German also Hallucinates! Inconsistency Detection in News Summaries with the Absinth Dataset

Laura Mascarell[†*], Ribin Chalumattu[†*], Annette Rios[§]
[†]ETH Zurich, [§]University of Zurich
lmascarell@inf.ethz.ch, cribin@inf.ethz.ch, arios@ifi.uzh.ch

Abstract

The advent of Large Language Models (LLMs) has led to remarkable progress on a wide range of natural language processing tasks. Despite the advances, these large-sized models still suffer from hallucinating information in their output, which poses a major issue in automatic text summarization, as we must guarantee that the generated summary is consistent with the content of the source document. Previous research addresses the challenging task of detecting hallucinations in the output (i.e. inconsistency detection) in order to evaluate the faithfulness of the generated summaries. However, these works primarily focus on English and recent multilingual approaches lack German data. This work presents ABSINTH, a manually annotated dataset for hallucination detection in German news summarization and explores the capabilities of novel open-source LLMs on this task in both fine-tuning and in-context learning settings. We open-source and release the ABSINTH dataset to foster further research on hallucination detection in German.

Keywords: Summarization, Natural Language Generation, Evaluation Methodologies, Corpora

1. Introduction

The field of natural language processing is currently undergoing a paradigm shift towards the use of Large Language Models (LLMs), showing a performance leap against the state-of-the-art pre-trained language models such as GPT-2 (Radford et al., 2019) by increasing their parameter scale (Zhao et al., 2023). Despite the emerging abilities of these LLMs, they are still prone to fabricate information, that is, to hallucinate. In particular for text summarization, there is no guarantee that the information in the generated summary is faithful to the source document (Tam et al., 2023).

Most of the research on inconsistency detection in summarization is focused on English, relying on annotated data that is not available in other languages (Goyal and Durrett, 2021; Kryscinski et al., 2020; Durmus et al., 2020). Recently, Qiu et al. (2023) and Gekhman et al. (2023) propose multilingual approaches and evaluate them on the XLSum (Hasan et al., 2021) and mFace (Aharoni et al., 2023) datasets, respectively. Even though these datasets comprise 44 languages, they do not include German, making it infeasible to assess inconsistency detection in this language.

It is important to highlight that there is not yet a consensus in the research community on the appropriate level of granularity for tackling this task. For the sake of simplicity, some existing benchmarks provide overall summary-level annotations of faithfulness (Li et al., 2023; Clark et al., 2023; Aharoni et al., 2023), thus making it challenging to pinpoint where the hallucination occurs. Furthermore, all hallucinations often fall under the same category. Maynez et al. (2020) distinguish between intrinsic and extrinsic hallucinations, as those that are counterfactual and add information to the source, respectively (see Table 1), allowing for a more fine-grained approach to hallucination detection.

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).

In this paper, we present ABSINTH, the first summarization dataset that is manually annotated for inconsistency detection in German.[1] ABSINTH consists of 4,314 summary sentence-level annotations that differentiate between intrinsic and extrinsic hallucinations. Additionally, the dataset comprises the outputs of multiple summarization models, ranging from the state-of-the-art pre-trained language models for German summarization to the latest prompt-based LLMs such as GPT-4 (OpenAI, 2023) and the open-source LLama 2 (Touvron et al., 2023). Finally, we assess the ability of recent open-source LLMs at detecting hallucination using our data and experiment with both fine-tuning and in-context learning to adapt the models to our task. We compare their performance with conventional transformer models such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) fine-tuned on the task. Our results show that mBERT achieves the best overall performance, whereas there is room for improvement with LLMs.

*Equal contribution

[1] ABSINTH GitHub repository

LREC-COLING 2024, pages 7696–7706

2. The ABSINTH Dataset

ABSINTH is a dataset of German news articles and their generated summaries that is manually annotated for hallucination detection. In particular, ABSINTH consists of 4,314 article-summary sentence pairs with the associated label *Faithful*, *Intrinsic*, or *Extrinsic* hallucination. In this section, we describe the construction of the dataset (Section 2.1), the annotation task (Section 2.2), and the final steps to build the dataset (Section 2.3).

2.1. Dataset Construction

Our hallucination dataset comprises a random sample of 200 articles from the *20Minuten* (Tannon Kew et al., 2023) test set split[2] and seven summaries per article that we generate using different models and approaches. These models include the multilingual transformer-based models mBART (Liu et al., 2020) and mLongT5 (Uthus et al., 2023) fine-tuned on the *20Minuten* training data. While mBART has been widely used for German summarization (Liu et al., 2020), mLongT5 has been recently introduced to handle long text inputs.

Furthermore, we consider the latest prompt-based LLMs, namely GPT-4 (OpenAI, 2023) and the open-source Llama 2 models (Touvron et al., 2023). Within the Llama 2 family, we employ Stable Beluga 2, a Llama 2 model with 70b parameters fine-tuned on an Orca style dataset (Mukherjee et al., 2023), along with a smaller Llama 2 model with 7b parameters that we fine-tune on *20Minuten*.

Finally, we employ GPT-4 to generate additional hallucinated instances. To ensure that they are not straightforward to identify, we enforce intrinsic and extrinsic hallucinations that adhere to the context of the source article. We therefore provide both article and reference summary and design a prompt for each hallucination type as follows: To generate intrinsic hallucinations, our prompt instructs the model to subtly alter the reference summary such that each sentence in the summary becomes counterfactual to the article. In contrast, our prompt to generate extrinsic hallucinations instructs the model to add information in the reference summary that is not present in the article without deviating from the article topic (see prompts in Appendix B).

2.2. Annotation Task

We design a task to manually annotate our dataset for hallucination detection. More specifically, given an article A and a sentence of a generated summary s, the task is to assess the consistency of s with the content of the source article A. If s is entirely consistent with A, it must be annotated as *Faithful*. In contrast, if s contains hallucinated information, we distinguish between hallucinations that are counterfactual to the content of the article A (*Intrinsic Hallucination*) and those that add information and, therefore, cannot be verified against A (*Extrinsic Hallucination*). Finally, we provide a fourth label to indicate that s contains both intrinsic and extrinsic hallucinations. We then recruit a team of 12 native German speakers to annotate the data, such that every article-sentence summary pair is reviewed by three different annotators.

To ensure that the annotators follow our annotation scheme, we continuously evaluate their performance on a gold standard that we annotated internally. These gold annotations are equally distributed among the sets such that each set comprises 50 different articles. Additionally, the articles and summaries are randomly shuffled for each human annotator to avoid biases. The annotation of a full set takes eight hours, and they were asked to complete it throughout two consecutive days.

Besides the continuous evaluation, we also implemented the following strategies to ensure high-quality annotations and high-inter annotator agreement: (a) in-person training and clear annotation guidelines; (b) the use of an intuitive annotation framework; and (d) a fair pay that aligns with the hourly wage of teaching assistants. Overall, we obtain a Fleiss'κ (Fleiss, 1971) agreement of 0.81 when distinguishing between *Faithful* or *Hallucination* and 0.77 on the four labels, indicating a very high agreement. Previous work reports a lower κ of 0.65 with three annotators on a similar annotation task (Falke et al., 2019), which confirms the effectiveness of our annotation strategy.

Model	FT	R1↑	RL↑	ρ↓	snt	sum
mBART	20m	32.7	23.1	5.4	12	42
mLongT5	20m	33.5	23.9	8.3	13	43
GPT-4	-	31.9	21.2	3.1	23	72
GPT-4$_{ext}$	-	65.7	64.3	1.4	24	87
GPT-4$_{int}$	-	81.2	80.5	1.6	13	45
SBeluga2	-	33.9	22.5	3.5	20	53
Llama2$_{ft}$	20m	32.4	23.0	2.2	11	39

Table 2: Comparison of the summarization models in ABSINTH evaluated on the *20Minuten* test set in terms of rouge-1 and rouge-L scores. The high rouge scores of GPT-4$_{ext}$ and GPT-4$_{int}$ are due to applying the hallucination changes directly in the reference summary. The FT column indicates whether the model is fine-tuned on *20Minuten*. Higher values of the extractive fragment density ρ indicate higher extractiveness (Grusky et al., 2018). *snt* and *sum* are the average token length of the generated sentences and summaries, respectively.

[2] https://github.com/ZurichNLP/20Minuten/tree/main/SwissText_2023

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

Table 3: Class distribution in our ABSINTH dataset.

Gold Standard Three domain experts annotate a gold standard consisting of 25 random articles from our dataset and their corresponding generated summaries. Since each summary contains about three sentences, our gold standard comprises a total of 580 article-sentence summary pairs. The purpose of the gold standard is twofold: Firstly, to identify annotation challenges beforehand, and secondly, to promptly assist those annotators that need further clarification on the task. The Fleiss' κ agreement on *Faithful* or *Hallucination* and all four labels are 0.86 and 0.90, respectively. The experts reached a consensus on the final label for the instances with disagreement, except for three ambiguous instances that are discarded.

Intuitive Annotation Framework To annotate our dataset, we use doccano (Hiroki Nakayama et al., 2018), an open-source crowd-sourcing text annotation tool, and adapt the code to our task (see Appendix A). The framework also allows annotators to add comments such that we can gather more information to inspect ambiguous cases.

Continuous Evaluation We randomly intersperse our gold standard in the annotation data and monitor the performance of each annotator on the gold annotations to provide them with clarifications if necessary. Furthermore, our dataset contains 121 duplicated summary sentences as a result from generating multiple summaries of the same article. We also use these duplicates to monitor their performance. Essentially, if an annotator uses a different label for a duplicate, the annotator is possibly performing the overall task incorrectly. Ultimately, we had to replace one of the annotators due to bad performance on the gold standard samples even when there was no ambiguity.

2.3. Final Dataset

To build the final dataset, we discard 121 duplicates and 11 instances with the label *Intrinsic and Extrinsic*. We then assign the label with the majority vote to the rest. Figure 1 and Table 3 show the distribution of the classes across the models and dataset splits, respectively.[3] The test split contains our gold annotations and 25 additional articles, where at least one annotator disagrees on multiple instances, under the assumption that those samples are more challenging to predict. Additionally, the dataset includes a set of 71 instances with no agreement. To distinguish these instances from the actual test set, we mark them as 'full disagreement'.

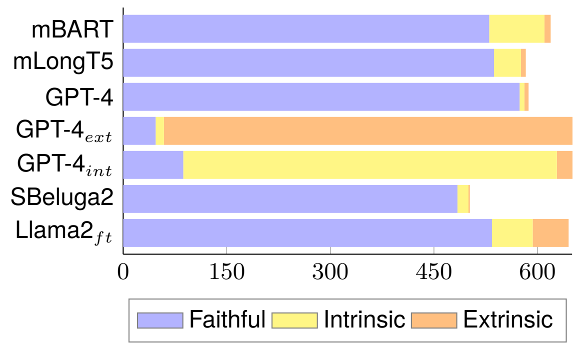

Figure 1: Class distribution for each summarization model in ABSINTH. The largest models GPT-4 and Stable Beluga 2 generate the least hallucinations. Since summaries are of different sentence length, the total of instances varies among models.

3. Inconsistency Detection Task

Our multi-classification task consists on predicting the faithfulness of a summary sentence to the source article (i.e. *Faithul*, *Intrinsic*, or *Extrinsic* hallucination) according to the definition in Section 2.2. We then assess the performance of different open-source LLMs on the ABSINTH test set in multiple settings, such as fine-tuning and in-context learning, where we extend the prompt with random examples on each label from the ABSINTH training data.

3.1. Models Selection

We adopt a wide range of models from the conventional mBERT and XLM-RoBERTa to a variety of open-source LLMs. Specifically, we consider the Llama 2 family (Touvron et al., 2023), which shows high performance on different tasks,[4] and experiment with base Llama 2 with 7b and 13b parameters. Additionally, we consider LeoLM 7b and 13b models, which adapt Llama 2 to German through continued pretraining on German data, and Mistral 7b, which outperforms Llama 2 on multiple benchmarks (Jiang et al., 2023).

3.2. Results

Table 4 shows the performance of the selected models in the zero-shot, three few-shot prompt-

[3]Test gold class distribution after discarding three ambiguous instances, 22 duplicates, and six instances with the label *Intrinsic and Extrinsic*.

[4]Open LLM Leaderboard

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.

ing, and prompt-based fine-tuning settings. We report macro-averaged F_1, class-wise F_1, and the balanced accuracy BACC scores—i.e. the average of accuracy scores from each class. The BACC scores are also adopted in the related work (Kryscinski et al., 2020; Laban et al., 2022) as they are not affected by the majority class (Hanselowski et al., 2018; Thölke et al., 2023).

We observe that the prompt-based LLMs improve the detection of intrinsic and extrinsic hallucination with the fine-tuning or the in-context learning setting, where models are prompted with three examples from our dataset. In particular, LeoLM 13b achieves the best performance, showing the benefits of further training on German data. However, LLMs exhibit a poor performance overall on this classification task. In contrast, the conventional transformer models mBERT and XLM-RoBERTa perform remarkably well, with mBERT achieving the best performance across the three classes. These results are consistent with Sun et al. (2023), where the authors claim that LLMs underperform fine-tuned models in text classification tasks.

Finally, we observe that the models are generally better at detecting extrinsic hallucinations than intrinsic hallucinations. The main difference between these types of hallucination is that the information labelled as extrinsic hallucination is not present in the source article. We suggest that in future work, LLMs could benefit from chain-of-thought prompting techniques that elicit reasoning in these models (Wei et al., 2022) to improve their prediction of intrinsic hallucinations.

4. Related Work

Previous work mostly focuses on the English language and implements inconsistency detection metrics in supervised and unsupervised settings (Huang et al., 2021). Whilst the former are trained on English datasets annotated for this task (Kryscinski et al., 2020; Goyal and Durrett, 2021), the latter adopts existing models trained for Natural Language Inference (NLI) or question answering to detect inconsistencies in summaries (Falke et al., 2019; Maynez et al., 2020; Laban et al., 2022; Durmus et al., 2020). Since these approaches rely on data and models that are limited to English, they cannot be directly applied to other languages. An exception is the XNLI dataset, the machine translated counterpart of the English NLI data. However, the dataset has been used in multilingual settings with unsatisfactory results (Qiu et al., 2023).

More recent research implements multilingual approaches instead. Qiu et al. (2023) leverage machine translation to generate a multilingual labeled summarization dataset for inconsistency detection. To annotate the dataset, their approach combines the predictions of several inconsistency metrics for English. Similarly, Gekhman et al. (2023) annotate a multilingual training dataset using FLAN-PaLM 540b (Chung et al., 2022), a LLM fine-tuned on the NLI task. Both approaches use their own synthetic dataset to fine-tune the multilingual pre-trained models BERT (Devlin et al., 2019) and T5 (Xue et al., 2021), respectively, and evaluate their performance on mFace, a multilingual test set for factual consistency evaluation of abstractive summarization (Aharoni et al., 2023). Although mFace comprises 44 languages, it does not include German. Other approaches use ChatGPT[5] to evaluate factual inconsistency (Luo et al., 2023; Li et al., 2023). However, the accuracy is only slightly above random chance. Additionally, Aiyappa et al. (2023)

[5]https://openai.com/chatgpt

argue against using ChatGPT for evaluation, as we cannot guarantee that there is no training-test contamination. In contrast, our work compares the performance of recent open-source LLMs in both fine-tuning and in-context learning settings.

5. Conclusion

Due to the lack of German data for inconsistency detection, we present the ABSINTH dataset, a collection of German news articles and their generated summaries that has been manually annotated for this task. The dataset provides summary sentence-level annotations that distinguish between hallucinations that are counterfactual to the article (intrinsic) and those that add information not present in the source (extrinsic), allowing for a more fine-grained approach to detecting hallucination.

We then evaluate the performance of novel open-source LLMs on this classification task using our data and experiment with different settings including few-shot prompting and prompt-based fine-tuning. Whilst LLMs improve their performance with fine-tuning or three-shot prompting, they exhibit a poor overall performance. Our results show that the conventional transformer model mBERT significantly outperforms the prompt-based models.

We expect this work to supplement and foster research on detecting hallucination that includes the German language, and we are excited to further explore this direction in future work.

6. Ethics Statement

We obtained the corresponding exemption determination (EK-2023-E-3) from the Ethics Commission of ETH Zurich university to perform the annotation task as it does not pose any risk for the annotators. In addition, the annotations were anonymously collected and no conclusions can be drawn about any specific annotator.

The summaries in ABSINTH are automatically generated, and we did not check them for problematic content such as hate speech or biases. Nevertheless, we do not anticipate further ethical issues besides those already identified in text generation (Smiley et al., 2017; Kreps et al., 2022).

7. Limitations

The articles used to create ABSINTH are part of the *20Minuten* dataset (Rios et al., 2021). We use the SwissText_2023 test split (Tannon Kew et al., 2023), since this version has been filtered for duplicates and overlap with mc4 (Raffel et al., 2020), a multilingual dataset commonly used for pretraining LLMs. However, since the dataset and the original news articles are available online, it is still possible that some of the newer LLM's might have seen these articles as part of their pre-training. The annotated dataset is limited to news articles in Standard Swiss German from one particular news outlet, 20Minuten. The articles are in general relatively short and informal in style, but cover a wide range of topics. Models trained for faithfulness assessment on this data might not perform as well on longer, more complex texts.

8. Acknowledgements

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

9. Bibliographical References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-yeol Ahn. 2023. Can we trust the evaluation on ChatGPT? In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 47–54, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020.

Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. *arXiv preprint arXiv:2303.15621*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. *arXiv preprint arXiv:2305.13632*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

Charese Smiley, Frank Schilder, Vassilis Plachouras, and Jochen L. Leidner. 2017. Say the right thing right: Ethics issues in natural language generation systems. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 103–108, Valencia, Spain. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Philipp Thölke, Yorguin-Jose Mantilla-Ramos, Hamza Abdelhedi, Charlotte Maschke, Arthur Dehgan, Yann Harel, Anirudha Kemtur, Loubna Mekki Berrada, Myriam Sahraoui, Tammy Young, et al. 2023. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *NeuroImage*, 277.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Uthus, Santiago Ontanon, Joshua Ainslie, and Mandy Guo. 2023. mLongT5: A multilingual and efficient text-to-text transformer for longer sequences. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9380–9386, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Advances in Neural Information Processing Systems.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

10. Language Resource References

Hiroki Nakayama et al. 2018. *doccano*. distributed via github, 1.8.4. PID https://github.com/doccano/doccano.

Tannon Kew et al. 2023. *20Minuten: A Multi-task News Summarisation Dataset for German*. Department of Computational Linguistics, University of Zurich, distributed via github, 1.0. PID https://github.com/ZurichNLP/20Minuten.

A. Annotation Framework

We extend doccano and implement a user interface to annotate article-summary sentence pairs. See an example in Figure 2.

B. Prompts

Table 7 lists all the prompts that we use in this work. We design these prompts using gpt-prompt-engineer.[6]

[6] `https://github.com/mshumer/gpt-prompt-engineer`

C. Technical Details

We use the HuggingFace Trainer API to fine-tune all models for summarization and inconsitency detection with ABSINTH. We train the LLMs with 4-bit QLoRA (Dettmers et al., 2024) on an Nvidia A100-80GB GPU and the smaller language models with default fine-tuning on an Nvidia 3090 GPU. We set the temperature to 0 during inference for all LLMs.

C.1. Summary Generation Details

We train mBART,[7] mLongT5,[8] and LLama 2 7b[9] on *20Minuten* to generate summaries for the ABSINTH dataset—see fine-tuning details in Table 5. During inference, we apply beam search and a beam size of 3 with mBART and mLongT5, and greedy decoding with Llama 2 7b. To generate summaries with GPT-4, we use OpenAI API[10] and the `gpt-4-0613` snapshot from June 13th, 2023 with a context window of 8,192 tokens. Lastly, we use the Lm-Eval-Harness framework (Gao et al., 2023) to generate zero-shot summaries with Stable Beluga 2 with a context window of 4,096.[11] Table 6 lists the prompts used to generate the summaries.

C.2. Inconsistency Detection Details

We evaluate all LLMs using the Lm-Eval-Harness framework on the ABSINTH test split. Specifically, we evaluate zero-shot and few-shot with the following model checkpoints from HuggingFace: Llama 2 7b,[12] Llama 2 13b,[13] LeoLM-Mistral 7b,[14] LeoLM 7b,[15] and LeoLM 13b.[16] In the few-shot setting, we randomly select 3 samples (i.e. one per label) from the training split and shuffle them. Finally, we further fine-tune mBERT,[17] XLM-RoBERTa,[18] and LeoLM 13b on the ABSINTH training split. Table 5 and Table 7 provide the fine-tuning details and the corresponding prompts, respectively.

[7] facebook/mbart-large-cc25
[8] agemagician/mlong-t5-tglobal-base
[9] NousResearch/Llama-2-7b-hf
[10] `https://platform.openai.com/`
[11] stabilityai/StableBeluga2
[12] NousResearch/Llama-2-7b-hf
[13] NousResearch/Llama-2-13b-hf
[14] LeoLM/leo-mistral-hessianai-7b
[15] LeoLM/leo-hessianai-7b
[16] LeoLM/leo-hessianai-13b
[17] google-bert/bert-base-multilingual-cased
[18] FacebookAI/xlm-roberta-base

Model	Training Set	Epochs	Learning Rate	Batch Size	Context Window
LLama 2 7b*	*20Minuten*	5	$2e-4$	8	4,096
mBart	*20Minuten*	10	$3e-5$	32	1,024
mLongT5	*20Minuten*	10	$3e-5$	32	2,048
LeoLM 13b*	ABSINTH	1	$2e-4$	8	4,096
mBERT	ABSINTH	5	$2e-5$	32	512
XLM-RoBERTa	ABSINTH	5	$2e-5$	32	512

Table 5: Model fine-tuning details. The asterisk (*) indicates that the model is fine-tuned with QLoRA.

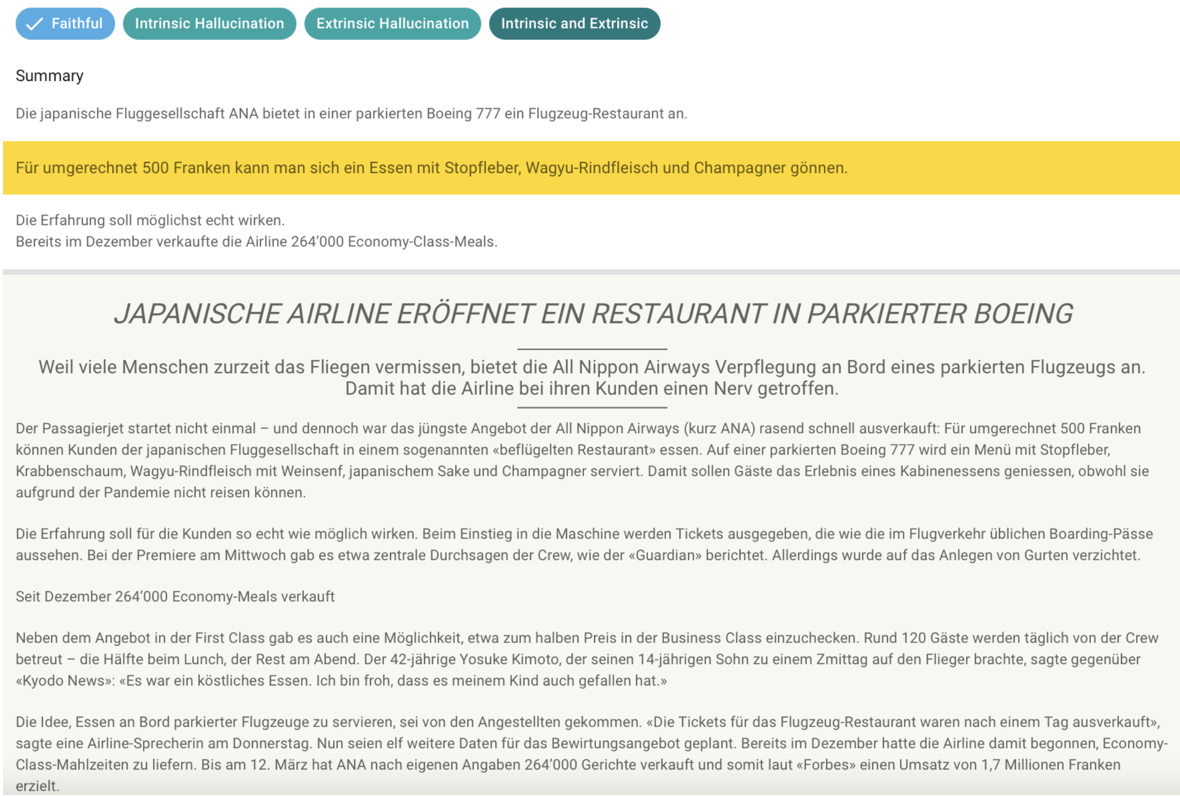

Figure 2: User interface of the annotation framework. We provide the article and all summary sentences. The interface highlights the summary sentence that is currently being reviewed.

GPT-4 Summarization

Provide a concise, 3-sentence summary of the following news article. The summary MUST be written in German.
Article: {article}

GPT-4 Intrinsic Hallucination

Given the news article and its reference summary, subtly alter every sentence of the summary to introduce EXACTLY ONE varied misrepresentations—such as incorrect entities, dates, or details without diverging drastically from the original structure.
Article:{article}
Summary:{summary}

GPT-4 Extrinsic Hallucination

For each sentence in the provided summary of the news article, embed a distinctive, external detail not present in the original article. Every modified sentence should contain this additional information. Ensure these insertions are credible and do not clash with the article's facts.
Article:{article}
Summary:{summary}

Stable Beluga 2 Summarization

System:
You are StableBeluga, an AI that follows instructions extremely well. Help as much as you can. Reply only German.
User: Generate a summary in German for the following article. The summary should be around 2 to 3 sentences.
Article: {article}
Assistant:

Llama 2 7b Summarization

Instruction: Generate a summary in German for the provided article. The summary should be around 2 to 3 sentences.
Article: {article}
Assistant:

Llama 2 7b 20Minuten Fine-tuning

Instruction: Generate a summary in German for the provided article. The summary should be around 2 to 3 sentences.
Article: {article}
Assistant:
{summary}

Table 6: List of all prompts that we use to summarize the articles of the ABSINTH dataset, generate intrinsic and extrinsic hallucinations with GPT-4, and fine-tune Llama 2 7b on *20Minuten*.

LeoLM 13b ABSINTH Fine-tuning
Instruction: Analyze whether the given sentence is faithful to the article. If the sentence solely conveys information that comes directly from the article, without any additions or omissions, respond with ‘Faithful’. If the sentence contains information that is in direct contradiction to the article, respond with ‘Intrinsic Hallucination’. If the sentence introduces information or details that are not explicitly mentioned in the article itself, respond with ‘Extrinsic Hallucination’. Article: {article} Sentence: {sentence} ### Answer: {label}

LLM Inconsistency Detection
Analyze whether the given sentence is faithful to the article. If the sentence solely conveys information that comes directly from the article, without any additions or omissions, respond with ‘Faithful’. If the sentence contains information that is in direct contradiction to the article, respond with ‘Intrinsic Hallucination’. If the sentence introduces information or details that are not explicitly mentioned in the article itself, respond with ‘Extrinsic Hallucination’. Article: {article} Sentence: {sentence} Label:

Table 7: Prompts used on the inconsistency detection task with LLMs.