German also Hallucinates!

Inconsistency Detection in News Summaries with the Absinth Dataset

Laura Mascarell[1], Ribin Chalumattu[1], Annette Rios[2]

[1] ETH Zurich
[2] University of Zurich

ETH zürich

University of Zurich[UZH]

Hallucination in News Summarization

*Only** for English

[Qiu et al., 2023]
[Gekhman et al., 2023]

The Absinth Dataset

Hallucination Detection in German News Summarization

- 200 news articles and 4,314 summary-sentence level annotations
- Intrinsic and Extrinsic Hallucinations

Source: Prof. Park awarded Nobel Prize in Physics.
{**F**} Nobel Physics Prize goes to Prof. Park.
{**I**} Prof. Park awarded Nobel Prize in **Economics**.
{**E**} Prof. Park (**58**) awarded Nobel Prize in Physics.

Table 1: Examples faithful to the source (F), containing intrinsic (I), or extrinsic hallucinations (E).

The Absinth Dataset

Hallucination Detection in German News Summarization

- 200 news articles and 4,314 summary-sentence level annotations
- Intrinsic and Extrinsic Hallucinations
- Multiple summarization models
  - mBART
  - mLongT5
  - GPT-4
  - GPT-4 **Intrinsic** hallucinations
  - GPT-4 **Extrinsic** hallucinations
  - Llama 2 70B
  - Llama 2 7B

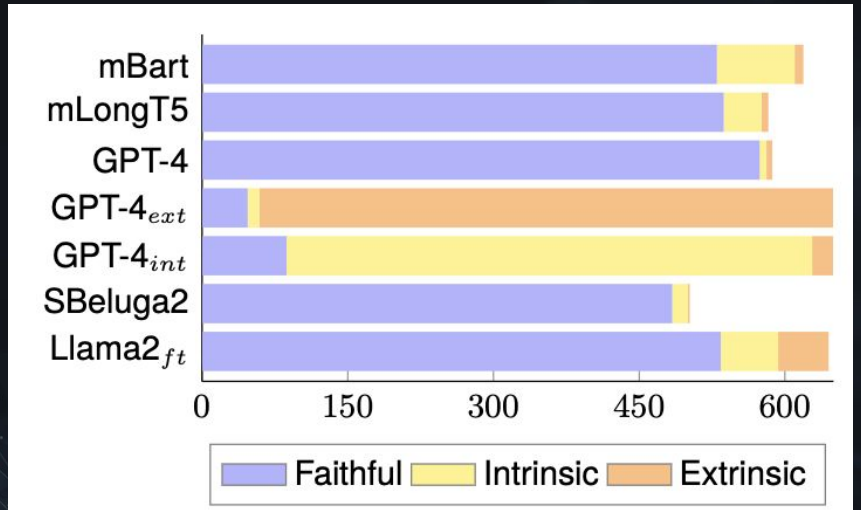


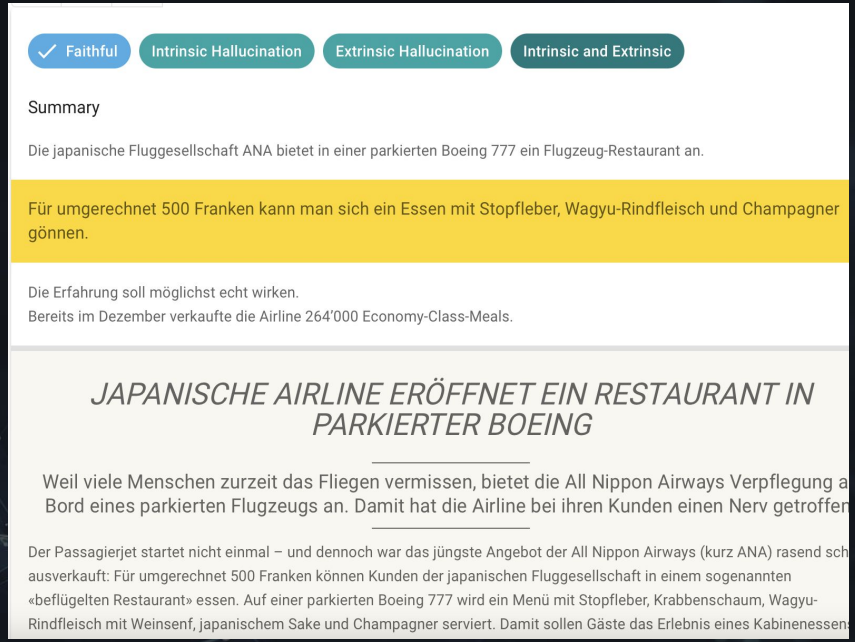
Figure 1: Class distribution for each summarization model in our dataset.

The Absinth Dataset

Annotation Strategy

- 12 German Native Speakers
- 50 articles/annotator (~8hours)
- In-person training
- Intuitive Annotation Framework
- Continuous evaluation on gold annotations
- Fair pair

➔ IRA: 0.81 F/H and 0.77 all labels

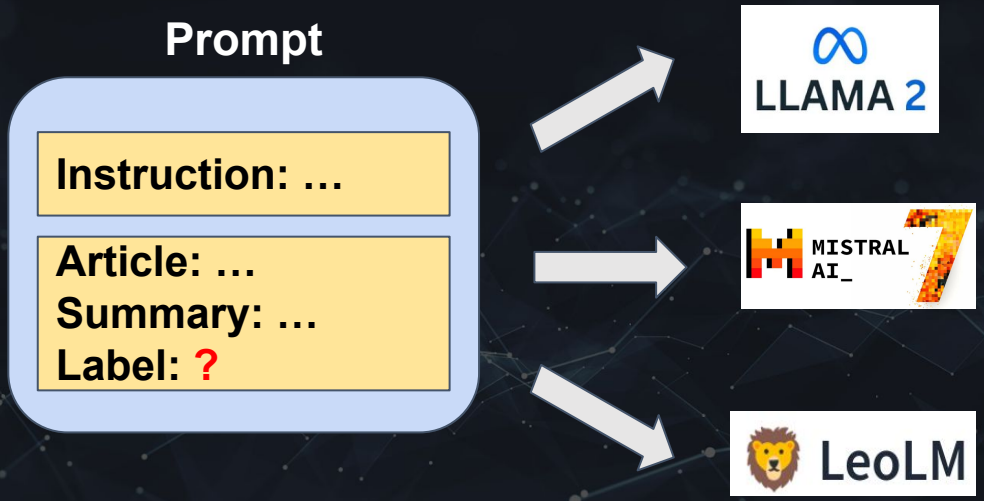


Doccano Annotation Framework

Hallucination Detection Task

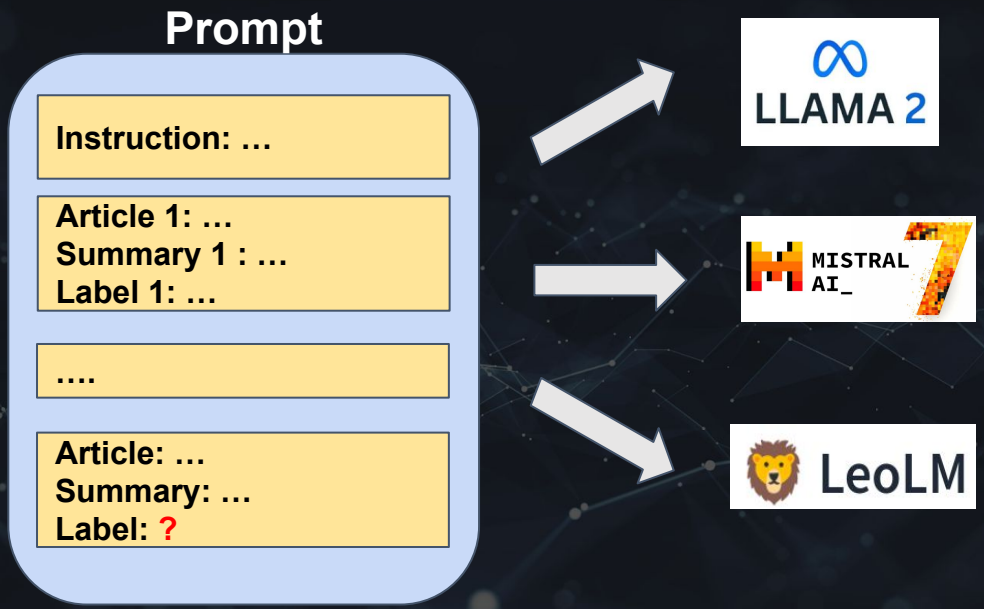
Hallucination Detection Experiments

- Evaluation of Language Models on Absinth Testset
- Experiments with LLMs: Zero-shot



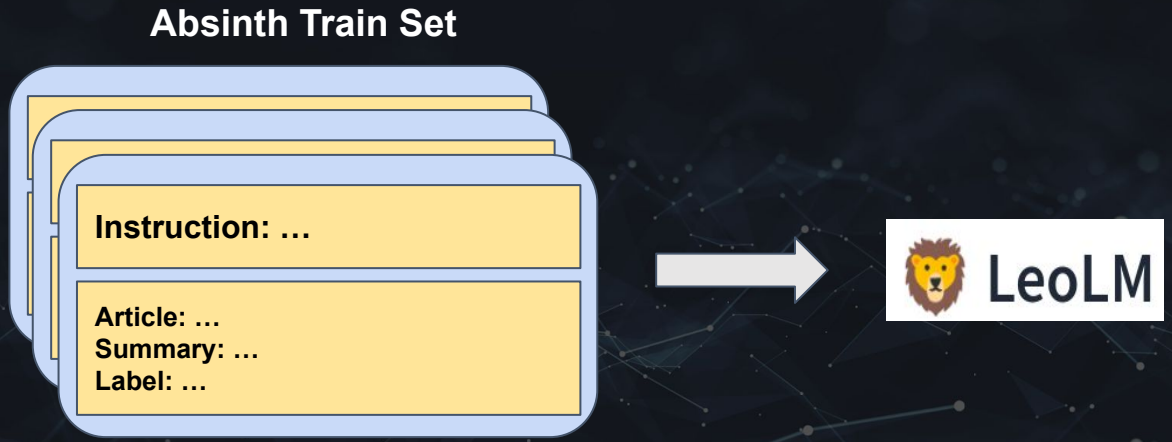
Hallucination Detection Experiments

- Evaluation of Language Models on Absinth Testset
- Experiments with LLMs: Few-shot (3)



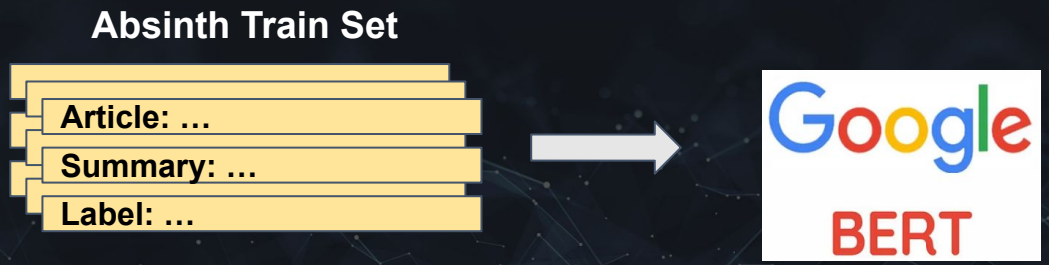
Hallucination Detection Experiments

- Evaluation of Language Models on Absinth Testset
- Experiments with LLMs: Instruction Tuning



Hallucination Detection Experiments

- Goal: Evaluation of Language Models on Absinth Testset
- Experiments with Fine Tuning multi-lingual BERT (mBERT)



Hallucination Detection Results

Model	Setting	F_1 macro	F_1 Faithful	F_1 Intrinsic	F_1 Extrinsic	BACC
Llama2 7b	zero-shot	0.265	0.776	0.019	0.0	0.335
Llama2 7b	few-shot (3)	0.226	0.318	*0.308*	0.052	*0.344*
Llama2 13b	zero-shot	0.258	0.774	0.0	0.0	0.332
Llama2 13b	few-shot (3)	*0.280*	0.290	*0.315*	*0.237*	*0.375*
LeoLM-mistral 7b	zero-shot	0.143	0.077	0.054	0.299	0.327
LeoLM-mistral 7b	few-shot (3)	*0.281*	*0.415*	*0.103*	*0.326*	*0.385*
LeoLM 7b	zero-shot	0.274	0.467	0.326	0.028	0.377
LeoLM 7b	few-shot (3)	0.103	0.0	0.0	*0.310*	0.333
LeoLM 13b	zero-shot	0.258	0.773	0.0	0.0	0.331
LeoLM 13b	few-shot (3)	*0.372*	0.554	*0.241*	0.321	0.419
LeoLM 13b	fine-tuning	*0.483*	**0.886**	0.029	*0.533*	*0.530*
mBERT	fine-tuning	**0.740**	**0.882**	**0.564**	**0.780**	**0.732**
XLM-RoBERTa	fine-tuning	0.642	0.861	0.352	0.714	0.624

Table 4: Macro-averaged F_1, class-wise F_1, and BACC scores averaged over three seeds in different settings—i.e. fine-tuning, zero-shot, and three few-shot prompting—on our inconsistency detection task. We highlight the improvements over the corresponding zero-shot. The overall best performance is in bold.

Hallucination Detection Insights

- Encoder-only mBERT achieves better performance than 7B and 13B LLMs
- Few-shot Prompting improved Intrinsic and Extrinsic Hallucination Detection
- Higher Detection Accuracy for Extrinsic Hallucination than Intrinsic Hallucination

Open Source Links

- Absinth Dataset
  https://github.com/mediatechnologycenter/Absinth
- LREC-COLING paper
  https://arxiv.org/abs/2403.03750
- Open Source Models on Hugging Face
  - mBERT - Fine Tuned on Absinth
  - Leo-Mistral - Instruction Tuned on Absinth