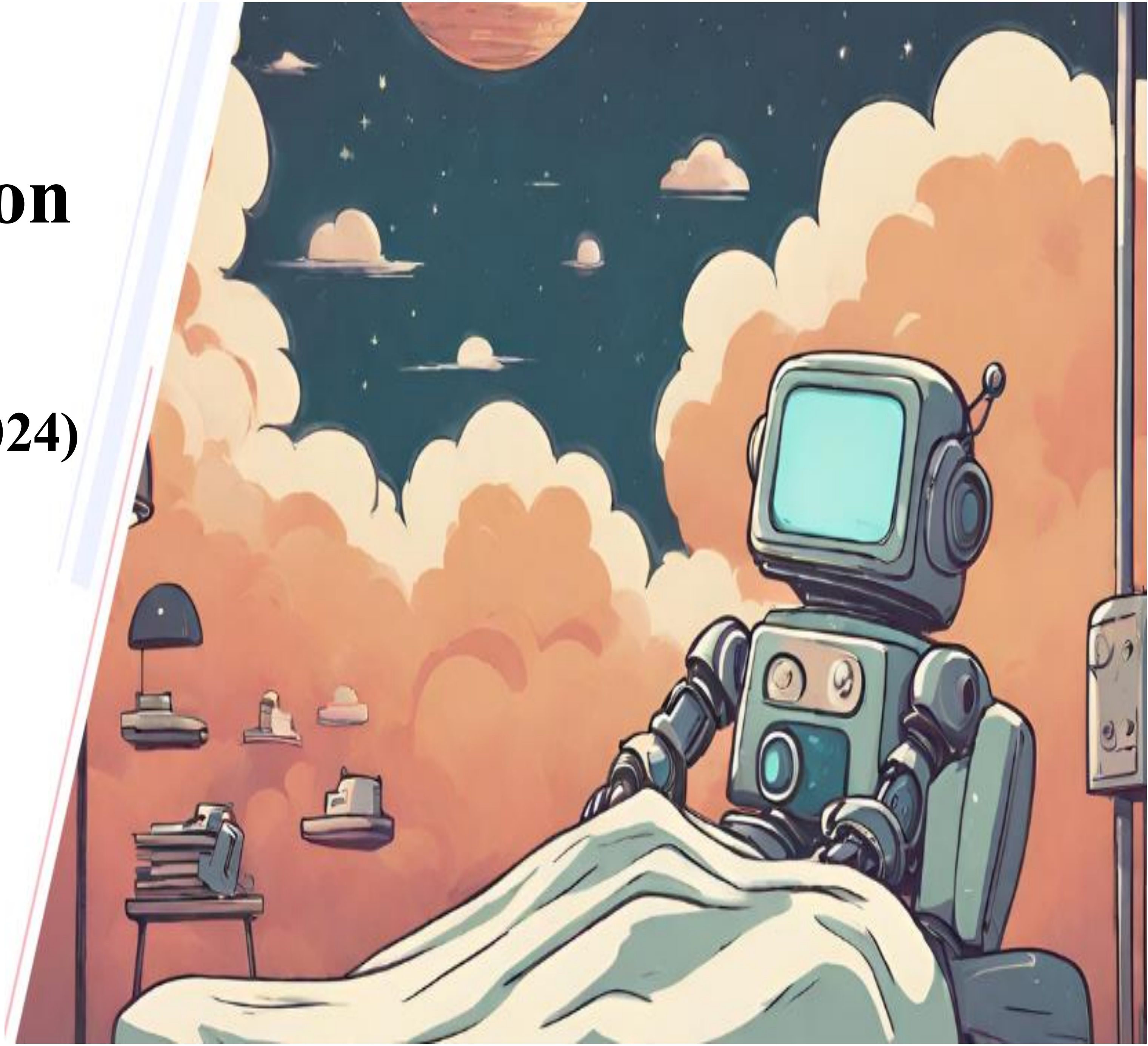
Hallucination Detection

(ELOQUENT Lab @ CLEF 2024)



Philipp Schaer and Narjes Nikzad

08-04-24 – Cologne, Germany
https://ir.web.th-koeln.de

Technology
Arts Sciences
TH Köln

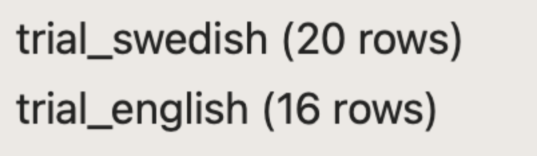
Hallucination Generation

Generation step: Given a source sentence, generate two LLM hypotheses:

- **<hypothesis+>** that is a correct translation/paraphrase of the source, and
- **<hypothesis->** that is a hallucinated translation/paraphrase of the source.

- Paraphrase: English, Swedish
- Machine Translation: English-German, English-French, French-English, German-English

Hallucination Generation



trial_swedish (20 rows)

trial_english (16 rows)

Example for the **paraphrase** task:

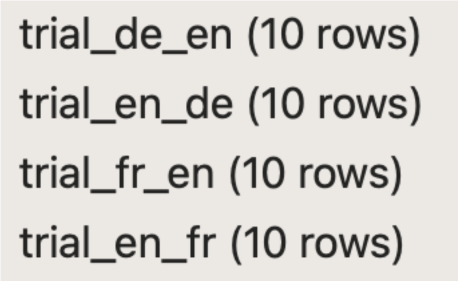
Given the src below, generate a paraphrase hypothesis hyp+ that is supported by src and a second paraphrase hyp- that is not supported by src.

src: *The fact is that a key omission from the proposals on agricultural policy in Agenda 2000 is a chapter on renewable energy.*

Expected output:

Paraphrase hypothesis supported by src (hyp+): One notable absence in the agricultural policy proposals of Agenda 2000 is a section addressing renewable energy.

Paraphrase hypothesis not supported by src (hyp-): Agenda 2000 lacks comprehensive measures for addressing climate change impacts within agricultural policy, which could significantly hinder the transition to renewable energy sources.



trial_de_en (10 rows)

trial_en_de (10 rows)

trial_fr_en (10 rows)

trial_en_fr (10 rows)

Hallucination Generation

Example for the **translation** task:

Given the src below, generate a translation hypothesis hyp+ that is supported by src and a second translation hyp- that is not supported by src.

src: Es ist der Sitz des Bezirks Zerendi in der Region Akmola.

target language: *English*

Expected output:

Translation hypothesis supported by src (hyp+):
It is the seat of the district of Zerendi in Akmola region.

Translation hypothesis not supported by src (hyp-):
It will be the seat of the Zerendi District in Akmola Region.

Submission Fromat

- We expect you to submit your solutions in a .csv format using a google form. More details about the google form will be announced early next week.

- We ask the participants to submit a `.csv` file with the answers and the prompts they used (for hallucination detection).

Submission dates

1. **May 1st:** Participants submit their results for the **hallucination generation step**
2. **May 3rd:** Organisers collect the submitted outputs of the hallucination generation step and redistribute them to the participants so that they can get started with the **cross-model evaluation phase** of the hallucination generation step
3. **May 7th:** Participants submit their results for the **hallucination detection step**
4. **May 10th:** Participants submit their results for the **cross-model evaluation of the hallucination generation step**

LLMs

- Gemma

Link

- Llama3

Link