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

Zdeněk Kasner
Generating texts with open LLMs
Generating texts with open LLMs

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

- **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:
```...```
{DATA}
```...
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}:

"
Generating texts with open LLMs

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

  Mountain Brook, Alabama is expected to have a mild and mostly cloudy start to the week, with temperatures ranging from 4.78°C to 11.47°C. Rain is expected on Monday and Tuesday. (...
Evaluating semantic accuracy with GPT-4
Evaluating semantic accuracy with GPT-4

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

<table>
<thead>
<tr>
<th>Error</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>INCORRECT</td>
<td>The fact in the text contradicts the data.</td>
</tr>
<tr>
<td>NOT_CHECKABLE</td>
<td>The fact in the text cannot be checked given the data.</td>
</tr>
<tr>
<td>MISLEADING</td>
<td>The fact in the text is misleading in the given context.</td>
</tr>
<tr>
<td>OTHER</td>
<td>The text is problematic for another reason, e.g. grammatically or stylistically incorrect, irrelevant, or repetitive.</td>
</tr>
</tbody>
</table>

Example data

- **Nokia 3310** | **color**: black, blue, grey | **display**: 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)

Evaluating semantic accuracy with GPT-4

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Human annotations ((E_{\text{human}}))</th>
<th>GPT-4 annotations ((E_{\text{GPT}}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>llama2</td>
<td>The weather in Niemegk, 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. From January 3rd to January 7th, the city of Niemegk 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 12.1 m/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 1.34mm. The wind speed will be 4.31m/s with a gust of 11.5m/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 1.17mm. The wind speed will be 7.07m/s with a gust of 15.1m/s. On January 6th, the temperature will be 9.55°C with a feels-like temperature of 6.59°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 0.65m/s with a gust of 0.7m/s.</td>
<td>The weather in Niemegk, 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. From January 3rd to January 7th, the city of Niemegk 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 12.1 m/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 1.34mm. The wind speed will be 4.31m/s with a gust of 11.5m/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 1.17mm. The wind speed will be 7.07m/s with a gust of 15.1m/s. On January 6th, the temperature will be 9.55°C with a feels-like temperature of 6.59°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 0.65m/s with a gust of 0.7m/s.</td>
</tr>
</tbody>
</table>

Evaluating semantic accuracy with GPT-4
Conversational web navigation with LLMs
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 \(\rightarrow\) collecting a large-scale dataset with professional human annotators
Conversational web navigation with LLMs

- working with real-world websites

CI: 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: Undersstood. If you need further assistance, just let me know.

S2: Instructor says “Please share the link.”
Reference: Alright.
LLaMA-WL: Okay
GPT-4V: Here is the link to the discussion: [incorrect link]
Conversational web navigation with LLMs

- explicit prompt, explicit action space

https://arxiv.org/abs/2402.05930
Conversational web navigation with LLMs

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

<table>
<thead>
<tr>
<th>Models</th>
<th>Size</th>
<th>Intent</th>
<th>Element</th>
<th>Text</th>
<th>Overall Score</th>
<th>TEST\text{ood}</th>
<th>TEST\text{imp}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Zero-shot</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Llama-2</td>
<td>13B</td>
<td>43.5</td>
<td>4.9</td>
<td>1.4</td>
<td>5.2</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>GPT-3.5T</td>
<td>–</td>
<td>42.7</td>
<td>9.0</td>
<td>3.5</td>
<td>8.8</td>
<td>10.3</td>
<td></td>
</tr>
<tr>
<td>GPT-4T</td>
<td>–</td>
<td>41.8</td>
<td>11.2</td>
<td>6.9</td>
<td>11.0</td>
<td>12.2</td>
<td></td>
</tr>
<tr>
<td>GPT-4V</td>
<td>–</td>
<td>42.3</td>
<td>11.4</td>
<td>6.4</td>
<td>10.9</td>
<td>12.9</td>
<td></td>
</tr>
</tbody>
</table>

| **Finetuned** |      |        |         |      |               |                |               |
| Pix2Act    | 1.3B | 82.1   | 9.3     | 26.6 | 18.4          | 23.9           |               |
| S-LLaMA    | 2.7B | **84.7** | 25.3   | **29.2** | 27.6  | **37.4**     |               |
| 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       | 8B   | 80.9   | 17.8    | 24.5 | 22.2          | 30.9           |               |
| Llama-2    | 13B  | 83.0   | **25.7** | 28.7 | **27.8**      | 37.0           |               |
| GPT-3.5F   | –    | 78.5   | 21.1    | 23.8 | 23.3          | 30.8           |               |