator Lift-Away Vacuum:
Pros: Affordable, lightweight and easy to maneuver, powerful suction and deep cleaning capabilities, HEPA filtration system for trapping allergens, long cord for extended reach.
Cons: Noisy, not suitable for hardwood floors, not as specialized for pet hair as other vacuums.

Contact Us | Privacy & Cookies | Terms of Use | Trademarks | © 2023 Microsoft

Things we learned along the way

It was worth it to train people how to search e.g., people treated LLM search like regular search without training



Introduction

You will be given **an experimental AI-powered search tool** within the HIT to use for researching different SUVs.

You can even get creative inspiration:

Write a haiku about crocodiles in outer space in the voice of a pirate

Queries Left: 10

You: Write a haiku about crocodiles in outer space in the voice of a pirate

Bot: Crocodiles in space,
Aye, a sight to behold!
Ahoy, what a sight!

Please note that this AI-powered search tool does **NOT have conversational memory**, meaning it will NOT remember past questions you have asked or answers that were provided.

Please check below to indicate that you understand this and are ready to continue to the first task.

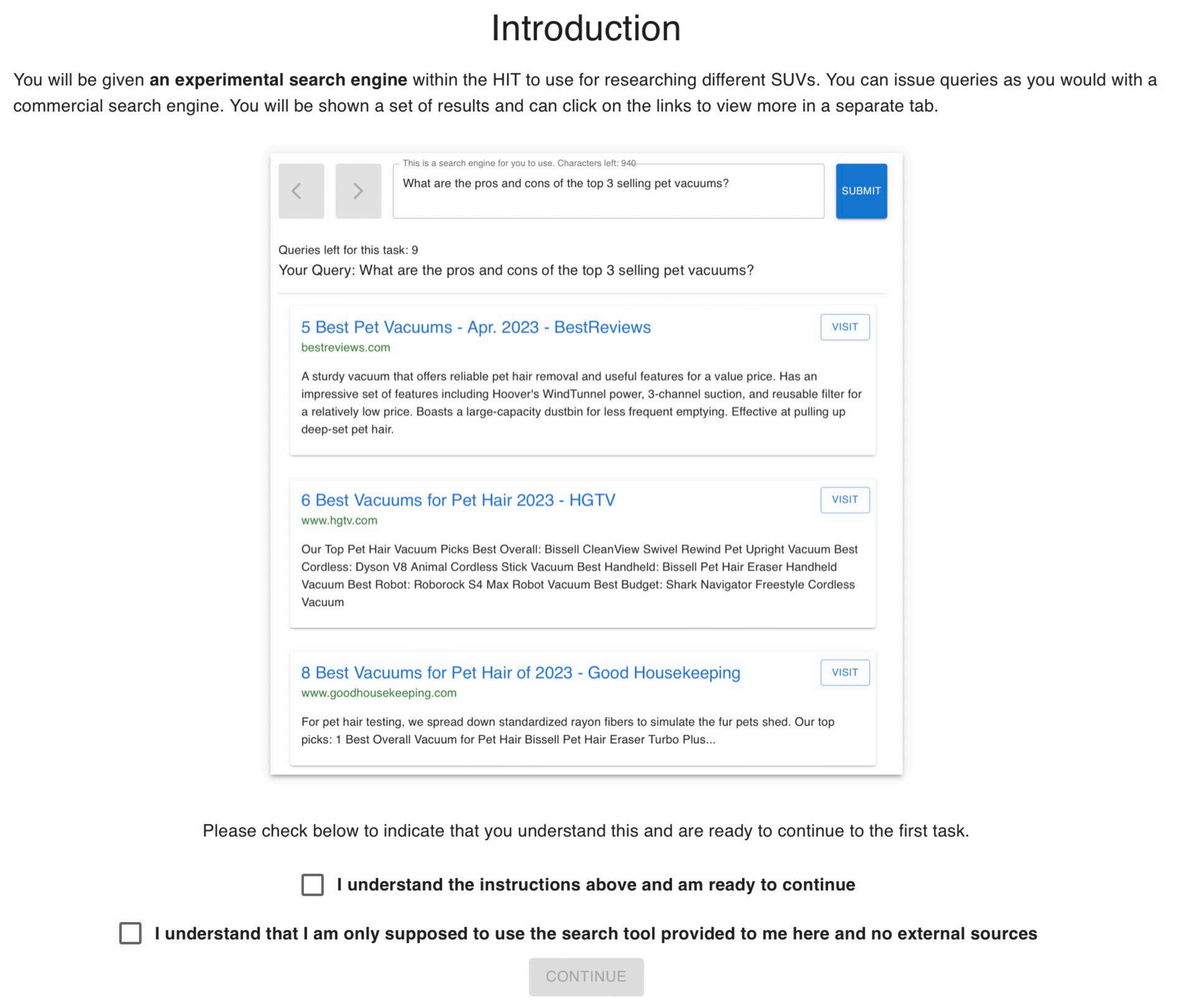
☐ **I understand the instructions above and am ready to continue**

☐ **I understand that I am only supposed to use the search tool provided to me here and no external sources**

Contact Us | Privacy & Cookies | Terms of Use | Trademarks | © 2023 Microsoft

Things we learned along the way

To be consistent, we showed people in the regular search condition how regular search looks



Introduction

You will be given **an experimental search engine** within the HIT to use for researching different SUVs. You can issue queries as you would with a commercial search engine. You will be shown a set of results and can click on the links to view more in a separate tab.

This is a search engine for you to use. Characters left: 940
What are the pros and cons of the top 3 selling pet vacuums?
SUBMIT

Queries left for this task: 9

Your Query: What are the pros and cons of the top 3 selling pet vacuums?

5 Best Pet Vacuums - Apr. 2023 - BestReviews
bestreviews.com

A sturdy vacuum that offers reliable pet hair removal and useful features for a value price. Has an impressive set of features including Hoover's WindTunnel power, 3-channel suction, and reusable filter for a relatively low price. Boasts a large-capacity dustbin for less frequent emptying. Effective at pulling up deep-set pet hair.

6 Best Vacuums for Pet Hair 2023 - HGTV
www.hgtv.com

Our Top Pet Hair Vacuum Picks Best Overall: Bissell CleanView Swivel Rewind Pet Upright Vacuum Best Cordless: Dyson V8 Animal Cordless Stick Vacuum Best Handheld: Bissell Pet Hair Eraser Handheld Vacuum Best Robot: Roborock S4 Max Robot Vacuum Best Budget: Shark Navigator Freestyle Cordless Vacuum

8 Best Vacuums for Pet Hair of 2023 - Good Housekeeping
www.goodhousekeeping.com

For pet hair testing, we spread down standardized rayon fibers to simulate the fur pets shed. Our top picks: 1 Best Overall Vacuum for Pet Hair Bissell Pet Hair Eraser Turbo Plus...

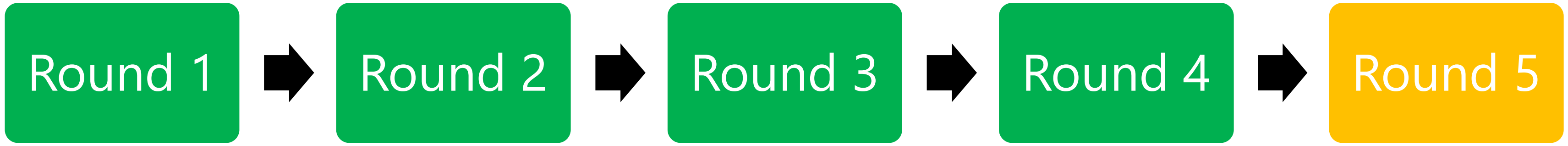
Please check below to indicate that you understand this and are ready to continue to the first task.

☐ **I understand the instructions above and am ready to continue**

☐ **I understand that I am only supposed to use the search tool provided to me here and no external sources**

One of these things is not like the others

- First four rounds are “routine”, in that search results should have similar accuracy in both conditions (4 random comparisons of 8 cars)



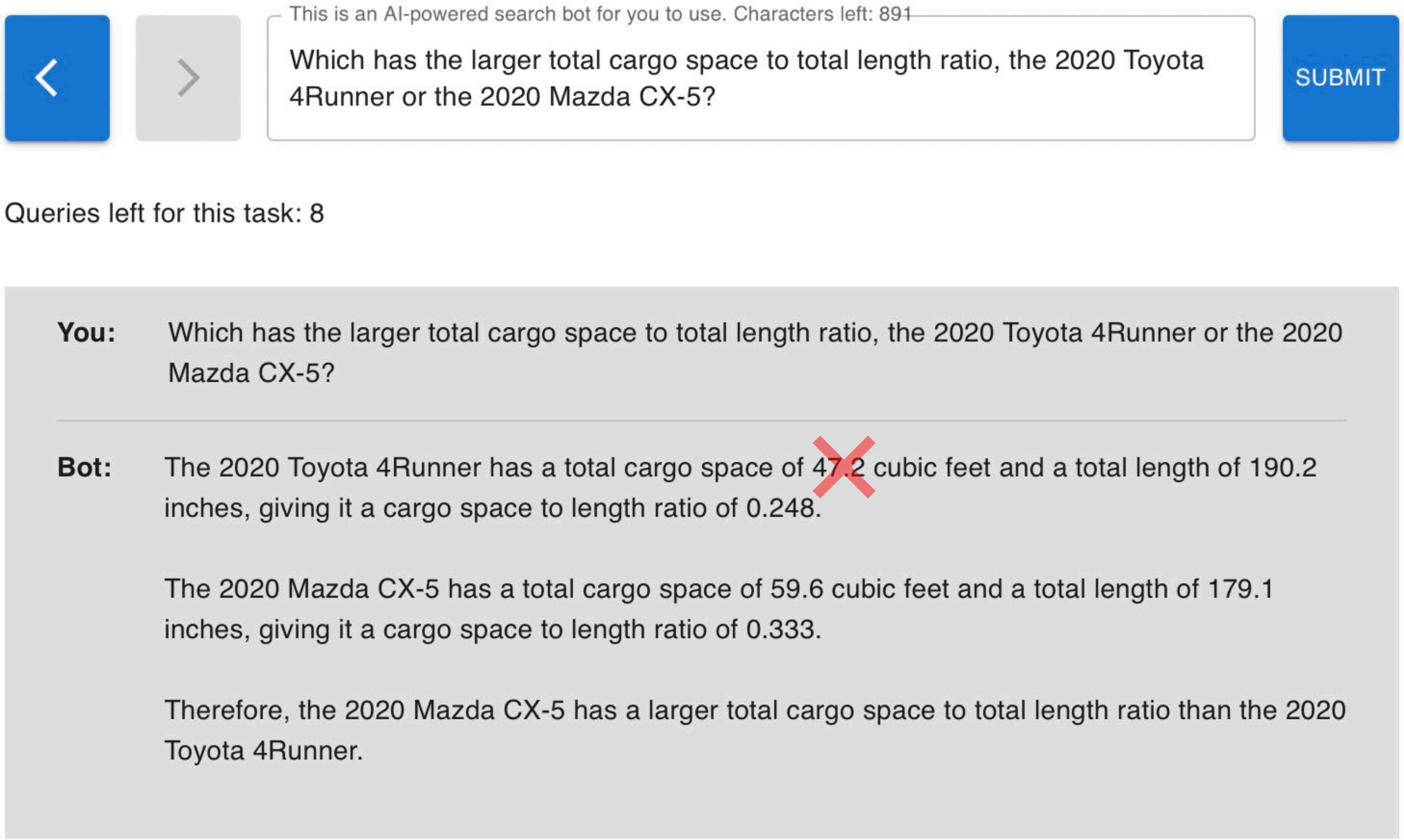
Round 1 → Round 2 → Round 3 → Round 4 → Round 5

- Last round is “challenging” for the LLM, yolked to a pair of cars for which the LLM tends to mistakenly report cargo space with seats up (instead of seats down)

80 Master Turkers, between subjects design

Round 5: Challenging for the LLM

Typically makes the *largest* vehicle look like the *smallest* one (actual value is 90 cu ft)


This is an AI-powered search bot for you to use. Characters left: 891
SUBMIT

Which has the larger total cargo space to total length ratio, the 2020 Toyota 4Runner or the 2020 Mazda CX-5?

Queries left for this task: 8

You: Which has the larger total cargo space to total length ratio, the 2020 Toyota 4Runner or the 2020 Mazda CX-5?

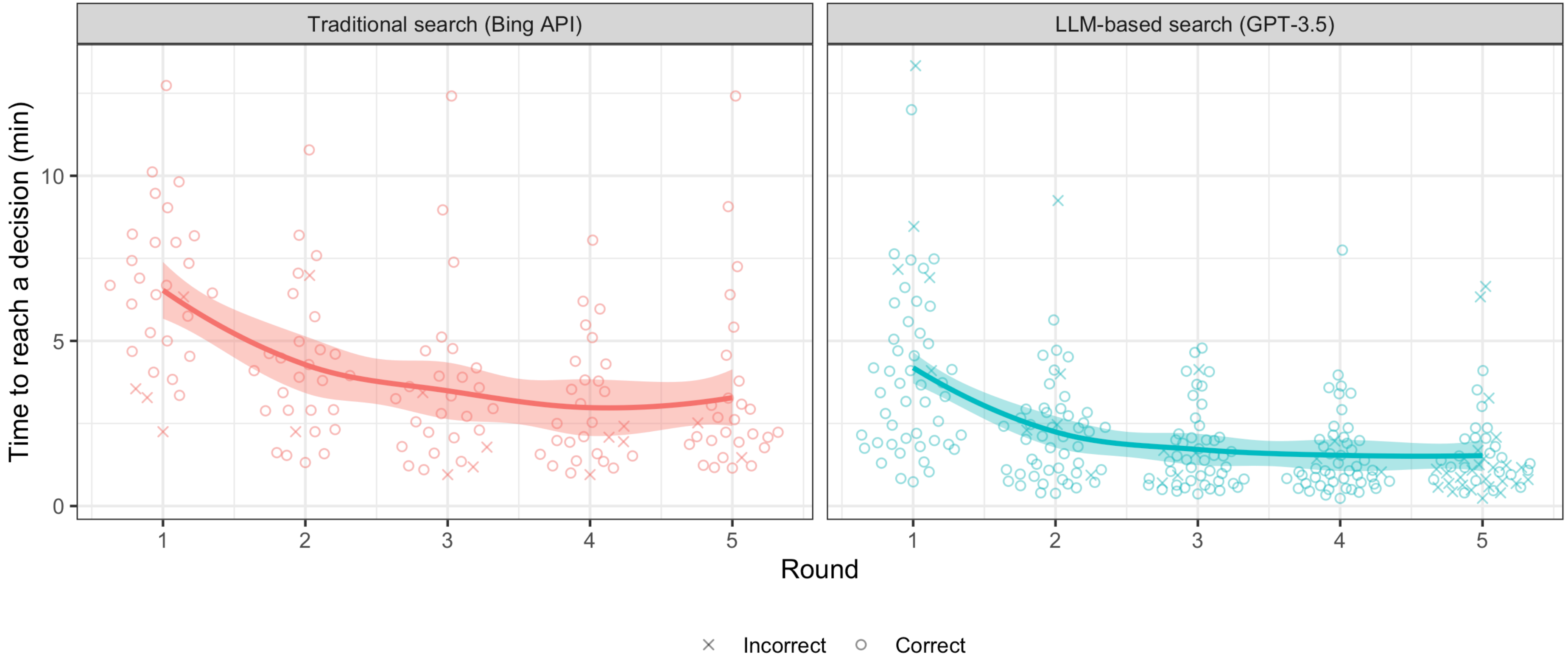
Bot: The 2020 Toyota 4Runner has a total cargo space of 47.2 cubic feet and a total length of 190.2 inches, giving it a cargo space to length ratio of 0.248.

The 2020 Mazda CX-5 has a total cargo space of 59.6 cubic feet and a total length of 179.1 inches, giving it a cargo space to length ratio of 0.333.

Therefore, the 2020 Mazda CX-5 has a larger total cargo space to total length ratio than the 2020 Toyota 4Runner.

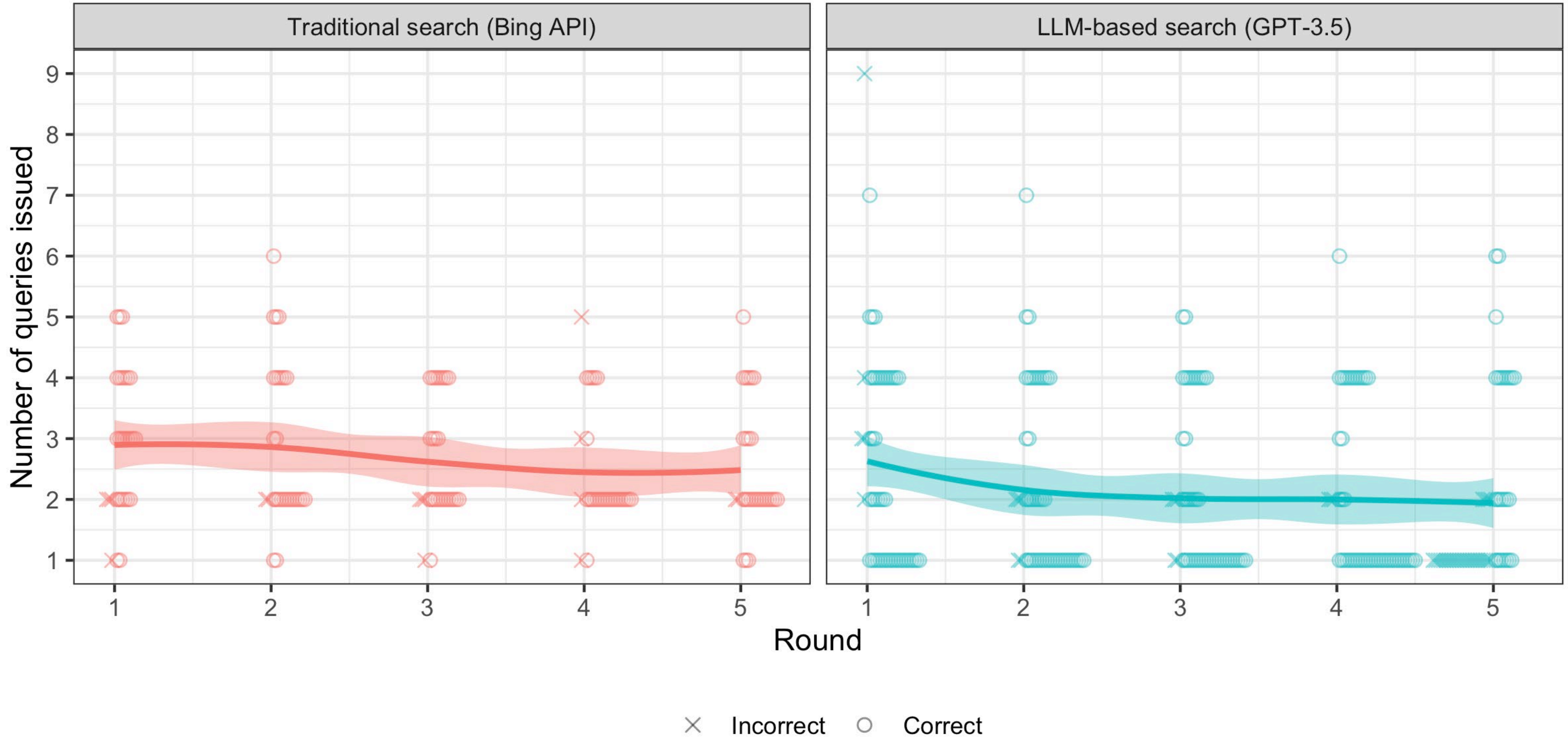
Results

Time taken



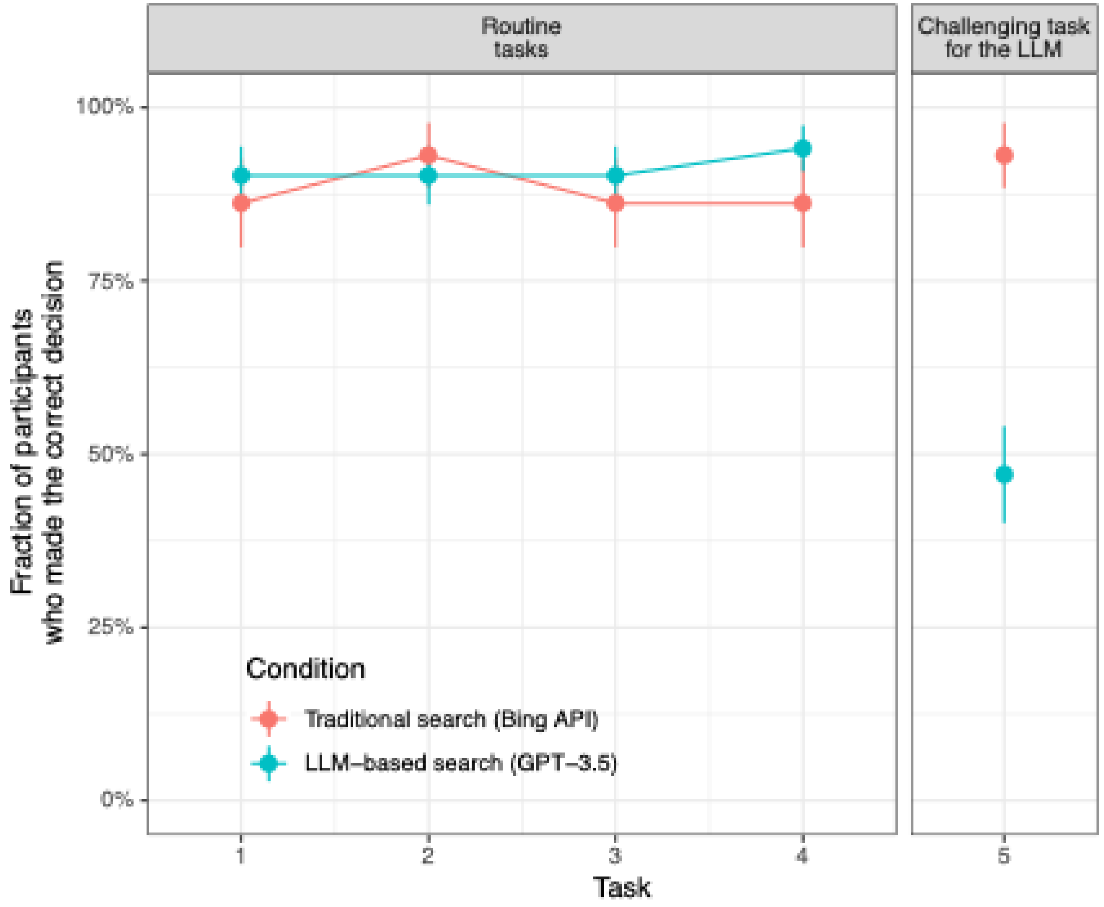
We see learning in both conditions, but LLM-based search takes roughly half the amount of time, on average

Number of queries



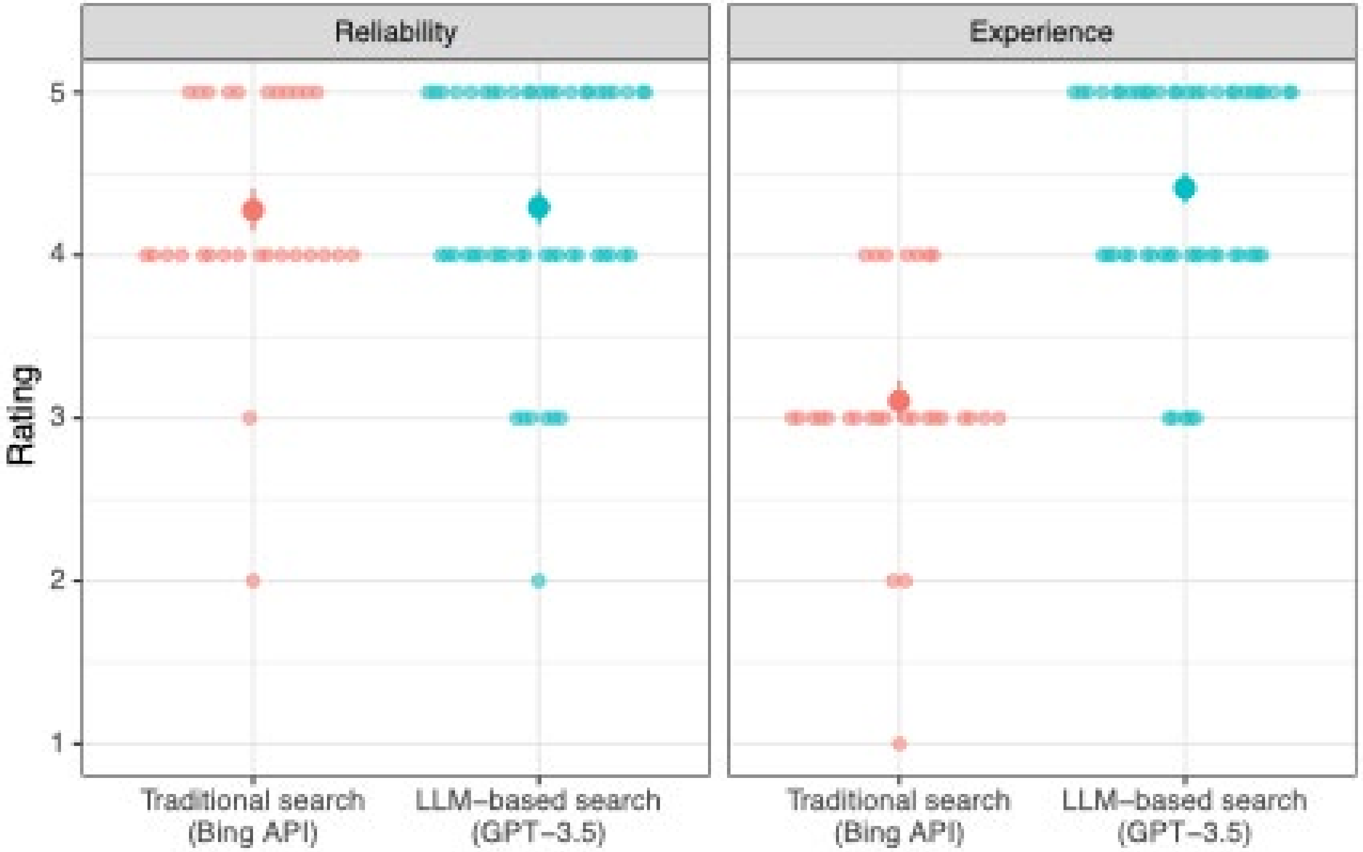
We see fewer than expected searches in traditional search,
but fewer still for LLM-based search

Accuracy



Accuracy is comparable between conditions for the first four tasks but much lower in the LLM-condition on the last task

Subjective evaluation



Participants rated both conditions similarly on reliability but preferred the LLM-based search experience

Participant feedback*

Mostly trusted the results, for better or worse

“The responses were fast and the bot seemed somewhat knowledgeable. Saved time that would have been needed to find the dimensions”

(but this participant got only 2 of 5 decisions correct!)

Noticed brittle results

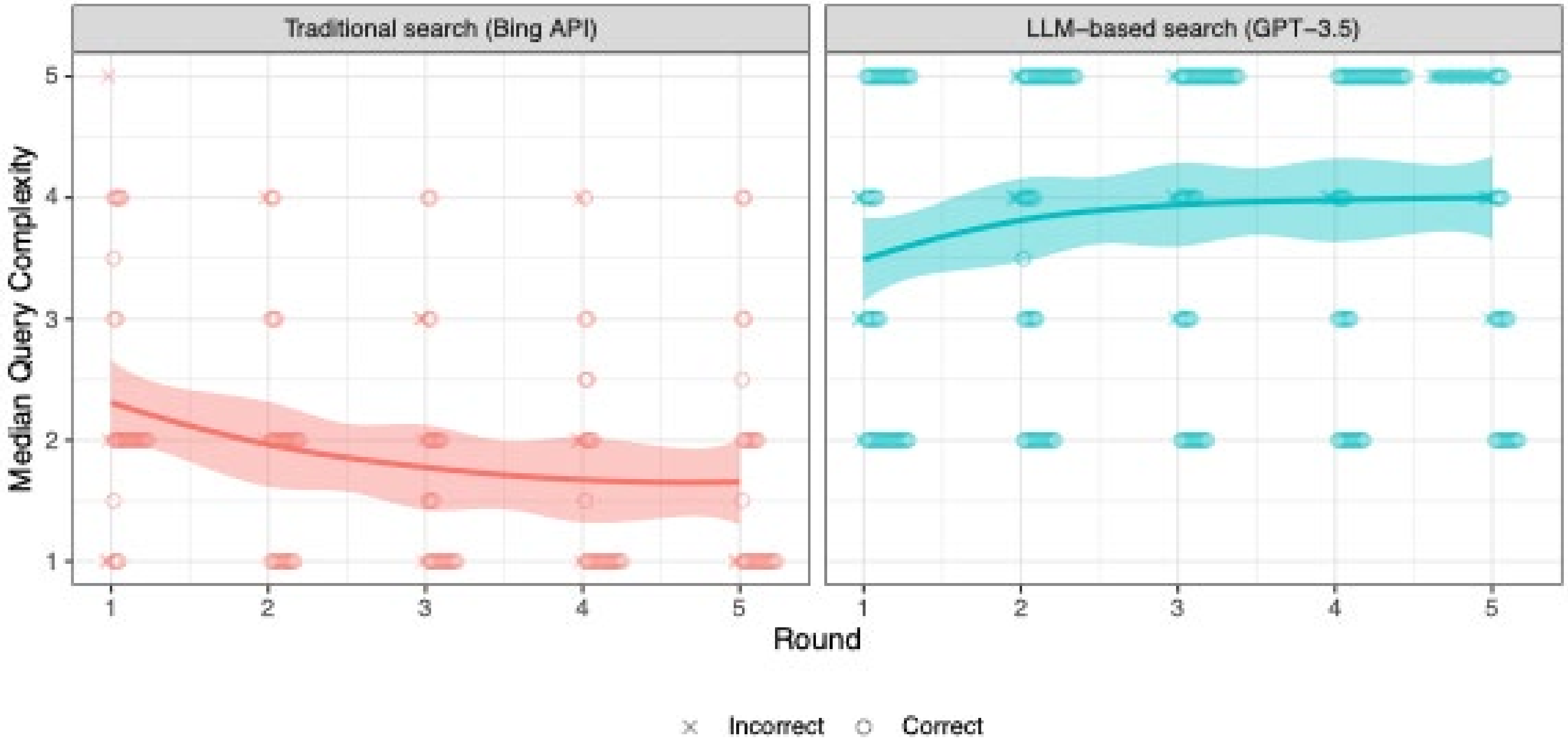
“The search tool was much easier to use and to understand compared to standard search engines. I was able to get an answer that was easy to understand and natural. The con was I needed to be careful in how I worded my question. One mistake could cause confusing results.”

Caught errors when cues were present

“It was nice that I could do a search for more than 1 piece of information at a time and compare it. The only downside was i wasn't sure how accurate the information i was being fed was. i believe it did say one car had more cargo space when the numbers it displayed said the opposite.”

***From pilot run with earlier model**

Complexity



Participants who issued less complex queries in round 5 with the LLM got more accurate results

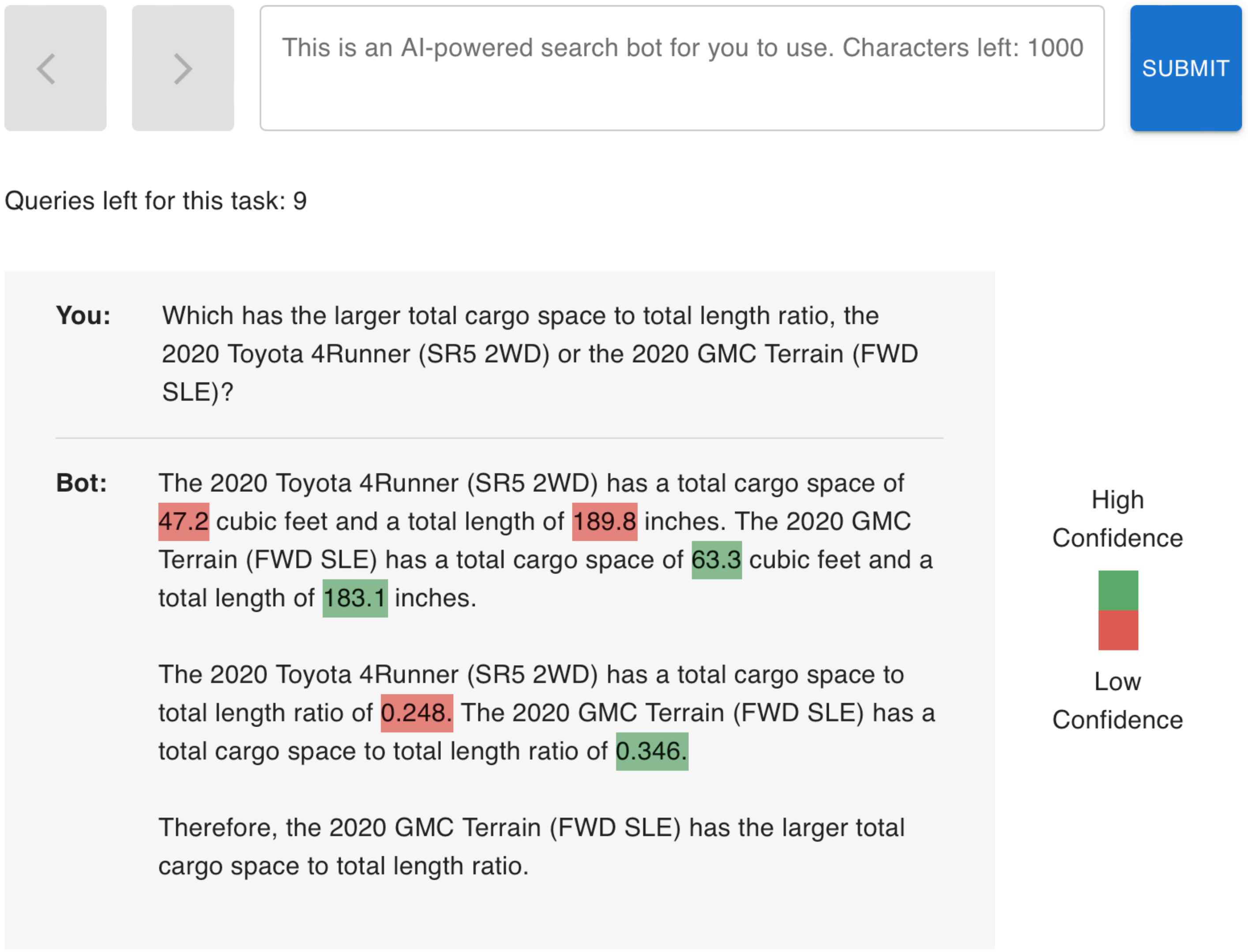
Displaying uncertainty

How do we compute and communicate calibrated uncertainty?

Colors are GPT-3 token probabilities

Red <= 50%, Green > 50%

Two color condition



This is an AI-powered search bot for you to use. Characters left: 1000

Queries left for this task: 9

You: Which has the larger total cargo space to total length ratio, the 2020 Toyota 4Runner (SR5 2WD) or the 2020 GMC Terrain (FWD SLE)?

Bot: The 2020 Toyota 4Runner (SR5 2WD) has a total cargo space of 47.2 cubic feet and a total length of 189.8 inches. The 2020 GMC Terrain (FWD SLE) has a total cargo space of 63.3 cubic feet and a total length of 183.1 inches.

The 2020 Toyota 4Runner (SR5 2WD) has a total cargo space to total length ratio of 0.248. The 2020 GMC Terrain (FWD SLE) has a total cargo space to total length ratio of 0.346.

Therefore, the 2020 GMC Terrain (FWD SLE) has the larger total cargo space to total length ratio.

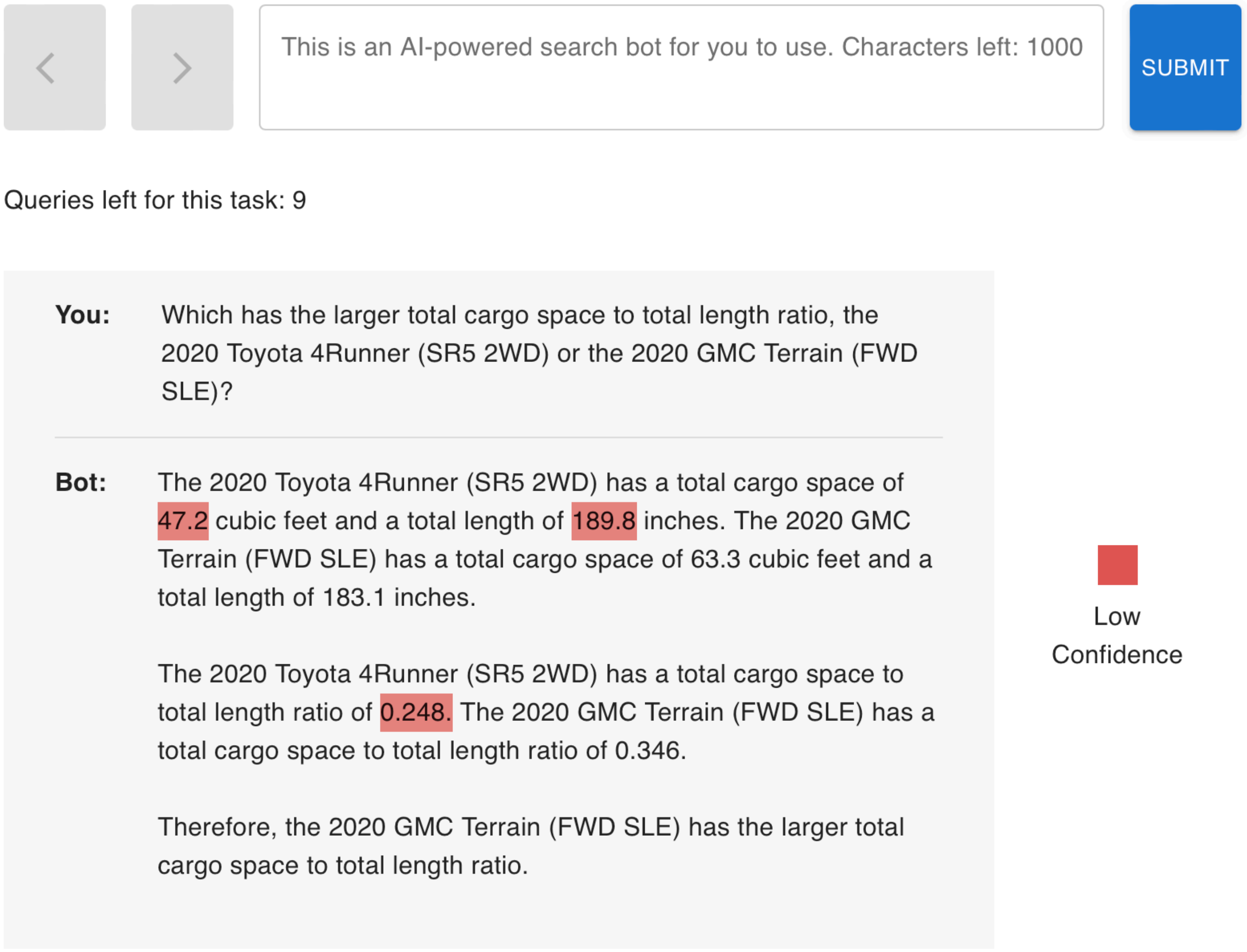
Displaying uncertainty

How do we compute and communicate calibrated uncertainty?

Colors are GPT-3 token probabilities

Red <= 50%

One color condition



This is an AI-powered search bot for you to use. Characters left: 1000

Queries left for this task: 9

You: Which has the larger total cargo space to total length ratio, the 2020 Toyota 4Runner (SR5 2WD) or the 2020 GMC Terrain (FWD SLE)?

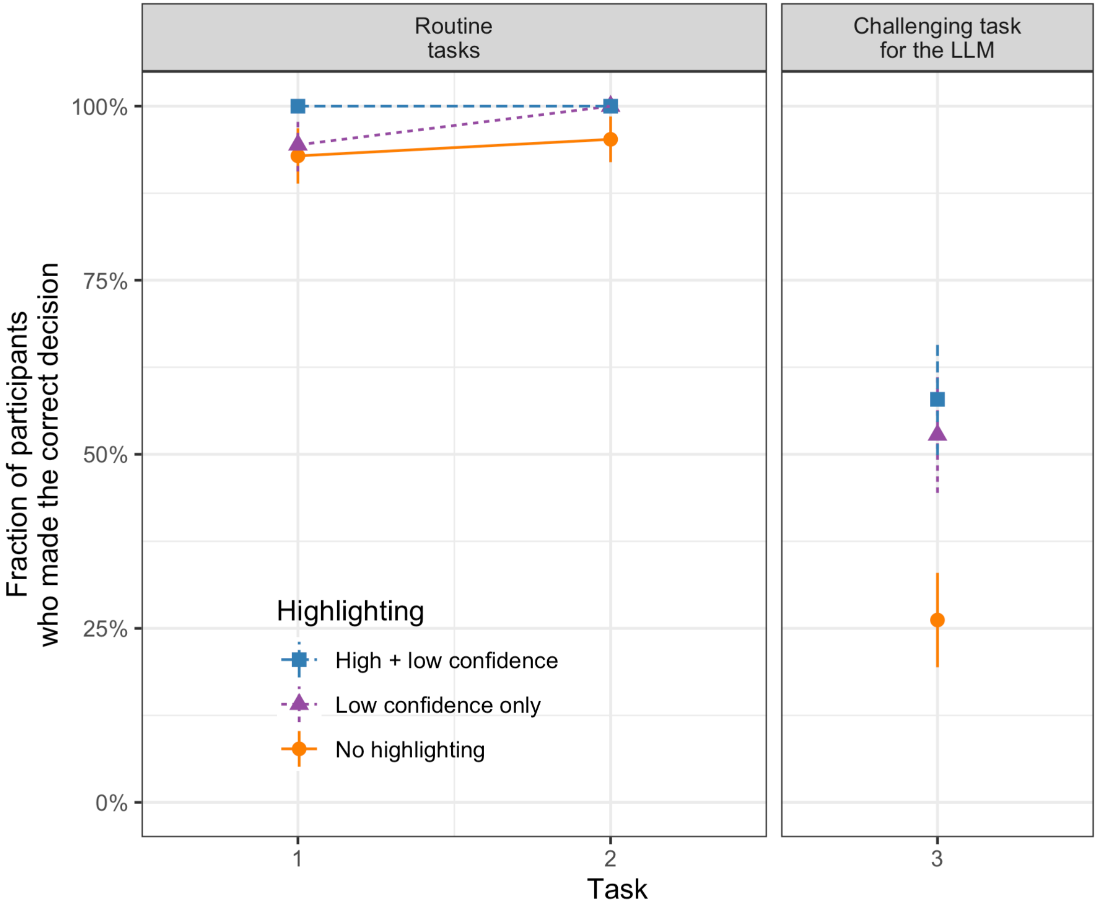
Bot: The 2020 Toyota 4Runner (SR5 2WD) has a total cargo space of 47.2 cubic feet and a total length of 189.8 inches. The 2020 GMC Terrain (FWD SLE) has a total cargo space of 63.3 cubic feet and a total length of 183.1 inches.

The 2020 Toyota 4Runner (SR5 2WD) has a total cargo space to total length ratio of 0.248. The 2020 GMC Terrain (FWD SLE) has a total cargo space to total length ratio of 0.346.

Therefore, the 2020 GMC Terrain (FWD SLE) has the larger total cargo space to total length ratio.

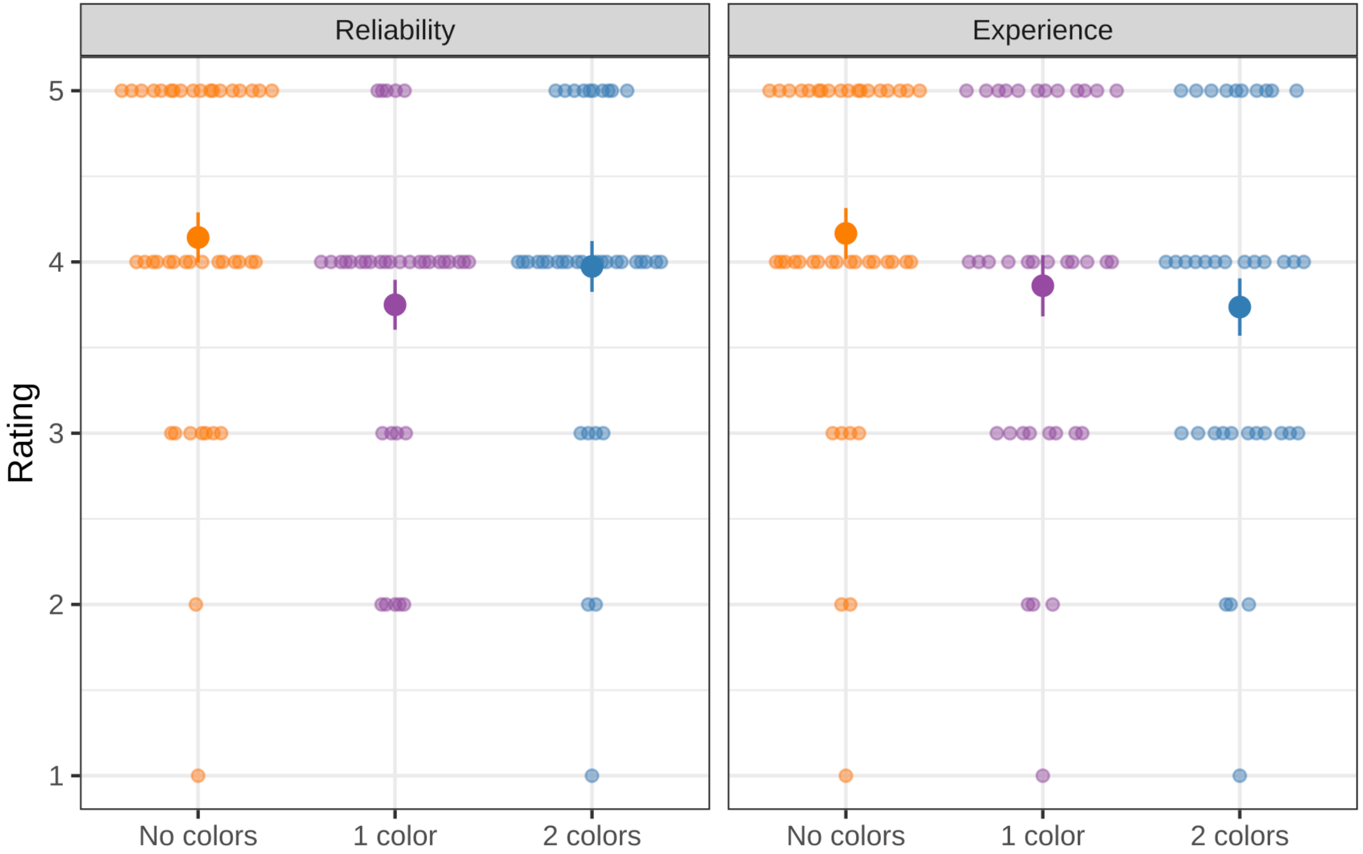
🟥 Low Confidence

Accuracy



Uncertainty highlighting doubled accuracy on the challenging task while maintaining comparable reliability/experience ratings

Subjective evaluation



Participants rated all conditions similarly on reliability and experience, with some preference for no uncertainty highlighting

Participant feedback*

Unclear how much to trust results without extra info

"They are very powerful, but you have to just assume that they are right, otherwise you save no time or effort at all because you have to research to verify."

(This participant was in the control condition)

Generally liked seeing confidence information

"I liked the fact that it highlighted text in a color that corresponded to a set confidence rating and gave me the answers that I needed quickly without giving me unnecessary information."

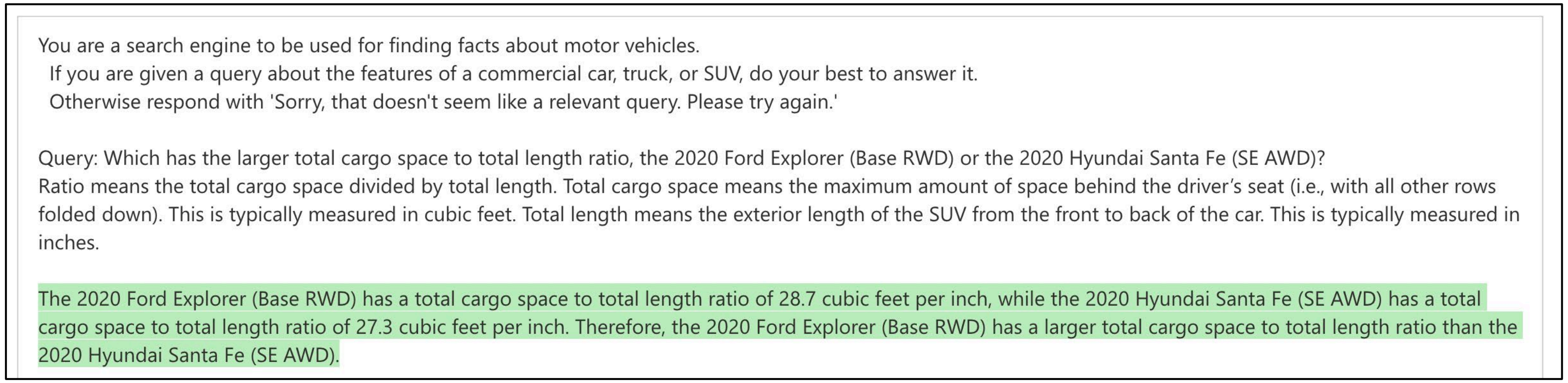
(Correctly) made them question reliability

"What was hard was the confidence level wasn't always good, so I was concerned I wasn't using accurate information to come up with the ratios."

***From pilot run with GPT-3 token probabilities**

Things we learned along the way

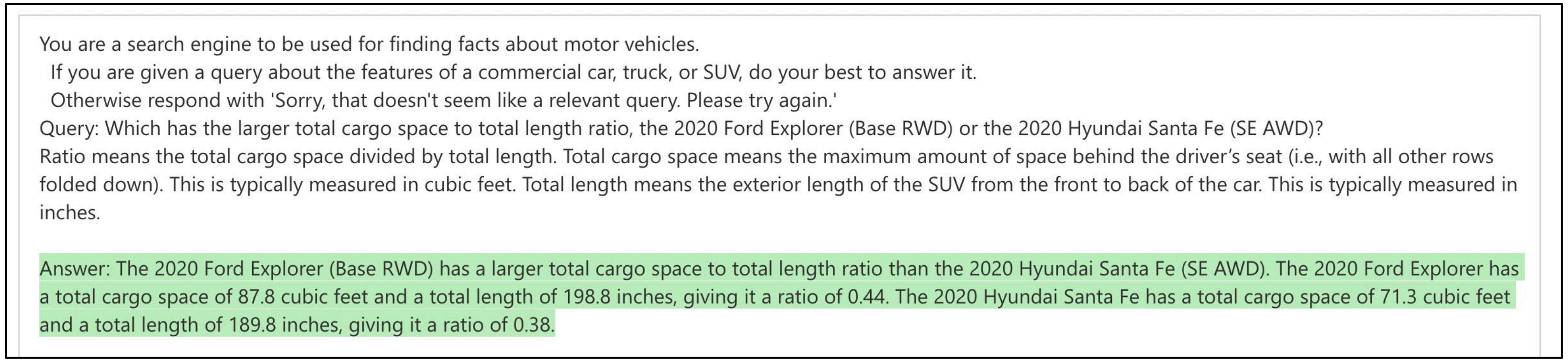
LLMs can be extremely sensitive to small changes wording (even spacing!), although newer models are better than older ones



You are a search engine to be used for finding facts about motor vehicles.
If you are given a query about the features of a commercial car, truck, or SUV, do your best to answer it.
Otherwise respond with 'Sorry, that doesn't seem like a relevant query. Please try again.'

Query: Which has the larger total cargo space to total length ratio, the 2020 Ford Explorer (Base RWD) or the 2020 Hyundai Santa Fe (SE AWD)?
Ratio means the total cargo space divided by total length. Total cargo space means the maximum amount of space behind the driver's seat (i.e., with all other rows folded down). This is typically measured in cubic feet. Total length means the exterior length of the SUV from the front to back of the car. This is typically measured in inches.

The 2020 Ford Explorer (Base RWD) has a total cargo space to total length ratio of 28.7 cubic feet per inch, while the 2020 Hyundai Santa Fe (SE AWD) has a total cargo space to total length ratio of 27.3 cubic feet per inch. Therefore, the 2020 Ford Explorer (Base RWD) has a larger total cargo space to total length ratio than the 2020 Hyundai Santa Fe (SE AWD).



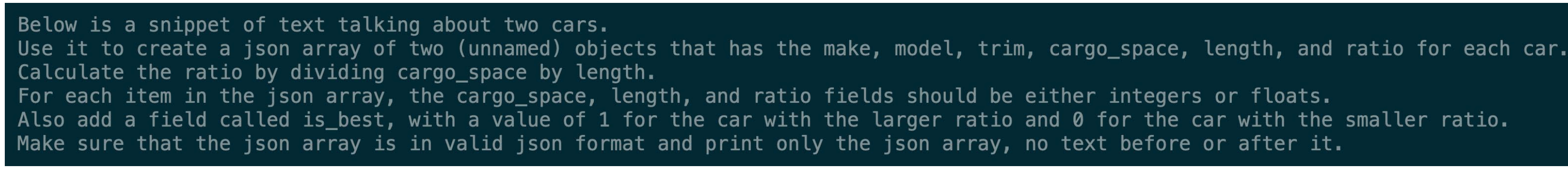
You are a search engine to be used for finding facts about motor vehicles.
If you are given a query about the features of a commercial car, truck, or SUV, do your best to answer it.
Otherwise respond with 'Sorry, that doesn't seem like a relevant query. Please try again.'
Query: Which has the larger total cargo space to total length ratio, the 2020 Ford Explorer (Base RWD) or the 2020 Hyundai Santa Fe (SE AWD)?
Ratio means the total cargo space divided by total length. Total cargo space means the maximum amount of space behind the driver's seat (i.e., with all other rows folded down). This is typically measured in cubic feet. Total length means the exterior length of the SUV from the front to back of the car. This is typically measured in inches.

Answer: The 2020 Ford Explorer (Base RWD) has a larger total cargo space to total length ratio than the 2020 Hyundai Santa Fe (SE AWD). The 2020 Ford Explorer has a total cargo space of 87.8 cubic feet and a total length of 198.8 inches, giving it a ratio of 0.44. The 2020 Hyundai Santa Fe has a total cargo space of 71.3 cubic feet and a total length of 189.8 inches, giving it a ratio of 0.38.

Things we learned along the way

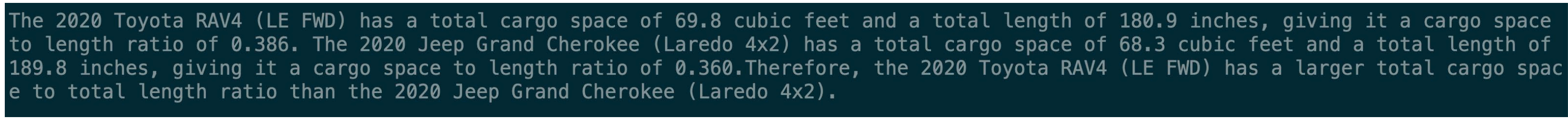
LLMs accelerated our own research and analyses, e.g., extremely useful for parsing and reformatting natural language data

Pre-prompt:



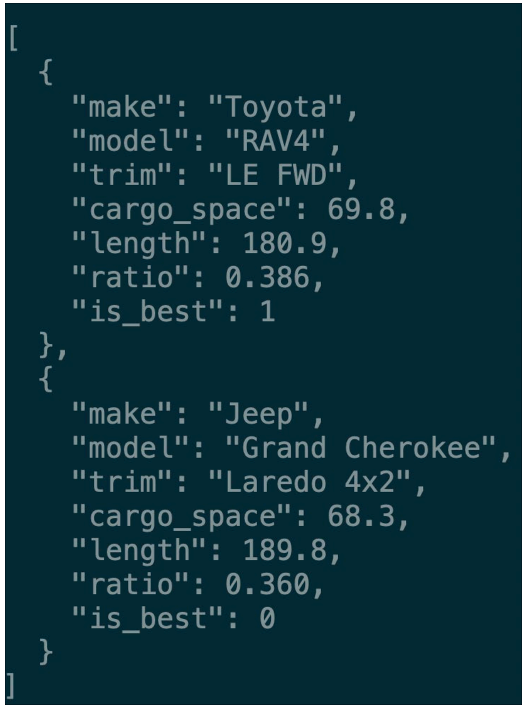
```
Below is a snippet of text talking about two cars.
Use it to create a json array of two (unnamed) objects that has the make, model, trim, cargo_space, length, and ratio for each car.
Calculate the ratio by dividing cargo_space by length.
For each item in the json array, the cargo_space, length, and ratio fields should be either integers or floats.
Also add a field called is_best, with a value of 1 for the car with the larger ratio and 0 for the car with the smaller ratio.
Make sure that the json array is in valid json format and print only the json array, no text before or after it.
```

Input:



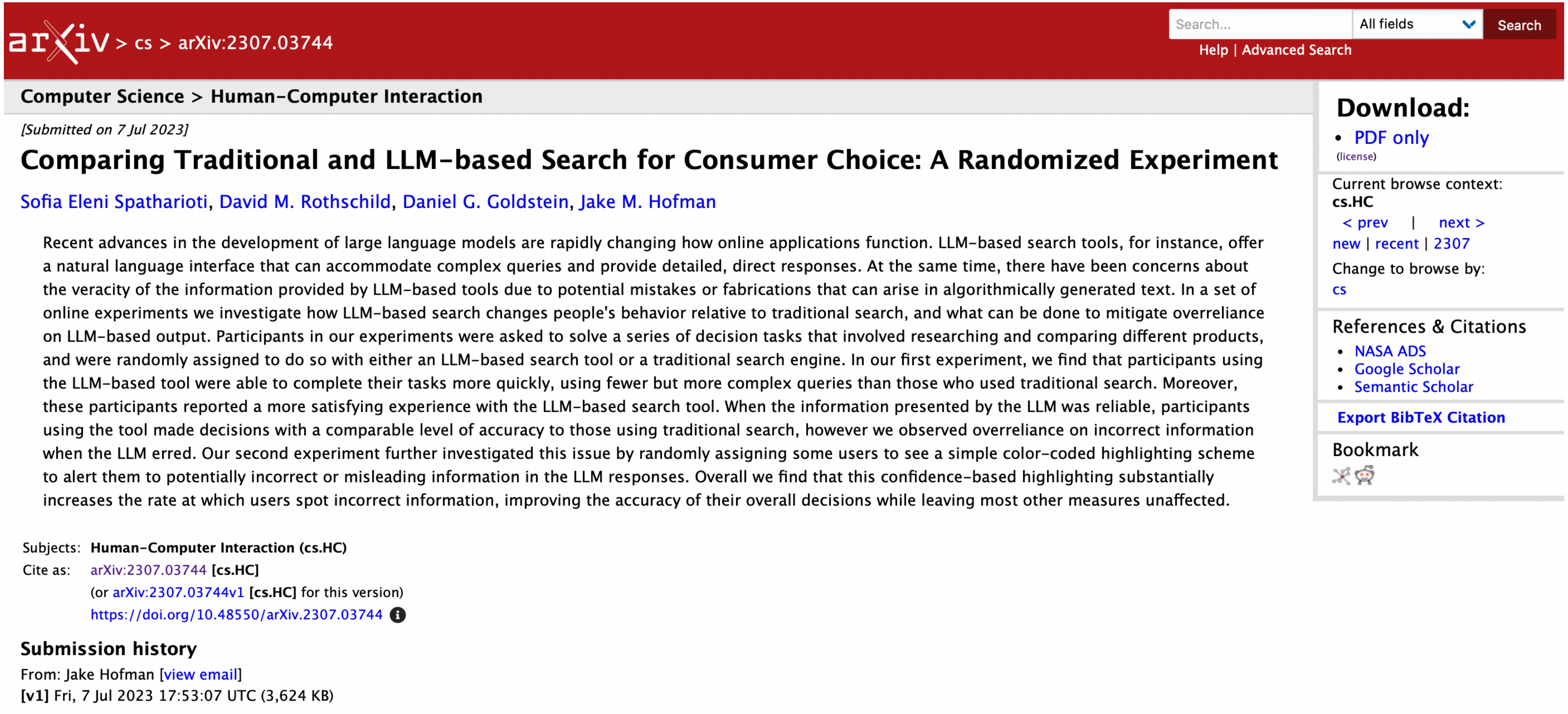
```
The 2020 Toyota RAV4 (LE FWD) has a total cargo space of 69.8 cubic feet and a total length of 180.9 inches, giving it a cargo space
to length ratio of 0.386. The 2020 Jeep Grand Cherokee (Laredo 4x2) has a total cargo space of 68.3 cubic feet and a total length of
189.8 inches, giving it a cargo space to length ratio of 0.360.Therefore, the 2020 Toyota RAV4 (LE FWD) has a larger total cargo spac
e to total length ratio than the 2020 Jeep Grand Cherokee (Laredo 4x2).
```

Output:



```
[
  {
    "make": "Toyota",
    "model": "RAV4",
    "trim": "LE FWD",
    "cargo_space": 69.8,
    "length": 180.9,
    "ratio": 0.386,
    "is_best": 1
  },
  {
    "make": "Jeep",
    "model": "Grand Cherokee",
    "trim": "Laredo 4x2",
    "cargo_space": 68.3,
    "length": 189.8,
    "ratio": 0.360,
    "is_best": 0
  }
]
```

Pre-print



Computer Science > Human-Computer Interaction

[Submitted on 7 Jul 2023]

Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment

Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, Jake M. Hofman

Recent advances in the development of large language models are rapidly changing how online applications function. LLM-based search tools, for instance, offer a natural language interface that can accommodate complex queries and provide detailed, direct responses. At the same time, there have been concerns about the veracity of the information provided by LLM-based tools due to potential mistakes or fabrications that can arise in algorithmically generated text. In a set of online experiments we investigate how LLM-based search changes people's behavior relative to traditional search, and what can be done to mitigate overreliance on LLM-based output. Participants in our experiments were asked to solve a series of decision tasks that involved researching and comparing different products, and were randomly assigned to do so with either an LLM-based search tool or a traditional search engine. In our first experiment, we find that participants using the LLM-based tool were able to complete their tasks more quickly, using fewer but more complex queries than those who used traditional search. Moreover, these participants reported a more satisfying experience with the LLM-based search tool. When the information presented by the LLM was reliable, participants using the tool made decisions with a comparable level of accuracy to those using traditional search, however we observed overreliance on incorrect information when the LLM erred. Our second experiment further investigated this issue by randomly assigning some users to see a simple color-coded highlighting scheme to alert them to potentially incorrect or misleading information in the LLM responses. Overall we find that this confidence-based highlighting substantially increases the rate at which users spot incorrect information, improving the accuracy of their overall decisions while leaving most other measures unaffected.

Subjects: **Human-Computer Interaction (cs.HC)**
Cite as: arXiv:2307.03744 **[cs.HC]**
(or arXiv:2307.03744v1 **[cs.HC]** for this version)
https://doi.org/10.48550/arXiv.2307.03744

Submission history

From: Jake Hofman [view email]
[v1] Fri, 7 Jul 2023 17:53:07 UTC (3,624 KB)

https://arxiv.org/abs/2307.03744

Activity

In your group, use https://chat.openai.com to

- Take the role of a consumer in a product area you know very well and ask for advice on how to choose

Debrief

What did you think of the buying guide information it provided?

What would be a fair way to advertise on this information?

Activity

In your group, use https://chat.openai.com to

- Take the role of a consumer and have the AI play the role of a salesperson for your company.
    - Ask it which of your products it would recommend for you
    - See if it answers product information questions correctly

Debrief

What did you think of its ability as a salesperson?

Thank you!

- Dan Goldstein

  dgg@microsoft.com

- With:

  Jake Hofman, David Rothschild, Sophie Spathariotti, & Harsh Kumar