

# Cost and gain of AI

Dr. rer. nat. Marco Selig



# hollistic costs of ...

- AI service & licence /
- infrastructure (e.g., hardware)
- operation (e.g., power)
- **data** (& statistics)
- staff (& training)
- (scientific) resilience
- data protection & information security



- efficiency  
**information gain**

quantifies the amount of knowledge is gained

# cost(-risk)-gain tradeoff

economics

?



**information gain**  
quantifies the amount of knowledge is gained

# cost(-risk)-gain tradeoff



Is some **cloud** AI  
more costly but  
even more informative?

Is some **local** AI  
less costly but  
equally informative?



# cost(-risk)-gain tradeoff



Is some **cloud** AI  
more costly but  
even more informative?

Is some **local** AI  
less costly but  
equally informative?

... for a **concrete** use case



# information gain (IG)



... is defined by the relative entropy and thus equals the **Kullback-Leibler divergence**, which ...

$$\text{IG} = H(A, B) - H(B) = D_{\text{KL}}(P(\cdot|A) || P(\cdot|B)) = \int dx \, P(x|A) \log \frac{P(x|A)}{P(x|B)}$$

- is non-negative, asymmetric, and
- does NOT satisfy the triangle inequality

# bits of information



image of a coin

444 x 444 pixels x 24 BPP (8 per RGB channel)

**= 4 731 264 bits of data**



# bits of information



image of a coin

444 x 444 pixels x 24 BPP (8 per RGB channel)

**= 4 731 264 bits of data**

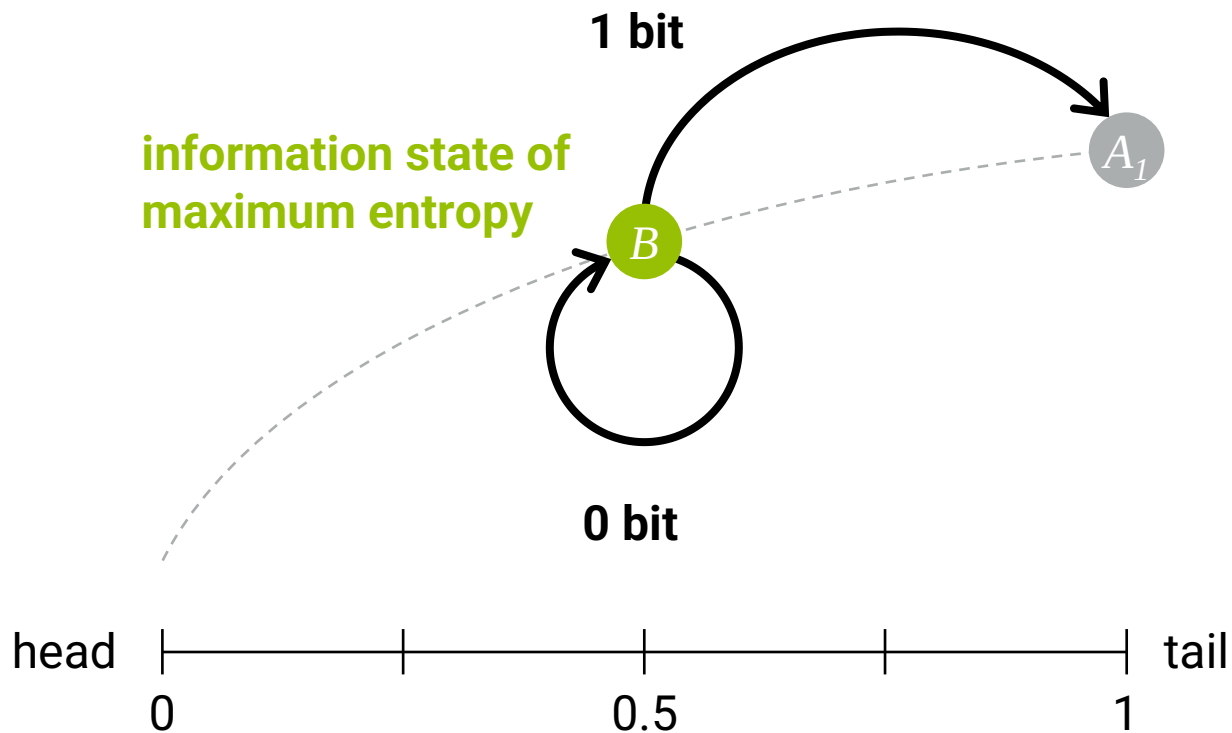
Heads or tails?

Head.

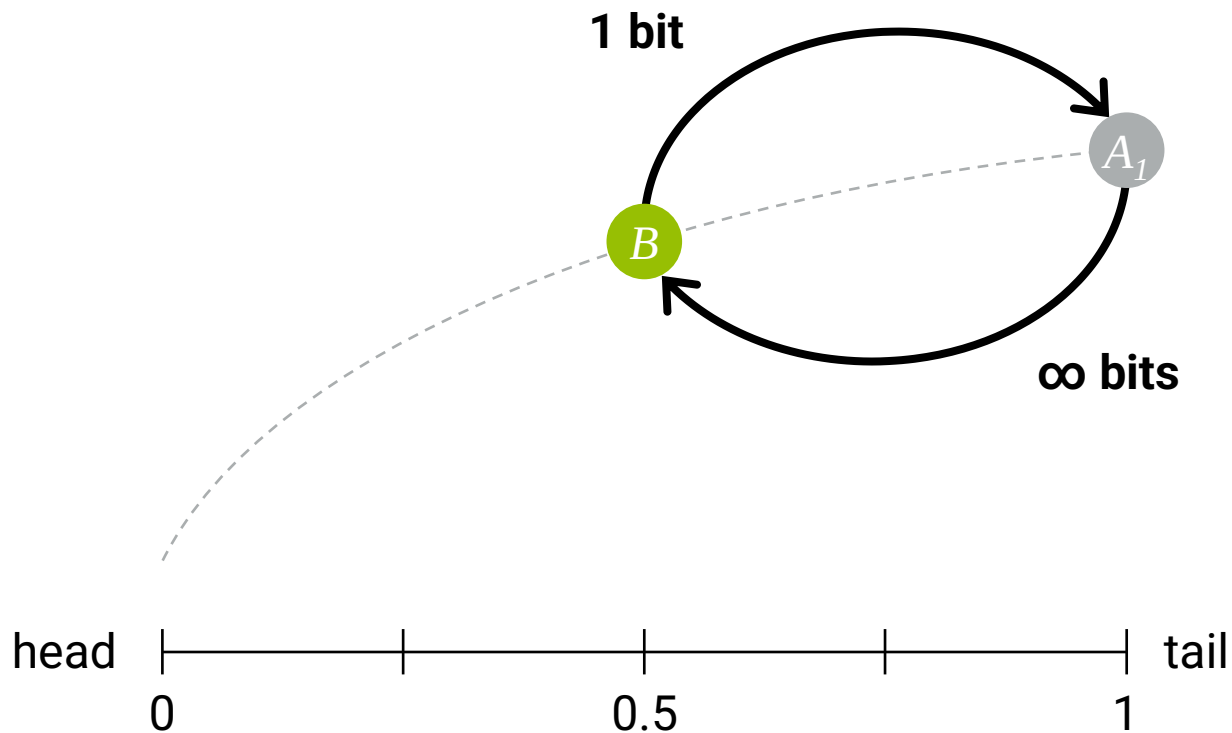
**= 1 bit of information**



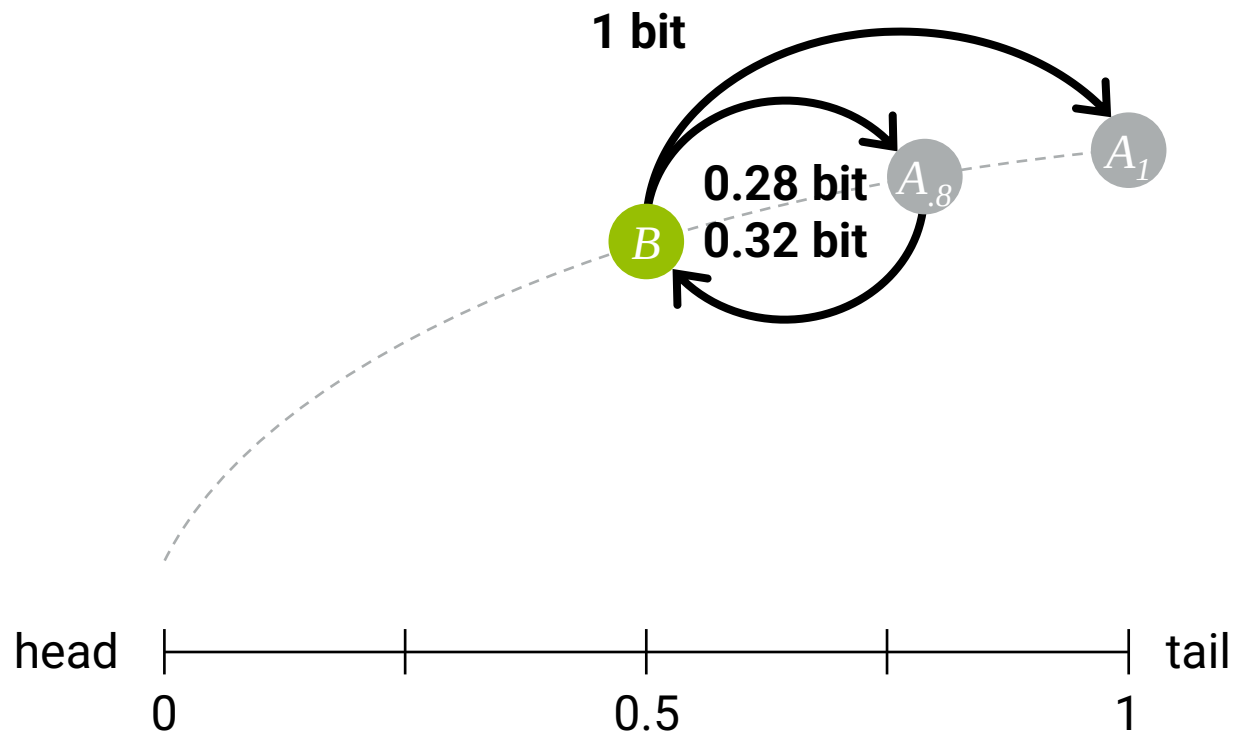
# bits of information gain



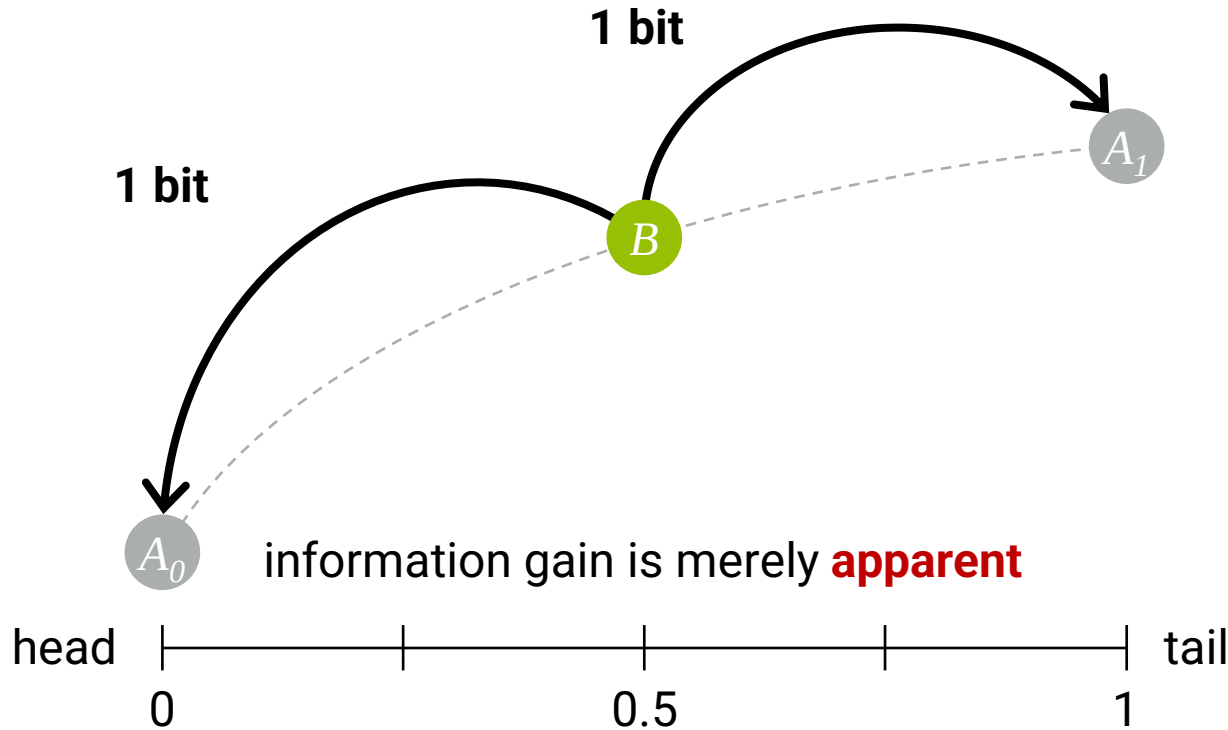
# bits of information gain



# bits of information gain



# bits of information gain



# achieved information gain (AIG)



... is defined by the difference in relative entropy  
with respect to some ideal reference information state

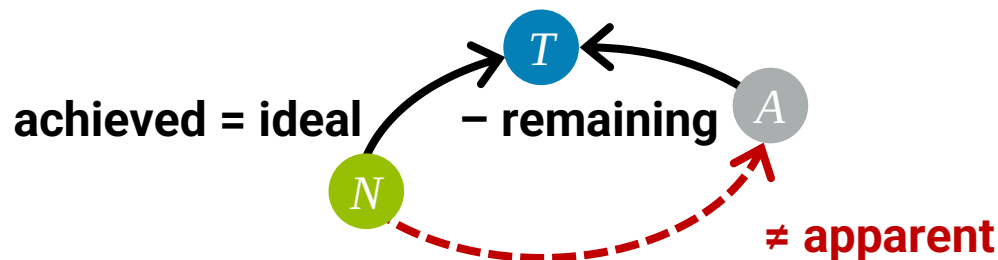
$$\text{AIG} = D_{\text{KL}}(P(\cdot|T)||P(\cdot|N)) - D_{\text{KL}}(P(\cdot|T)||P(\cdot|A)) = \int dx \, P(x|T) \log \frac{P(x|A)}{P(x|N)}$$

# achieved information gain (AIG)

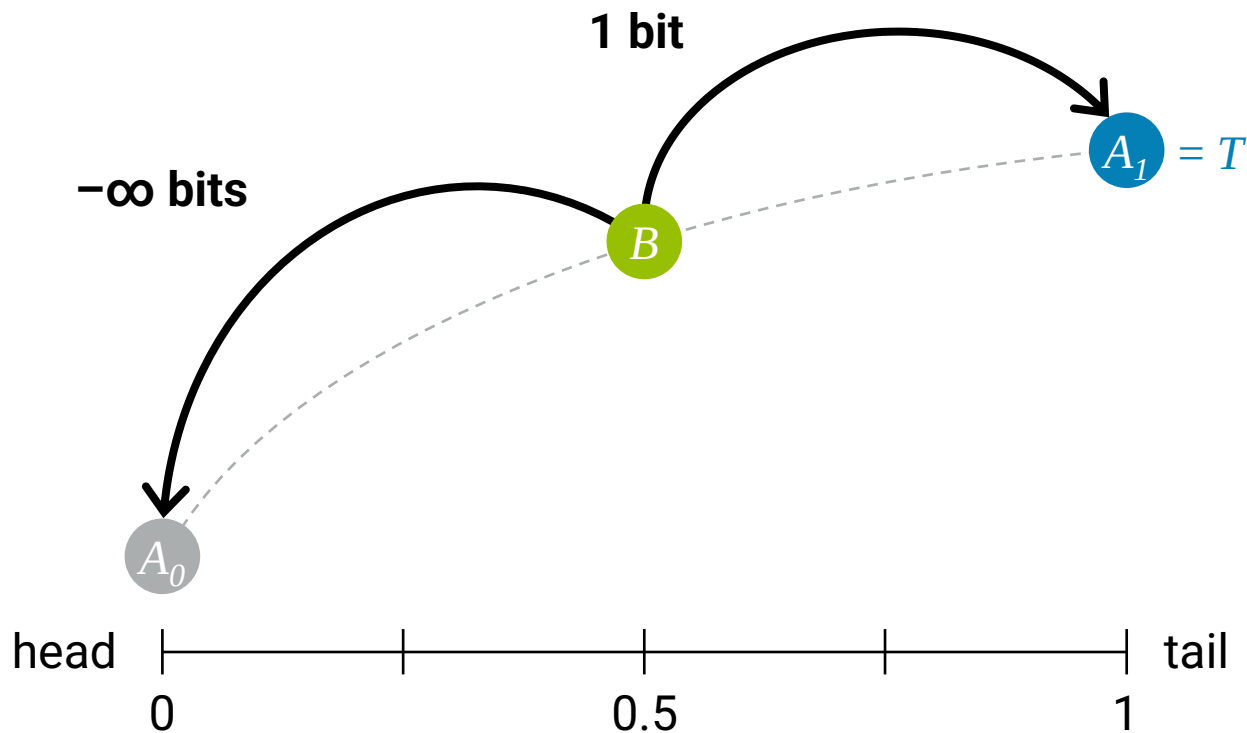


... is defined by the difference in relative entropy  
with respect to some ideal reference information state

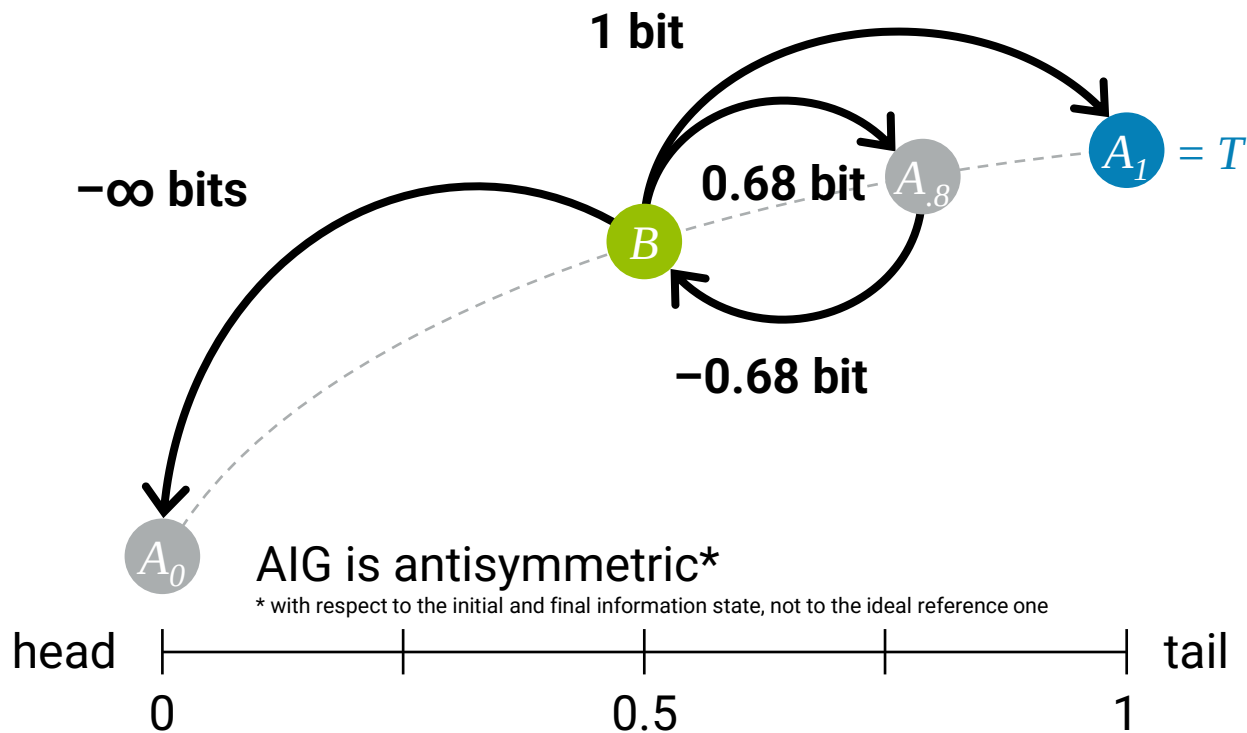
$$\text{AIG} = D_{\text{KL}}(P(\cdot|T)||P(\cdot|N)) - D_{\text{KL}}(P(\cdot|T)||P(\cdot|A)) = \int dx \, P(x|T) \log \frac{P(x|A)}{P(x|N)}$$



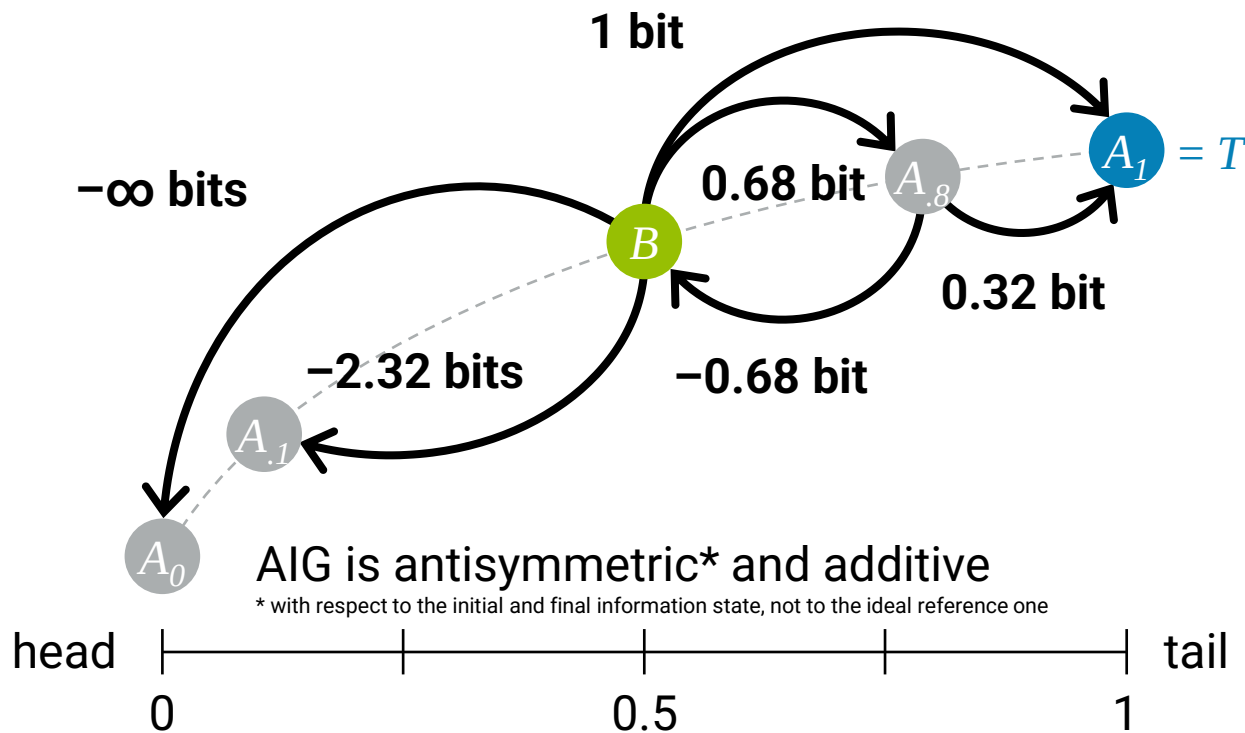
# bits of achieved information gain



# bits of achieved information gain



# bits of achieved information gain



# how to toss LLMs



*“Is it legally sound?”*

# how to toss LLMs



## 1. Q&A benchmark dataset

- general benchmark or
  - concrete dataset; e.g.,  
portfolio of documents
- 
- utilise a supreme LLM
  - involve experts

# how to toss LLMs



## 1. Q&A benchmark dataset

- general benchmark or
- concrete dataset; e.g., portfolio of documents

- 
- utilise a supreme LLM
  - involve experts

## 2. atomic information extraction

- atomise answers into indivisible and self-sufficient statements
- compare by number

- 
- utilise a supreme LLM as a judge

# how to toss LLMs



## 1. Q&A benchmark dataset

- general benchmark or
  - concrete dataset; e.g., portfolio of documents
- 

- utilise a supreme LLM
- involve experts

## 2. atomic information extraction

- atomise answers into indivisible and self-sufficient statements
  - compare by number
- 

- utilise a supreme LLM as a judge

---

cost of data

# LLM confusion matrix



|                        |                           |
|------------------------|---------------------------|
| $a$<br>found/TP        | $t - a$<br>missed/FN      |
| $h$<br>hallucinated/FP | $\infty$<br>irrelevant/TN |

$t$  atomic information in “ground truth”

$a + h$  atomic information in answer

# LLM confusion matrix



|                        |                           |
|------------------------|---------------------------|
| $a$<br>found/TP        | $t - a$<br>missed/FN      |
| $h$<br>hallucinated/FP | $\infty$<br>irrelevant/TN |

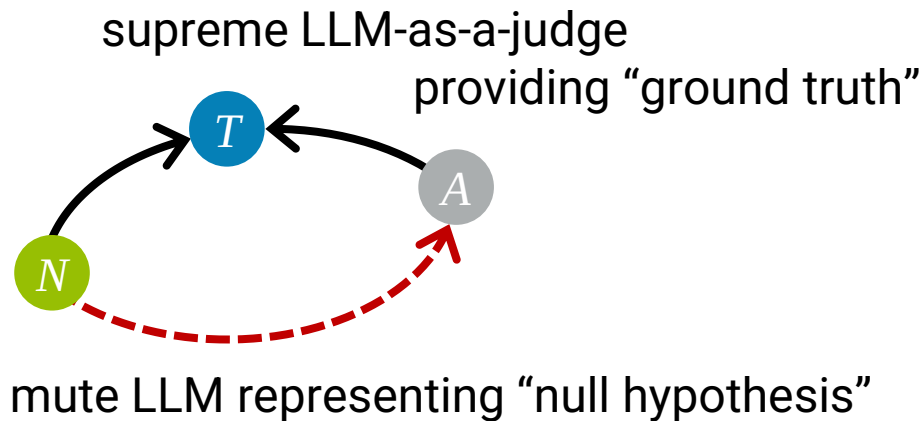
$t$  atomic information in “ground truth”

$a + h$  atomic information in answer

# LLM confusion matrix & AIG

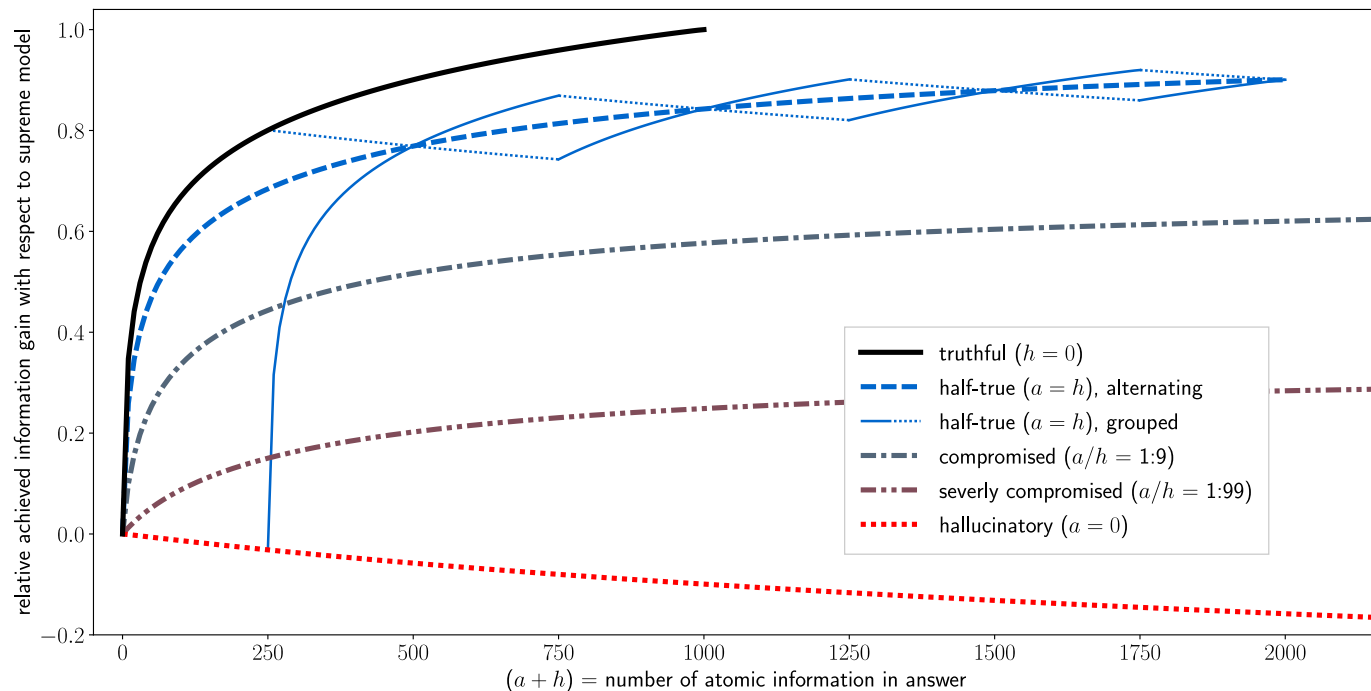


|                        |                           |
|------------------------|---------------------------|
| $a$<br>found/TP        | $t - a$<br>missed/FN      |
| $h$<br>hallucinated/FP | $\infty$<br>irrelevant/TN |



$$\text{AIG} = \frac{t}{t + 3\varepsilon} \log_2 \left( \frac{t + \varepsilon}{\varepsilon} \right) - \frac{\varepsilon}{t + 3\varepsilon} \log_2 \left( \frac{\varepsilon^2 (t + h + 3\varepsilon)^2}{(t + 3\varepsilon)(t - a + \varepsilon)(h + \varepsilon)} \right)$$

# answer mixture & AIG



# Gemini 3.5 Flash vs. Gemma-4



- proprietary, cloud-scale
- about **250B–300B** (A**10B–16B**)
- 1M token context window

- open-weight
- **26B** (A**4B**)
- 256k token context window

# Gemini 3.5 Flash vs. Gemma-4



- proprietary, cloud-scale
- about **250B–300B** (A**10B–16B**)
- 1M token context window

- open-weight
- **26B** (A**4B**)
- 256k token context window

## AIG $\approx$ 2.85 bits (67.8%)

These values are preliminary results. They have neither undergone rigorous quality assurance nor formal peer review. Due to the high volatility of this work in progress, the final results may differ fundamentally, up to and including the complete reversal or invalidation of current trends.

# conclusion



- The costs associated with AI are multifaceted; in particular, the **costs of data** should be taken into account.
- The **achieved information gain** (AIG) is a viable measure for updates of information states with respect to some ideal reference.
- The executive summary reads:  
*“Compared to the reference AI \_\_\_, a conversation with large language model \_\_\_ leaves one with \_\_\_ percent of the knowledge (at \_\_\_ percent of the cost).”*

**Deutsches Biomasseforschungszentrum**

gemeinnützige GmbH



Dr. rer. nat. Marco Selig

Tel.: +49 (0)341 2434-584

E-Mail: [datalab@dbfz.de](mailto:datalab@dbfz.de)

**DBFZ Deutsches  
Biomasseforschungszentrum  
gemeinnützige GmbH**

Torgauer Straße 116

D-04347 Leipzig

Tel.: +49 (0)341 2434-112

E-Mail: [info@dbfz.de](mailto:info@dbfz.de)

[www.dbfz.de](http://www.dbfz.de)