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

Hallucination in News Summarization

Language models are still prone to hallucinate information. In particular for text summarization, there is no guarantee that the summary is faithful to the source article.

Following Maynez et al. (2020) [1], we distinguish between intrinsic and extrinsic hallucinations:

- **Intrinsic** hallucinations are **counterfactual** to the source article.
- **Extrinsic** hallucinations **add information** not present in the source.

	Prof. Park awarded Nobel Prize in Physics.
Faithful	Nobel Physics Prize goes to Prof. Park.
Intrinsic	Prof. Park awarded Nobel Prize in Economics.
Extrinsic	Prof. Park (58) awarded Nobel Prize in Physics.

Table 1: Examples faithful to the source, containing intrinsic, or extrinsic hallucinations.

The ABSINTH Dataset

ABSINTH is a manually annotated summarization dataset for **inconsistency detection in German**, consisting of 4,314 summary sentence-level annotations that consider **intrinsic and extrinsic hallucinations**.

Split	Faithful	Extrinsic	Intrinsic
Train	1,957	512	522
Validation	132	42	28
Test Gold	353	92	104
Test Crowd	351	112	100

Table 2: Class distribution in our ABSINTH dataset.

The dataset comprises a sample of **200 articles** from 20Minuten [3] and **seven summaries per article**:

- Multilingual encoder-decoder models **mBART** and **mLongT5**
- LLM models **GPT-4**, **Llama2-7b**, and **Stable Beluga2**
- GPT-4 to enforce **intrinsic** and **extrinsic** hallucinations

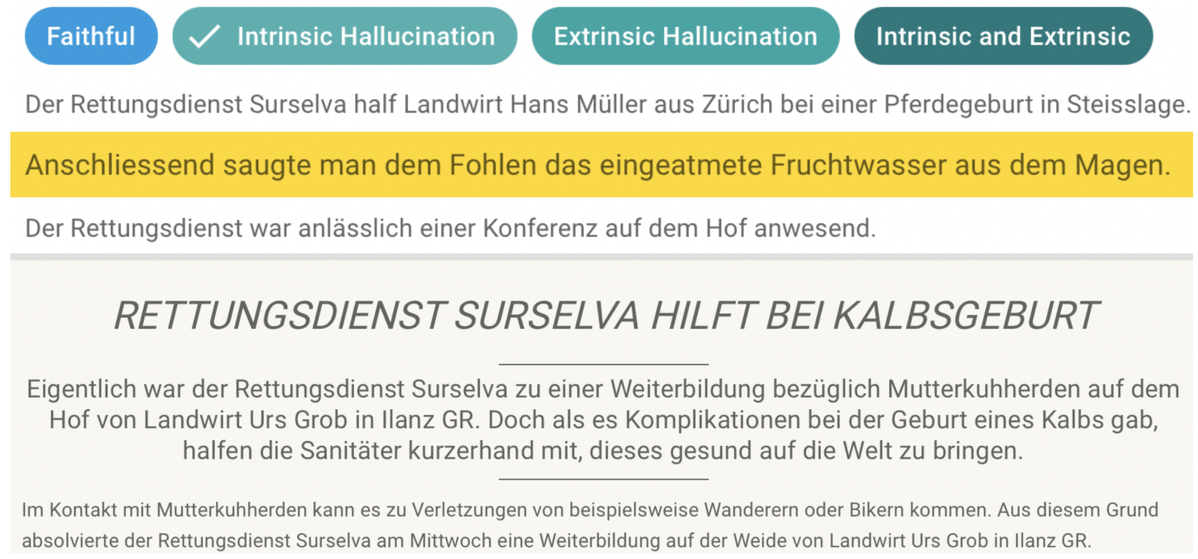
GPT-4	Der Samstag könnte ein sehr gefährlicher Tag werden.
GPT-4$_{int}$	Der Sonntag könnte ein sehr sicherer Tag werden.
GPT-4$_{ext}$	Der Samstag, Halloween, könnte ein sehr gefährlicher Tag werden.

SCAN ME

GitHub: mediatechnologycenter/Absinth

This project is supported by Ringier, TX Group, NZZ, SRG, VSM, viscom, the ETH Zurich Foundation, and the Swiss Innovation Agency Innosuisse under grant agreement number PFFS-21-47.

Manual Annotation Task



- Team of 12 native German speakers
- Final IRA: 0.81 for Faithful/Hallucination and 0.77 for all labels

Annotation Strategy

- 50 articles per annotator (duration 8h)
- In-person training and clear annotation guidelines
- Use of intuitive annotation framework: Doccano [2]
- Continuous evaluation on gold standard

Inconsistency Detection on ABSINTH Testset

Multiclass classification task: Given a source article and a summary sentence, classify the sentence as either *Faithful*, *Extrinsic* or *Intrinsic*.

Model	Setting	F_1 macro	F_1 Faithful	F_1 Intrinsic	F_1 Extrinsic	BACC
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**

Table 3: 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.

Conclusion

- mBERT achieves better performance than 7B and 13B LLMs.
- Few-shot(3) and fine-tuning improves detection of intrinsic and extrinsic hallucinations over zero-shot.
- Models are generally better at detecting extrinsic hallucination over intrinsic hallucination.

References

[1] Joshua Maynez et al. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of ACL*, July 2020.

[2] Hiroki Nakayama et al. doccano, 2018.

[3] Tannon Kew et al. 20Minuten: A Multi-task News Summarisation Dataset for German, 2023.