

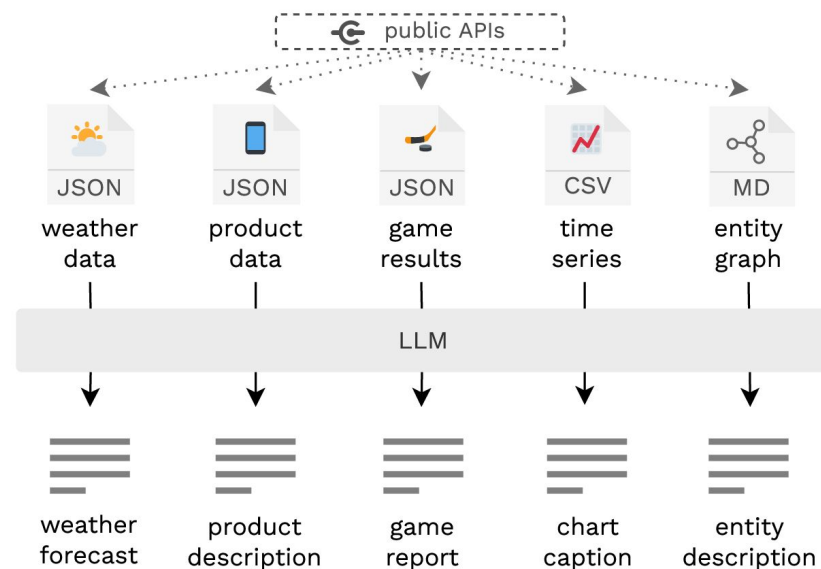
How to get the model to do what we want: Generation, evaluation, web navigation

Zdeněk Kasner



Generating texts with open LLMs

- **data-based reports** (similar to the weather forecast assignment)
- three **open LLMs** (Llama2, Mistral, Zephyr) + ChatGPT (GPT-3.5)



- **zero-shot prompting**

- only prompt + initial tokens
- no examples (context too long), no finetuning (missing references)
- reports very fluent and natural, but with **many semantic errors**

Prompt

```
Based on the given data:
```

```
```\n
```

```
{DATA}
```

```
```\n
```

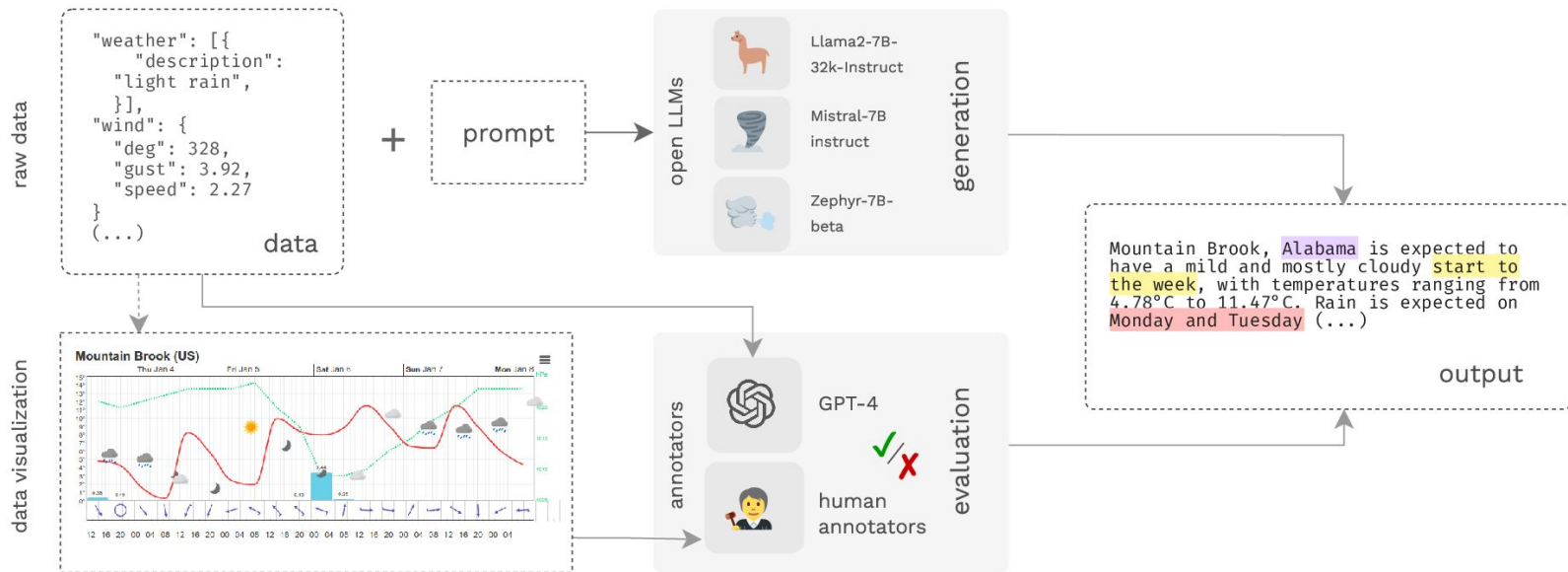
```
Your task is to write a brief, fluent, and coherent single-paragraph {output_type} in natural language. The text should be balanced and neutral. Make sure that all the facts mentioned in the text can be derived from the input data, do *not* add any extra information.
```

Start with

```
Sure! Here is the {output_type}:
```

```
```\n
```

- **evaluation** (GPT-4 → next few slides, human annotators)



# Evaluating semantic accuracy with GPT-4

- automatic annotations of **semantic errors** in the system outputs

---

Error	Description
INCORRECT	The fact in the text contradicts the data.
NOT_CHECKABLE	The fact in the text cannot be checked given the data.
MISLEADING	The fact in the text is misleading in the given context.
OTHER	The text is problematic for another reason, e.g. grammatically or stylistically incorrect, irrelevant, or repetitive.

---

<b>Example</b> <i>data</i> <i>text</i>	<b>Nokia 3310</b>   <i>color</i> : black, blue, grey   <i>display</i> : 320x240px Nokia 3310 is produced in Finland and features a 320x320 display. It is available in black color. The data seem to provide only partial information about the phone.
----------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

---

- **very explicit prompt + system message**
- **JSON outputs**
- **GPT-4 is the only model which is good at following fine-grained instructions (as of now)**

## System Message

You are an expert data-to-text error annotation system. You understand structured data and you can correctly operate with units and numerical values. You are designed to output token-level annotations in JSON.

## Prompt

```
Given the data:
...
data
...
Annotate all the errors in the following text:
...
text
...
Output the errors as a JSON list "errors" in which each object contains fields "reason", "text",
and "type". The value of "text" is the text of the error. The value of "reason" is the reason
for the error. The value of "type" is one of 0, 1, 2, 3 based on the following list:
- 0: Incorrect fact: The fact in the text contradicts the data.
- 1: Not checkable: The fact in the text cannot be checked in the data.
- 2: Misleading: The fact in the text is misleading in the given context.
- 3: Other: The text is problematic for another reason, e.g. grammatically or stylistically
incorrect, irrelevant, or repetitive.
The list should be sorted by the position of the error in the text.
Example:
data:
...
[["Aditi Bhagwat", "occupation", "television actor"], ["Aditi Bhagwat", "date of birth", "18
January 1981"]]
...
text:
...
Aditi Bhagwat, born on January 18, 1991, used to be a popular Indian television actor. The data
comes from a knowledge graph.
...
output:
... "errors": [{"reason": "The data mentions that the actor was born on 1981", "text": "1991",
"type": 0, "reason": "Misleadingly suggests that the actor is not alive", "text": "used to
be", "type": 2, "reason": "Popularity is not mentioned in the data", "text": "popular", "type": 1,
"reason": "Nationality is not mentioned in the data", "text": "Indian", "type": 1, "reason": "The
note is superfluous", "text": "The data comes from a knowledge graph.", "type": 3}
...
Note that some details may not be mentioned in the text: do not count omissions as errors. Also
do not be too strict: some facts can be less specific than in the data (rounded values, shortened
or abbreviated text, etc.), do not count these as errors. If there are no errors in the text,
"errors" will be an empty list.
```



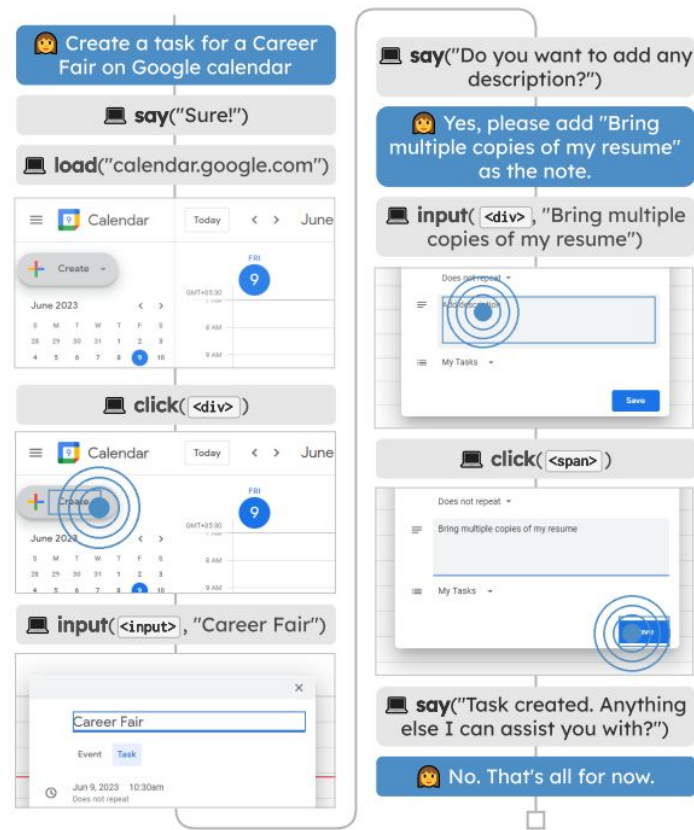
# Evaluating semantic accuracy with GPT-4

- moderate correlations with human annotators
- providing “reasons” for the annotations

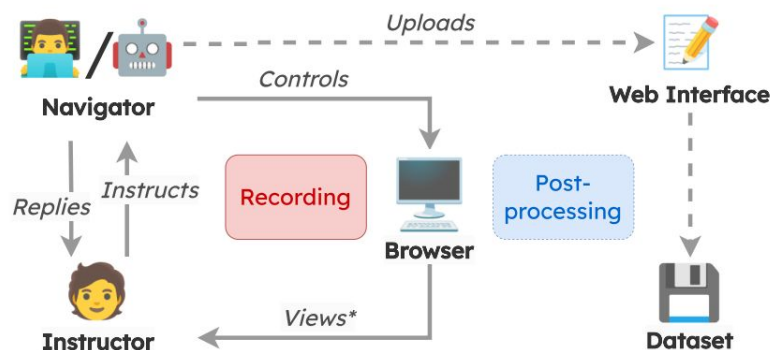
Model	Human annotations ( $\mathcal{E}_{hum}$ )	GPT-4 annotations ( $\mathcal{E}_{gpt}$ )
Llama2	The weather in Niemegek, Germany, will be cloudy and rainy for the next few days. On January 3rd, the temperature will be around 9.55°C, with a feels-like temperature of 6.59°C. The weather will be light rain, with a wind speed of 6.24 m/s and a gust of 12.1 m/s. The rain will last for 3 hours, and the pressure will be 991 hPa.	The weather in Niemegek, Germany, will be cloudy and rainy for the next few days. On January 3rd, the temperature will be around 9.55°C, with a feels-like temperature of 6.59°C. The weather will be light rain, with a wind speed of 6.24 m/s and a gust of 12.1 m/s. The rain will last for 3 hours, and the pressure will be 991 hPa.
Mistral	From January 3rd to January 7th, the city of Niemegek in Germany will experience a range of weather conditions. On January 3rd, the temperature will be 9.55°C with a feels like temperature of 6.59°C. There will be light rain with a 3-hour precipitation of 1.06mm. The wind speed will be 6.24m/s with a gust of 12.1m/s. On January 4th, the temperature will be 8.2°C with a feels like temperature of 5.64°C. There will be light rain with a 3-hour precipitation of 1.94mm. The wind speed will be 4.31m/s with a gust of 11.6m/s. On January 5th, the temperature will be 6.37°C with a feels like temperature of 2.25°C. There will be light rain with a 3-hour precipitation of 1.13mm. The wind speed will be 7.07m/s with a gust of 15.51m/s. On January 6th, the temperature will be 2.17°C with a feels like temperature of -3.55°C. There will be overcast clouds with a 3-hour precipitation of 0.31mm. The wind speed will be 3.54m/s with a gust of 7.11m/s. On January 7th, the temperature will be -0.89°C with a feels like temperature of -0.89°C. There will be overcast clouds with a 3-hour precipitation of 0.31mm. The wind speed will be 0.63m/s with a gust of 0.7m/s.	From January 3rd to January 7th, the city of Niemegek in Germany will experience a range of weather conditions. On January 3rd, the temperature will be 9.55°C with a feels like temperature of 6.59°C. There will be light rain with a 3-hour precipitation of 1.06mm. The wind speed will be 6.24m/s with a gust of 12.1m/s. On January 4th, the temperature will be 8.2°C with a feels like temperature of 5.64°C. There will be light rain with a 3-hour precipitation of 1.94mm. The wind speed will be 4.31m/s with a gust of 11.6m/s. On January 5th, the temperature will be 6.37°C with a feels like temperature of 2.25°C. There will be light rain with a 3-hour precipitation of 1.13mm. The wind speed will be 7.07m/s with a gust of 15.51m/s. On January 6th, the temperature will be 2.17°C with a feels like temperature of -3.55°C. There will be overcast clouds with a 3-hour precipitation of 0.31mm. The wind speed will be 3.54m/s with a gust of 7.11m/s. On January 7th, the temperature will be -0.89°C with a feels like temperature of -0.89°C. There will be overcast clouds with a 3-hour precipitation of 0.31mm. The wind speed will be 0.63m/s with a gust of 0.7m/s.
Zephyr	Over the next five days, the weather in Niemegek, Germany will be characterized by fluctuating temperatures and varying precipitation. On Thursday, January 3, light rain and a low temperature of 9.55°C can be expected, with a chance of rain continuing into Friday. Snowfall is possible on Friday, with a low temperature of -0.89°C. Saturday will be overcast with a low temperature of 0.03°C, and light rain is forecasted for Saturday evening. Sunday will be overcast as well, with a low temperature of -2.23°C. Monday will see a drop in temperatures, with a low of -4.07°C and scattered clouds. Temperatures will continue to fall throughout the week, with lows of -5.19°C and -9.51°C on Tuesday and Wednesday, respectively. Wind speeds will vary from 3.1 to 15.51 meters per second, with gusts reaching up to 15.51 meters per second on Monday.	Over the next five days, the weather in Niemegek, Germany will be characterized by fluctuating temperatures and varying precipitation. On Thursday, January 3, light rain and a low temperature of 9.55°C can be expected, with a chance of rain continuing into Friday. Snowfall is possible on Friday, with a low temperature of -0.89°C. Saturday will be overcast with a low temperature of 0.03°C, and light rain is forecasted for Saturday evening. Sunday will be overcast as well, with a low temperature of -2.23°C. Monday will see a drop in temperatures, with a low of -4.07°C and scattered clouds. Temperatures will continue to fall throughout the week, with lows of -5.19°C and -9.51°C on Tuesday and Wednesday, respectively. Wind speeds will vary from 3.1 to 15.51 meters per second, with gusts reaching up to 15.51 meters per second on Monday.

# Conversational web navigation with LLMs

- LLM acting as an agent **controlling a web browser**
- the model needs to interact with the user in natural language



- no data → collecting a **large-scale dataset** with professional human annotators

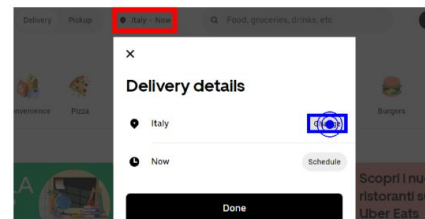


- working with  
real-world  
websites



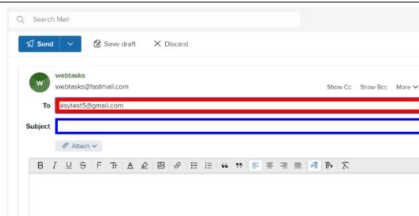
**C1:** Instructor wants Navigator to open a specific tab on a News site, i.e., "Sportsday on 28 May 2023 at 4.15 AM".

**GPT-4V (R)** clicks on an incorrect (3:30AM) tab, even though the instructor requested a different time (4:15AM).  
**LLaMA-WL (B)** clicks on the correct 4:15AM tab.



**C2:** Instructor requests the location on a food delivery site to be set to Las Vegas. The Delivery details window is already open.

**GPT-4V (R)** attempts to exit the Delivery details page and reopen it, which could potentially lead to a loop.  
**LLaMA-WL (B)** correctly clicks on the *Change* button.

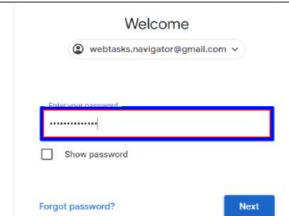


**T1:** Compose a "Invitation to Collaboration" email.

**Reference (B):** Invitation to Collaboration  
**GPT-4V (R):** Leon Tales  
**LLaMA-WL (B):** Invitation to Collaboration

**S1:** Instructor says "Like & Bookmark this listing."

**Reference:** Alright  
**LLaMA-WL:** Alright  
**GPT-4V:** Understood. If you need further assistance, just let me know.



**T2:** Open Google translate and sign in using the following credentials: [email] [password]

**Reference (B):** [password]  
**GPT-4V (R):** [email]  
**LLaMA-WL (B):** [password]

**S2:** Instructor says " Please share the link."

**Reference:** Alright.  
**LLaMA-WL:** Okay.  
**GPT-4V:** Here is the link to the discussion: [incorrect link]

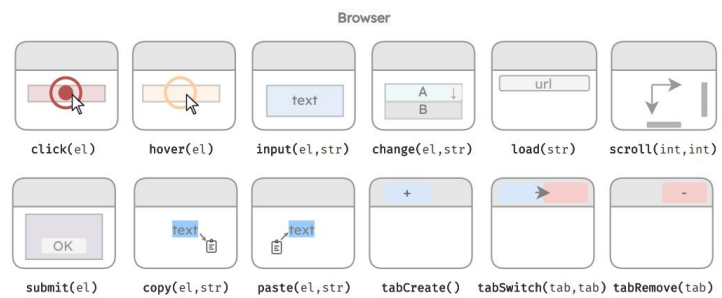
- explicit prompt, explicit action space

```
(html(body(div class="container"(div class="row"(div class="col hd...tems-center"(div class="hdr...container"(a class="rc-link"
→ onclick="if (!...Flyout() data-webtasks-id="7697...-4cf1"(span class="text" data-webtasks-id="6c7f...-4dce"EXPLORE)(i class="fa
→ ency-down" data-webtasks-id="e7...-48a5"))(div class="rc-flyout")))))(div (div class="dialog-off...main-canvas"(div
→ class="homepage"(div style="background-image: ...png!);" class="ency-loaded"(div class="ency-loaded mask-hero")(h4
→ data-webtasks-id="1ea...d-4e30"The World's #1 Online Encyclopedia)(div class="clear-both hero"(div class="ency-hero-search"(form
→ action="https://www.../gsearch" method="get" data-webtasks-id="c7f...-4ab2"(div class="js...o-label"
→ data-webtasks-id="8d8...97-477a"(span class="field-prefix"(input class="button j... form-submit" type="submit" value="")
→ (input title=" class="searchbox form-search form-input" placeholder="What do you want to learn today?" type="search" name="q"
→ value="" size="15" maxlength="128" data-webtasks-id="67a2...-4fab" spellcheck="false"(span class="field-suffix"(i class="fa
→ ency-close")))(div class="form-actions...-wrapper"(input class="button j... form-submit" type="submit" value="Search"
→ data-webtasks-id="fedfb...-42b3")))(div class="clear-both hero footer-copy"(a href="/about" data-webtasks-id="bf33...44f0"Read
→ more) about our content and why so many people love it.)))))((div class="adth...ive-sticky" style="min-height: 90px;"
→ closable="true"(div style="border: 0pt none;")(span class="adthrive-close" data-webtasks-id="0ff...-4c2a"x))))))
```

Above are the pruned HTML contents of the page. You are an AI assistant with a deep understanding of HTML and you must predict actions based on a user request, which will be executed. Use one of the following, replacing [] with an appropriate value:  
→ change(value=[str], uid=[str]); click(uid=[str]); load(url=[str]); say(speaker="navigator", utterance=[str]); scroll(x=[int], y=[int]); submit(uid=[str]); text\_input(text=[str], uid=[str]);  
The user's first and last 4 utterances are: [00:05] Hello ;  
Viewport size: 746h x 1536w ;  
Only the last 5 turns are provided.

```
Here are the top candidates for this turn: (uid = 67e2a5fb-8b1d-41a0) [[tag]] input [[xpath]] /html/body/...[1]/input [[bbox]] x=419.6
→ y=461.0 width=477.6 height=89.6 [[Attributes]] title=' values: .. want to learn today?'
(uid = fedfb512-949e-42b3) [[tag]] input [[xpath]] /html/body/...[2]/input [[bbox]] x=915.6 y=461.0 width=185.6 height=89.6
→ [[Attributes]] type='submit' ...mit form-submit'
(uid = 67fbc11c-0949-4ab2) [[tag]] form [[xpath]] /html/body/...div[3]/form [[bbox]] x=419.6 y=461.0 width=680 height=88 [[Attributes]]
→ method='get' data...com/gsearch' [[children]] div div
(uid = 6c7fef1f-f640-4dce) [[tag]] span [[xpath]] /html/body/.../a/span [[text]] EXPLORE [[bbox]] x=1240.5 y=28.6 width=54.1 height=30
→ [[Attributes]] class='text' data...menu-menu'
(uid = 0ffc6f0e-808a-4c2a) [[tag]] span [[xpath]] /html/body/div[5]/span [[text]] x=1485.9 y=665.6 width=23.3 height=21.6
→ [[Attributes]] class='ad...a-4c2a'
(uid = 8d8af84-5b97-477a) [[tag]] div [[xpath]] /html/body/.../div[1] [[text]] [[bbox]] x=419.6 y=461.0 width=476 height=88
→ [[Attributes]] data-webtasks...no-label' [[children]] span input
(uid = 1ea51e98-3fcd-4e30) [[tag]] h4 [[xpath]] /html/body/...1/h4 [[text]] The World's #1 Online Encyclopedia [[bbox]] x=33 y=163
→ width=143.2 height=43.2 [[Attributes]] data-webtasks...d-4e30'
(uid = 769785af-485e-4cf1) [[tag]] a [[xpath]] /html/body/...[2]/a [[bbox]] x=1240.5 y=28.6 width=74.1 height=30 [[Attributes]]
→ id='r... toggleFlyout()' [[children]] span i
(uid = e7b7879f-45ae-48a5) [[tag]] i [[xpath]] /html/body/.../a/i [[bbox]] x=1294.6 y=33.6 width=20 height=20 [[Attributes]]
→ class='fa...e-48a5'
(uid = bf33a062-fb67-44f0) [[tag]] a [[xpath]] /html/body...4/p/a [[text]] Read more [[bbox]] x=567.0 y=641.0 width=69.3 height=16
→ [[Attributes]] href=...67-44f0'
```

Assistant: say(speaker="navigator", utterance="Hi")  
User: say(speaker="instructor", utterance="Open Encyclopedia website.")  
Assistant: say(speaker="navigator", utterance="Yes, sure") load(url="https://www.encyclopedia.com/")  
User: say(speaker="instructor", utterance="Search for biotechnology") Please select the best action using the correct format, do not provide any other information or explanation.  
Assistant:



- models
  - zero-shot / finetuned LLMs
  - closed / open LLMs
- evaluation using **custom metrics**
  - text-only models  
(surprisingly) better than  
multimodal models

Models	Size	Intent	Element	Text	Overall Score	
		IM	IoU	F1	TEST <sub>OOD</sub>	TEST <sub>ID</sub>
<i>Zero-shot</i>						
Llama-2	13B	43.5	4.9	1.4	5.2	5.6
GPT-3.5T	–	42.7	9.0	3.5	8.8	10.3
GPT-4T	–	41.8	11.2	6.9	11.0	12.2
GPT-4V <sup>Ⓢ</sup>	–	42.3	11.4	6.4	10.9	12.9
<i>Finetuned</i>						
Pix2Act <sup>Ⓢ</sup>	1.3B	82.1	9.3	26.6	18.4	23.9
S-LLaMA	2.7B	<b>84.7</b>	25.3	<b>29.2</b>	27.6	<b>37.4</b>
MindAct	3B	80.1	17.7	23.4	21.9	25.7
Flan-T5	3B	81.6	22.1	26.4	25.2	31.1
Fuyu <sup>Ⓢ</sup>	8B	80.9	17.8	24.5	22.2	30.9
Llama-2	13B	83.0	<b>25.7</b>	28.7	<b>27.8</b>	37.0
GPT-3.5F	–	78.5	21.1	23.8	23.3	30.8