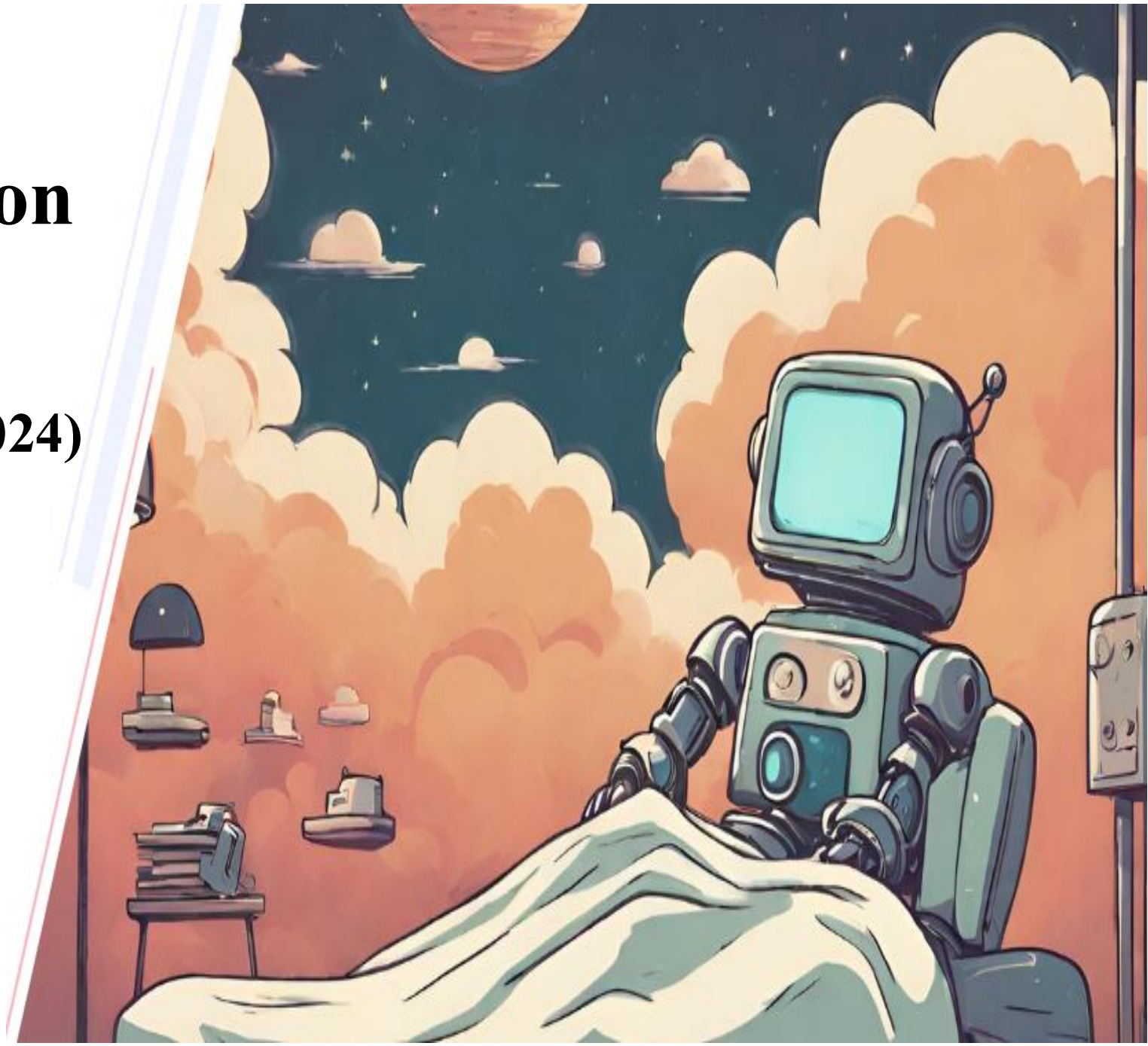
Hallucination Detection

(ELOQUENT Lab @ CLEF 2024)

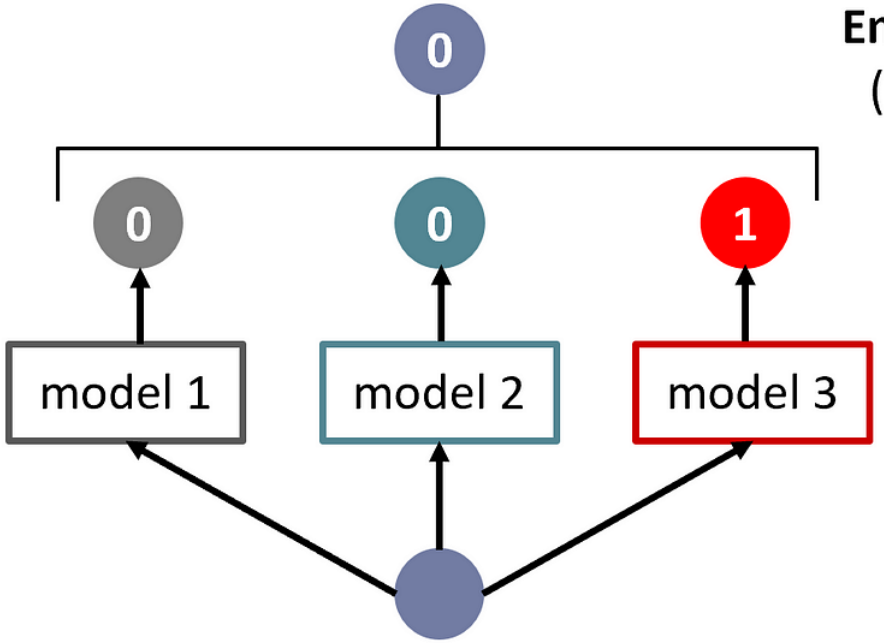
Philipp Schaer and Narjes Nikzad

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

Technology
Arts Sciences
TH Köln



Simple voting approach



Ensemble's prediction
(e.g., majority vote)

Predictions

Inverse Proportion approach for calculating weights

Goal: **assigning higher weights** to **LLMs with better F1-scores**.

- ✓ Split your trail dataset into training (for cot or few-shot learning) and validation sets.
- ✓ Evaluate the performance of each LLM on the validation set using F1-score.
- ✓ Calculate Weights:
  - ➢ Calculate the inverse of each LLM's F1-score: weight_i = 1 / (1- F1_score_i (where i represents the LLM).
  - ➢ Normalize the weights so they sum to 1: normalized_weight_i = weight_i / sum(weights).

Inverse Proportion approach for calculating weights

✓ Calculate Inverse Proportions:

- weight_GPT4 = 1 / (1-0.85) = 6.666
- weight_LaMDA3 = 1 / (1-0.78) = 4.545
- weight_Gemma = 1 / (1-0.82) = 5.555

✓ Sum the Weights:

total_weight = weight_GPT4 + weight_LaMDA3 + weight_Gemma
total_weight = 6.666 + 4.545 + 5.555 = 16.766

✓ Normalize Weights:

normalized_weight_GPT4 = weight_GPT4 / total_weight
normalized_weight_GPT4 = 6.666 / 16.766 ≈ 0.3975
normalized_weight_LaMDA3 ≈ 0.271
normalized_weight_Gemma ≈ 0.3313