NLI to the Rescue: Mapping Entailment Classes to Hallucination Categories in Abstractive Summarization

Naveen Badathala*, **Ashita Saxena***, **Pushpak Bhattacharyya**
Department of Computer Science and Engineering, IIT Bombay, India
{naveenbadathala, ashitasaxena, pb}@cse.iitb.ac.in

Abstract

In this paper, we detect hallucinations in summaries generated by abstractive summarization models. We focus on three types of hallucination *viz.* *intrinsic*, *extrinsic*, and *non-hallucinated*. The method used for detecting hallucination is based on textual entailment. Given a premise and a hypothesis, textual entailment classifies the hypothesis as *contradiction*, *neutral*, or *entailment*. These three classes of textual entailment are mapped to intrinsic, extrinsic, and non-hallucinated respectively. We fine-tune a RoBERTa-large model on NLI datasets and use it to detect hallucinations on the XSumFaith dataset. We demonstrate that our simple approach using textual entailment outperforms the existing factuality inconsistency detection systems by 12% and we provide insightful analysis of all types of hallucination. To advance research in this area, we create and release a dataset, *XSumFaith++*, which contains balanced instances of hallucinated and non-hallucinated summaries.

1 Introduction

Natural Language Generation (NLG) has made tremendous progress in neural text generation with the advent of large pre-trained language models like BERT (Devlin et al., 2018) and GPT Series (Radford et al., 2019; Brown et al., 2020; OpenAI, 2023). Although text generation using these models is fluent, it is often observed that the generated text is divergent or unfaithful to the source text (Kryściński et al., 2019; Wiseman et al., 2017; Dhingra et al., 2019). This problem of generating contradicting or irrelevant text is termed *hallucination* (Maynez et al., 2020). The state-of-the-art abstractive summarization systems can generate fluent summaries with high values of automatic evaluation metrics like ROGUE (Lin, 2004). However, the generated summaries are often inconsistent with respect to the original document and such summaries are said to be *hallucinated* (Kryscinski et al., 2020). These hallucinations can be *intrinsic* or *extrinsic*. We follow Maynez et al. (2020) for defining intrinsic and extrinsic hallucination in abstractive summarization. If the generated summary contradicts the source document, we refer to it as *intrinsic* hallucination. If the generated summary contains information that cannot be verified from the source document, we refer to it as *extrinsic* hallucination. When the summary is factually consistent with the document, we refer to the summary as *non-hallucinated*. In this work, we explore the usage of Natural Language Inference (NLI) for hallucination detection in abstractive summarization.

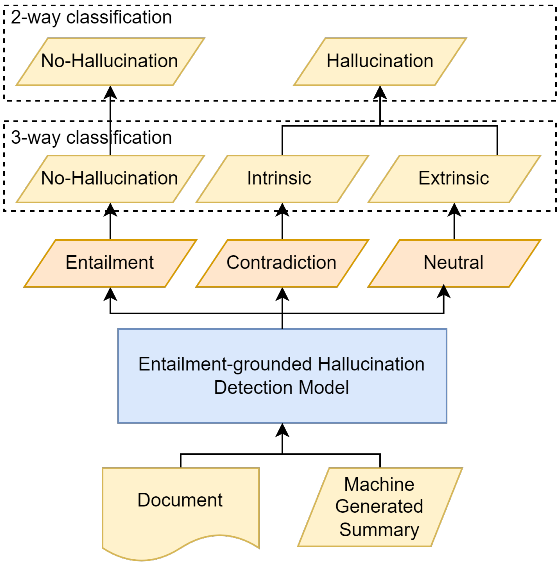


Figure 1: An overview of our entailment-grounded hallucination detection model. The model takes the machine-generated summary and the corresponding document as input and classifies the summary as hallucinated (intrinsic/extrinsic) or non-hallucinated.

Natural Language Inference (a.k.a textual entailment) was first studied in Dagan et al. (2005). Given a premise P, the task of NLI or textual entailment is to classify a hypothesis, H, as con-

[*] These authors contributed equally to this work

tradictory, neutral or entailed with respect to the premise P. We justify in Section 2.3 that these three classes of textual entailment can be mapped to intrinsic, extrinsic, and non-hallucinated respectively. We apply our approach by using models fine-tuned on the NLI task for hallucination classification. We evaluate this approach on the XSumFaith dataset (Maynez et al., 2020) which contains document-summary pairs with summaries labelled as intrinsic, extrinsic or non-hallucinated. Further, we observe that the XSumFaith dataset is heavily skewed towards hallucinated summaries (Section 4.2). To address this imbalance, we create and release XSumFaith++, a balanced extension of XSumFaith, which will aid further research in hallucination detection.

1.1 Motivation

Natural Language Generation (NLG) tasks, which are not open-ended, like document summarization (Nenkova and McKeown, 2011; See et al., 2017; Paulus et al., 2017), require models to be factual and/or faithful to the source text (Maynez et al., 2020). Despite recent improvements in generative models, most summarization systems are prone to hallucinations (Kryściński et al., 2019; Wiseman et al., 2017; Dhingra et al., 2019). Detecting the presence and the type of hallucination is the first step towards hallucination mitigation. Identifying whether a summary contains intrinsic or extrinsic hallucination will enable the development of effective mitigation methods targeted to fix a specific kind of hallucination.

Textual entailment provides a three-way classification given a premise-hypothesis pair *viz.* entailment, contradiction and neutral (Section 2.2). Exploring the usage of textual entailment with the document as premise and summary as hypothesis will help understand the relationship between the types of entailment and the types of hallucination.

1.2 Contributions

Our contributions are:

1. A novel **entailment-grounded** strategy to classify hallucination by mapping entailment classes *viz. contradiction*, *neutral* and *entailment* to *intrinsic*, *extrinsic* and *non-hallucination* respectively (Section 2.3).

2. Demonstrating the efficacy of the entailment-grounded mapping by using a RoBERTa-large model fine-tuned on the NLI datasets and testing it on the XSumFaith dataset (Maynez et al., 2020). We achieve a **12%** improvement over the state-of-the-art consistency detection models (Table 5).

3. A dataset, **XSumFaith++**[1], containing 22,669 balanced instances of hallucinated and non-hallucinated summaries which will aid further research in hallucination detection (Section 4.2). We prepare this dataset by correcting the gold summaries in the XSumFaith dataset (Maynez et al., 2020) and augmenting the dataset with 7,282 instances of non-hallucinated summaries (Table 2).

2 Hallucination and Textual Entailment

2.1 Hallucination

Hallucination, as a psychological term, refers to a perception that is unreal but looks real on the surface (Blom, 2010). In the same way, in NLG, the generated text may contain information that might look correct, but if we verify the information present, it might contain unfaithful or non-factual text. Hallucination is further divided into **intrinsic** and **extrinsic hallucination** (Maynez et al., 2020).

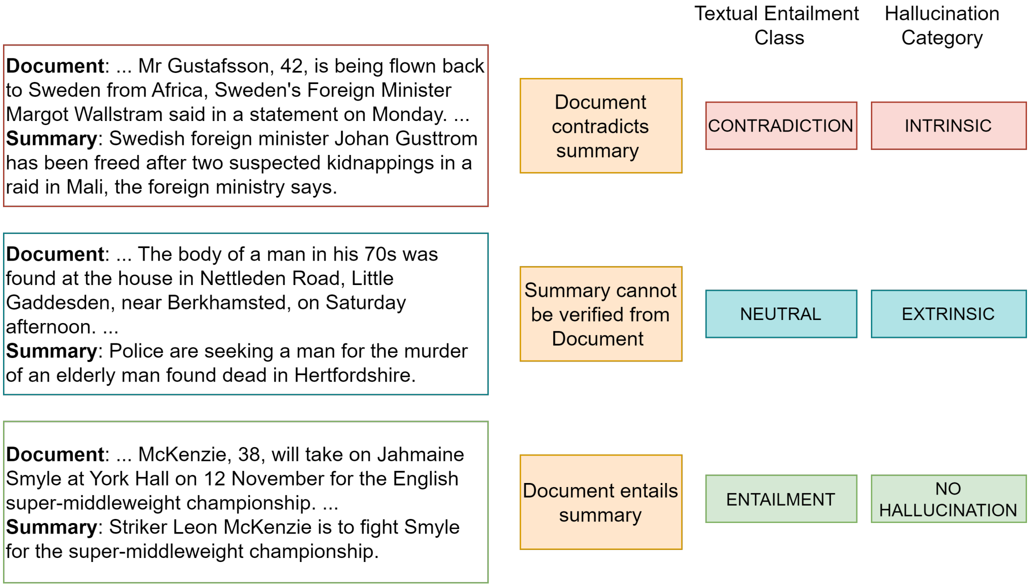
Intrinsic Hallucination: Hallucinations are said to be intrinsic when the generated output contradicts the source text. In abstractive summarization, if the generated summary contradicts the given source information or document, it is referred to as intrinsic hallucination.

Extrinsic Hallucination: This occurs when the generated output cannot be confirmed by the source information. In such cases, the generated output is not supported by or in contradiction with the source information. In abstractive summarization, extrinsic hallucinations arise when the summary neither supports nor contradicts the input document.

2.2 Textual Entailment

Three-way textual entailment classification was introduced by the fourth RTE (Recognizing Textual Entailment) challenge (Giampiccolo et al., 2008). The goal of this challenge was to make a 3-way classification of a given premise (P) and hypothesis (H) pair. The three classes *viz.* entailment,

[1] Code and data are available at: `https://github.com/naveen-badathala/NLI-Hallucination-Mapping`



Document: ... Mr Gustafsson, 42, is being flown back to Sweden from Africa, Sweden's Foreign Minister Margot Wallstram said in a statement on Monday. ...
Summary: Swedish foreign minister Johan Gusttrom has been freed after two suspected kidnappings in a raid in Mali, the foreign ministry says.

Document: ... The body of a man in his 70s was found at the house in Nettleden Road, Little Gaddesden, near Berkhamsted, on Saturday afternoon. ...
Summary: Police are seeking a man for the murder of an elderly man found dead in Hertfordshire.

Document: ... McKenzie, 38, will take on Jahmaine Smyle at York Hall on 12 November for the English super-middleweight championship. ...
Summary: Striker Leon McKenzie is to fight Smyle for the super-middleweight championship.

Figure 2: Examples of document-summary pairs showing the mapping of textual entailment classes to hallucination categories.

contradiction and unknown/neutral are defined by Giampiccolo et al. (2008) as follows:

- P entails H - in which case the pair is marked as ENTAILMENT
- P contradicts H - in which case the pair is marked as CONTRADICTION
- The truth of H can not be determined on the basis of P - in which case the pair is marked as UNKNOWN or NEUTRAL.

2.3 Mapping Textual Entailment to Hallucination Categories

Considering the document as premise and its summary as the hypothesis, the *definitions* of the three classes of textual entailment *viz. contradiction*, *neutral* and *entailment* (Section 2.2) correspond to *intrinsic*, *extrinsic* and *non-hallucination* respectively (Section 2.1). We further motivate this mapping by showing examples of extrinsic, intrinsic, and non-hallucinated instances along with the corresponding entailment classes in Figure 2. It can be observed that the presence and type of hallucination can be easily understood in terms of textual entailment.

3 Related Work

Natural Language Inference (NLI) has been widely applied to various NLP tasks like Question-Answering (QA) (Abacha and Demner-Fushman, 2019; Ben Abacha and Demner-Fushman, 2019; Pathak et al., 2021), Information Extraction (IE) (Clinchant et al., 2006; Wehnert et al., 2019) etc.

Prior work in understanding hallucination includes a survey by Ji et al. (2022) which gives an in-depth discussion of hallucination in various NLG tasks. Hallucination detection methods include token-level hallucination detection for various NLG tasks (Zhou et al., 2021; Rebuffel et al., 2022; Liu et al., 2022). Sentence-level hallucination detection was explored by Laban et al. (2022) using a zero-shot entailment metric. Prior work related to the usage of entailment in summarization includes Falke et al. (2019) which used entailment models to re-rank the generated summaries. Mrini et al. (2021) proposed a novel data-augmented and joint learning approach combining question summarization and Recognizing Question Entailment (RQE) in the medical domain. Louis and Maynez (2022) used an entailment-based self-training approach for abstractive opinion summarization.

Prior work related to factual consistency in abstractive summarization includes Kryscinski et al. (2020) which proposed a weakly-supervised model. Maynez et al. (2020) released a human-annotated hallucination dataset, XSumFaith, containing summaries generated by various abstractive summarization models. Goyal and Durrett (2020) decomposed text at the level of dependency arcs and proposed a DAE model trained on XSumFaith. Scialom et al. (2021) and Fabbri et al. (2022) used QA-based metrics to evaluate factual consistency in abstractive summaries.

In this paper, we explore the usage of textual entailment for a 3-class classification of hallucination in abstractive summarization. To the best of our knowledge, approaches for three-class classification of hallucination (*viz.* intrinsic, extrinsic and non-hallucinated) for the task of abstractive summarization have not been explored before.

4 Dataset

4.1 XSumFaith

The XSumFaith (e**X**treme **Sum**marization **Faith**fulness or XSF) dataset was released by Maynez et al. (2020). It contains 500 random news articles taken from the test set of XSum (Narayan et al., 2018) and for each article, it contains summaries generated using various abstractive summarizers. These summaries along with the corresponding gold summaries (taken from the XSum dataset) are manually labelled as intrinsic or extrinsic along with their spans.

To evaluate our approach of mapping entailment classes to hallucination categories, we use the XSumFaith dataset as it contains intrinsic and extrinsic labels for the machine-generated summaries.

4.2 XSumFaith++

While working with the XSumFaith dataset, we faced the following challenges:

- For the task of hallucination detection, we require the gold summaries to be free from any kind of hallucination so they can be used as non-hallucinated instances. It was found that **only** 23% of the gold summaries were free from hallucination (Maynez et al., 2020).
- The ratio of hallucinated and non-hallucinated instances is heavily skewed towards hallucinated summaries.

We tackle the challenges mentioned above in the following manner:

- We employ human annotators to modify the gold summaries containing hallucinations to free them from intrinsic or extrinsic hallucinations (Section 4.2.1).
- We add instances of non-hallucinated summaries to create a balanced dataset. For this purpose, we use the hallucination-free gold summaries and generate their paraphrases (Section 4.2.2).

As a result of this process, we get a balanced dataset containing hallucination-free gold summaries, which we refer to as **XSumFaith++**. Table 2 shows the distribution of intrinsic, extrinsic, and non-hallucinated instances in XSumFaith++.

4.2.1 Manual Editing of Summaries

We employed three annotators proficient in English. Two annotators were Master's students and one had M.A. in linguistics. Among the three annotators, two annotators were male and one was female all belonging to the age group of 24-30. The annotators were presented with an equal split of article documents and corresponding gold summaries containing hallucinations. They were also given the type of hallucination and the span which were taken from the XSumFaith dataset. The annotation guidelines along with examples are shown in Figure 3.

Human evaluation of the edited summaries: To ensure that the manually edited summaries are fluent, abstract and relevant to the document, we conduct a human evaluation of the modified summaries on three features using the five-point Likert Scale (Likert, 1932). The features on which these ratings are conducted are the following:

Fluency: The grammatical correctness of the summary.

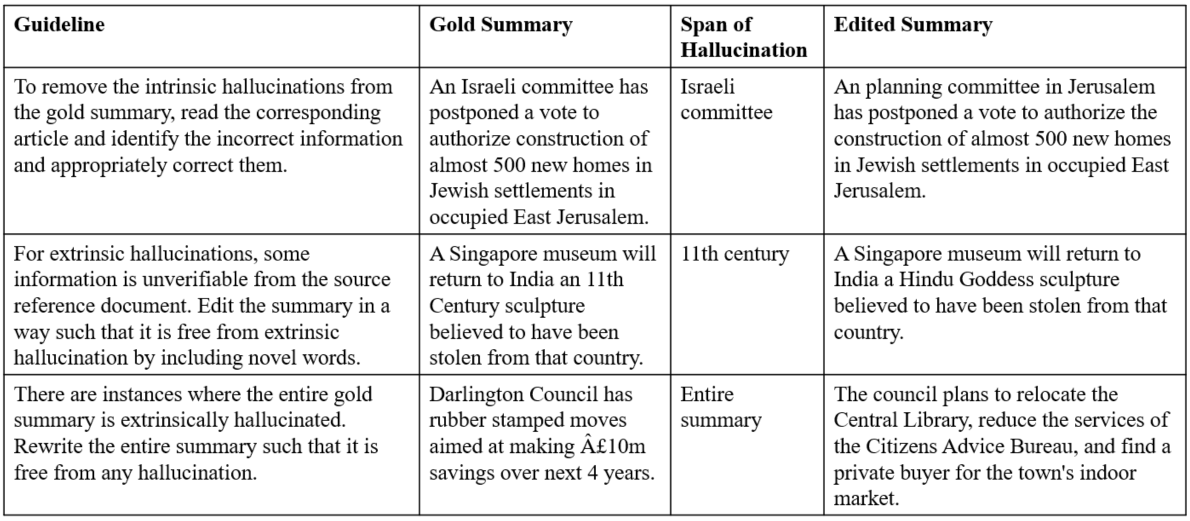
Relevance: The coverage of the theme of the corresponding document and the key points in the summary.

Abstractiveness: The usage of novel words or phrases in the summary.

To ensure fairness, each annotator is given summaries modified by the other two annotators. The average scores for these three features *i.e.,* fluency, relevance, and abstractiveness are **4.93**, **4.14**, and **4.42** respectively. The detailed ratings can be seen in Figure 4. All of the average scores for features are above the agreement mark which indicates the high quality of the edited summaries and the reliability of our XsumFaith++ dataset.

4.2.2 Data Augmentation using Paraphrasing

To balance the number of hallucinated and non-hallucinated instances, we augment the dataset. This augmentation is done by adding paraphrases for the 500 manually edited summaries. Each paraphrased summary is considered a non-hallucinated instance. To generate the paraphrases, we use the



Guideline	Gold Summary	Span of Hallucination	Edited Summary
To remove the intrinsic hallucinations from the gold summary, read the corresponding article and identify the incorrect information and appropriately correct them.	An Israeli committee has postponed a vote to authorize construction of almost 500 new homes in Jewish settlements in occupied East Jerusalem.	Israeli committee	An planning committee in Jerusalem has postponed a vote to authorize the construction of almost 500 new homes in Jewish settlements in occupied East Jerusalem.
For extrinsic hallucinations, some information is unverifiable from the source reference document. Edit the summary in a way such that it is free from extrinsic hallucination by including novel words.	A Singapore museum will return to India an 11th Century sculpture believed to have been stolen from that country.	11th century	A Singapore museum will return to India a Hindu Goddess sculpture believed to have been stolen from that country.
There are instances where the entire gold summary is extrinsically hallucinated. Rewrite the entire summary such that it is free from any hallucination.	Darlington Council has rubber stamped moves aimed at making Â£10m savings over next 4 years.	Entire summary	The council plans to relocate the Central Library, reduce the services of the Citizens Advice Bureau, and find a private buyer for the town's indoor market.

Figure 3: Annotation guidelines to edit gold summaries from the XSumFaith dataset containing hallucinations. Each annotation guideline is shown along with an example of a gold summary, the span containing hallucination and the edited hallucination-free summary.

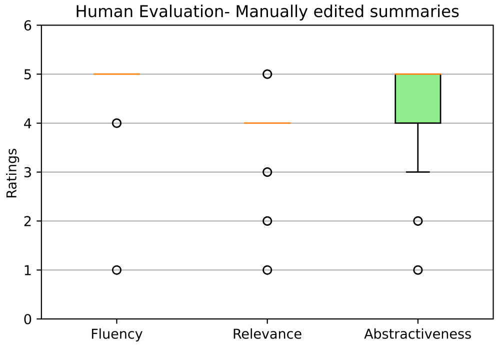


Figure 4: Box plot of human evaluators' ratings for three features: fluency, relevance and abstractiveness on manually edited summaries of XSumFaith++.

PEGASUS model (Zhang et al., 2020) fine-tuned for paraphrasing[2]. The PEGASUS model has a transformer-based architecture. In the pre-training task of PEGASUS, important sentences from a document are removed and masked. PEGASUS generates the missing sentence from the given sentences. PEGASUS can be fine-tuned for many applications. We use the PEGASUS model fine-tuned for the task of paraphrasing. Using this process, we add 7, 282 instances of non-hallucinated data.

Evaluation of the paraphrased summaries: To ensure that the paraphrasing step preserves the semantics of the original summary sentence, we compute BERTScore[3] (Zhang et al., 2019) and BARTScore[4] (Yuan et al., 2021). The BERTScore for the generated paraphrases in our dataset is found to be **0.96**. The BARTScore for the paraphrased sentences is found to be **-1.01** (a higher negative score is better).

To ensure that there are no hallucinations introduced by paraphrasing, we also perform a human evaluation of the paraphrased summaries. We employ the same three annotators (demographic details mentioned in Section 4.2.1). We randomly take 500 pairs of paraphrased summaries and the corresponding non-hallucinated summaries. These are annotated by giving a 0 or 1 score. A score of 1 is given if the paraphrased summary does not contain any hallucination and correctly aligns with the non-hallucinated summary, and 0 otherwise. These details were provided with annotation instructions before beginning the annotation process.

The Inter Annotator Agreement (IAA) was computed using **Fleiss' Kappa score** and pairwise **Cohen's Kappa**. The Fleiss' Kappa score is found to be **0.77** which shows substantial agreement. The pairwise Cohen's Kappa scores are reported in Table 1. The average Cohen's Kappa is found to be **0.78** which shows substantial agreement. More details of manual annotation are in Appendix A.1.

[2] https://huggingface.co/tuner007/pegasus_paraphrase

[3] BERTScore uses cosine similarity to match words in candidate and reference sentences by leveraging the pre-trained BERT contextual embeddings.

[4] BARTScore performs text evaluation as a text generation problem. Since the average log-likelihood is used for the target tokens, the BARTScores are less than zero.

	Cohen's Kappa
Annotator A and B	0.82
Annotator B and C	0.75
Annotator A and C	0.77
Average pair-wise score	**0.78**
Fleiss' Kappa score	**0.77**

Table 1: Fleiss' Kappa and Cohen's Kappa scores for manual annotations of paraphrased data.

4.2.3 Dataset Statistics

The distribution of intrinsic, extrinsic, and non-hallucinated instances in XSumFaith++ is given in Table 2. We split the XSumFaith++ dataset in the ratio of 60:20:20 for the train, dev, and test set respectively ensuring equal distribution of hallucination types present. The detailed instances of the split are shown in Table 3. We make sure that the test set instances are not present in either training or validation splits.

Hallucination Type	No. of instances
Intrinsic	$7,527$
Extrinsic	$7,860$
Non-Hallucinated	$7,282$
Total Instances	**22,669**

Table 2: Distribution of intrinsic, extrinsic, and non-hallucinated instances in XSumFaith++.

The instances of intrinsic hallucinations in XSumFaith++ are very low as can be seen from Table 3. To create more instances containing intrinsic hallucination, we employ a perturbation strategy similar to Kryscinski et al. (2020). We used spaCy [5] to identify the named entity types present in the summary and the corresponding document. We limit our perturbation to the following named entity types: GPE, PERSON, ORG, NORP and CARDINAL. This is done because we observe that 81% of the intrinsic hallucinations happen due to incorrect entities of these 5 types. After identifying the named entities in the summary, we replace them with other named entities of the same type from the document. This causes a direct contradiction in the summary with respect to the corresponding document which makes the summary intrinsically hallucinated. One example of replacing a named entity to create an intrinsically hallucinated summary is given below-

Non-hallucinated summary: *BBC Sport is showing coverage of the EuroBasket warm-up game between **Great Britain** and Greece at the Copper Box in London.*

Intrinsically hallucinated summary: *BBC Sport is showing coverage of the EuroBasket warm-up game between **Turkey** and Greece at the Copper Box in London.*

The named-entity ***Great Britain*** in the non-hallucinated summary is replaced with ***Turkey*** which was mentioned in the corresponding document. *Turkey* is a named-entity of the same types as *Great Britain* i.e., GPE.

We put a maximum limit of 14 replacements per named entity. This oversampling approach results in $4,990$ instances of intrinsic hallucinations. These instances are added to existing intrinsic hallucination data which adds up to $7,527$ instances. The data distribution after this oversampling is as follows: intrinsic (33.2%), extrinsic (34.7%), and non-hallucinated (32.1%). The detailed distribution along with train-dev-test split instances is shown in Table 3.

	Hallucination Type (%)	Train (60%)	Dev (20%)	Test (20%)
3-class	Intrinsic (14.4)	1509	536	492
	Extrinsic (44.6)	4542	1666	1652
	Non-Hallucinated (41.2)	4348	1483	1450
3-class (os)	Intrinsic (33.2)	4361	1701	1465
	Extrinsic (34.7)	4542	1666	1652
	Non-Hallucinated (32.1)	4348	1483	1450

Table 3: Distribution of Train, Dev, and Test instances in XSumFaith++ dataset. The last row, 3-class (os), shows the data distribution after oversampling the class of intrinsic hallucination.

5 Methodology

Using Textual Entailment for Hallucination Detection: Textual entailment or NLI (Natural Language Inference) is the task of classifying a hypothesis, H, as a contradiction, neutral, or entailment given a premise P. We argue that the three types of textual entailment can be mapped to the three types of hallucination *viz.* intrinsic, extrinsic, and non-hallucination (Section 2.3).

Task Formulation: For a document **d** (considered as the premise), its summary **s** (considered as

[5] We used spaCy v3.5.0 to identify NER categories.

the hypothesis) and a set of entailment labels y_0, y_1 and y_2 corresponding to contradiction, neutral, or entailment respectively, we can mathematically formulate the three-class hallucination classification as:

$$y^* = \underset{y \in \{0,1,2\}}{argmax}\, P(y|\mathbf{d}, \mathbf{s}; \theta) \quad (1)$$

$$P(y|\mathbf{d}, \mathbf{s}; \theta) = \rho(f(E(\mathbf{d}, \mathbf{s}))) \quad (2)$$

where E and f represent the transformer-based encoder trained for the NLI task and the feed-forward neural network (classification head) respectively, θ represents the weights from both E and f and ρ represents the softmax function. We use the multi-class cross-entropy loss as our loss function. We map the obtained entailment labels y_0, y_1 and y_2 to hallucination categories *viz.* intrinsic, extrinsic, and non-hallucination.

6 Experiments and Results

To perform the three-way classification, we use the following models which are fine-tuned on NLI datasets: DeBERTa-base[6] and RoBERTa-large[7]. We use two experimental settings: zero-shot (ZS) and fine-tuning (FT). In a zero-shot setting, we use the models fine-tuned on NLI directly for the task of hallucination detection without training them on any hallucination data. For the fine-tuning experiments, these DeBERTa and RoBERTa models are further fine-tuned using the XSumFaith++ dataset. The results are shown in Table 4 as DeBERTa (ZS), RoBERTa (ZS), DeBERTa (FT) and RoBERTa (FT). The results are shown on the test set of the XSumFaith++ dataset. Further details of the experimental setup are given in Appendix B.

RoBERTa-large model shows better performance in both zero-shot and fine-tuned experiment settings. The best performance is seen in the fine-tuned RoBERTa-large model with an F1 score of **0.81**. This three-way classification performance cannot be compared with the existing factual consistency detection models as all of them perform a binary classification (i.e., hallucinated vs. non-hallucinated). To compare the performance with the current models, we use our approach for binary classification (*viz.* hallucinated vs. non-hallucinated) by combining intrinsic and extrinsic categories into a single category, i.e., hallucinated.

Model	P	R	F1
DeBERTa-base (ZS)	0.69	0.64	0.66
RoBERTa-large (ZS)	0.77	0.73	0.75
DeBERTa-base (FT)	0.76	0.73	0.74
RoBERTa-large (FT)	**0.80**	**0.81**	**0.81**

Table 4: Results of zero-shot (ZS) experiments and fine-tuned experiments (FT) for 3-class hallucination classification. The results are shown on the test set of the XSumFaith++ dataset.

We consider the following baselines[8]:

- **QuestEval**: (Scialom et al., 2021) proposes a QA-based metric that aggregates answer overlap scores from selected spans. The questions are of two types - derived from the source and answered using the summary, and derived from the summary and answered using the source. We compute the best threshold value over the scores generated by the QuestEval model for all data instances. The scores above the threshold are treated as non-hallucinated and the scores below the threshold are treated as hallucinated.

- **SummaC-ZS**: (Laban et al., 2022) uses sentence-level entailment scores between the summary and the corresponding document. The maximum entailment score for each summary sentence is computed and the final score is calculated by averaging over all summary sentences.

- **SummaC-Conv**: (Laban et al., 2022) extends SummaC-ZS and creates a histogram by replacing the max operation with a binning of the entailment scores between the source sentences and summary sentence. The summary sentence scores are produced by passing the histogram through a 1-D convolution layer.

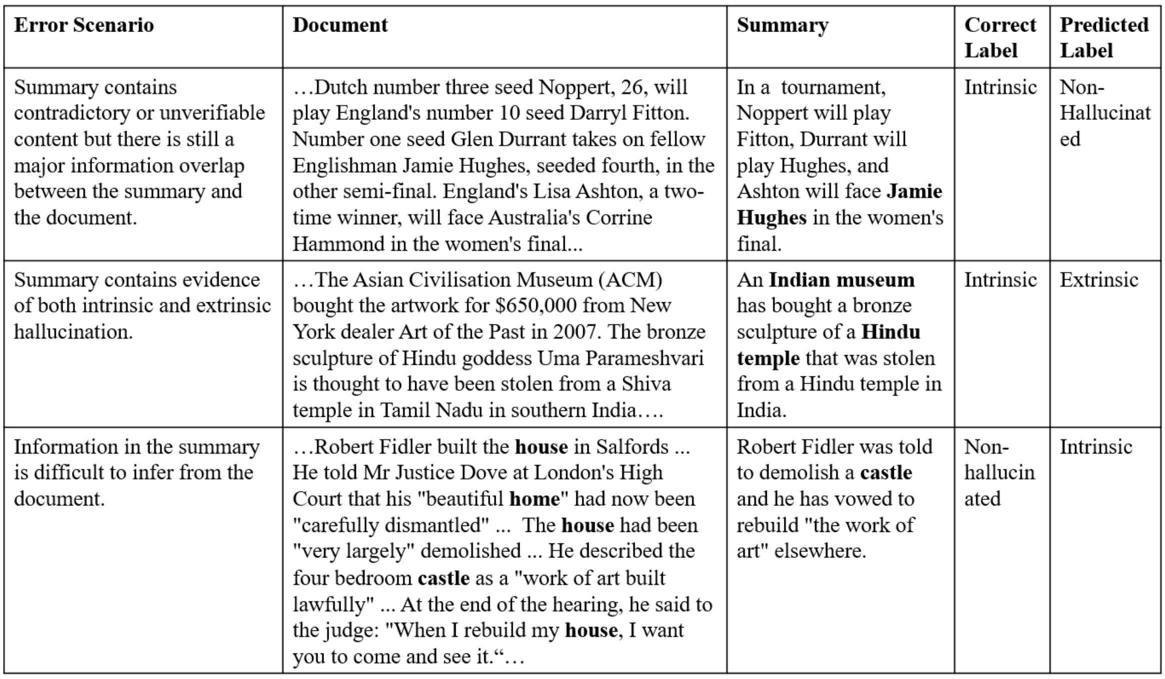
- **QAFactEval**: (Fabbri et al., 2022) is also a QA-based metric that includes optimized question-answering, generation, and answer-overlap components.

The results of this comparative study are shown in Table 5. The RoBERTa-large model outperforms the current factuality models by **12%** on the XSumFaith dataset and by **5.48%** on the XSumFaith++

[6] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c

[7] https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli

[8] We do not consider the DAE model (Goyal and Durrett, 2021) in this comparison as it is trained on XSumFaith itself.



Error Scenario	Document	Summary	Correct Label	Predicted Label
Summary contains contradictory or unverifiable content but there is still a major information overlap between the summary and the document.	…Dutch number three seed Noppert, 26, will play England's number 10 seed Darryl Fitton. Number one seed Glen Durrant takes on fellow Englishman Jamie Hughes, seeded fourth, in the other semi-final. England's Lisa Ashton, a two-time winner, will face Australia's Corrine Hammond in the women's final...	In a tournament, Noppert will play Fitton, Durrant will play Hughes, and Ashton will face **Jamie Hughes** in the women's final.	Intrinsic	Non-Hallucinated
Summary contains evidence of both intrinsic and extrinsic hallucination.	…The Asian Civilisation Museum (ACM) bought the artwork for $650,000 from New York dealer Art of the Past in 2007. The bronze sculpture of Hindu goddess Uma Parameshvari is thought to have been stolen from a Shiva temple in Tamil Nadu in southern India….	An **Indian museum** has bought a bronze sculpture of a **Hindu temple** that was stolen from a Hindu temple in India.	Intrinsic	Extrinsic
Information in the summary is difficult to infer from the document.	…Robert Fidler built the **house** in Salfords ... He told Mr Justice Dove at London's High Court that his "beautiful **home**" had now been "carefully dismantled" ... The **house** had been "very largely" demolished ... He described the four bedroom **castle** as a "work of art built lawfully" ... At the end of the hearing, he said to the judge: "When I rebuild my **house**, I want you to come and see it.“…	Robert Fidler was told to demolish a **castle** and he has vowed to rebuild "the work of art" elsewhere.	Non-hallucinated	Intrinsic

Figure 5: Error Analysis of results of RoBERTa-large NLI model fine-tuned on XSumFaith++ for 3-class hallucination classification. This figure shows three types of error scenarios along with an example of each scenario containing a document, summary, correct label and the predicted label.

dataset which proves the efficacy of the zero-shot hallucination classification using models fine-tuned for NLI (Table 5).

Model	XSF	XSF++
QuestEval	55.99	52.71
SummaC-ZS	62.42	68.44
SummaCConv	63.51	61.11
QAFactEval	65.02	74.64
RoBERTa (ZS)	**76.98**	**80.12**

Table 5: Comparison of balanced accuracy with current consistency checking models (in %). The models are tested on the entire XSumFaith (XSF) and XSumFaith++ (XSF++) datasets in a zero-shot (ZS) manner.

7 Qualitative Analysis

As seen from the results (Tables 4 and 5), models fine-tuned on NLI datasets are very effective in detecting hallucination. An example where the RoBERTa-large model correctly detects the presence of intrinsic hallucination is:

Document: *... Mr Gustafsson, 42, is being flown back to Sweden from Africa, Sweden's Foreign Minister Margot Wallstram said in a statement on Monday. ...*

Generated Summary: ***Swedish foreign minister*** *Johan* ***Gusttrom*** *has been freed after two suspected kidnappings in a raid in Mali, the foreign ministry says.*

Correct label: *Intrinsic*

Predicted label: *Intrinsic*

To understand where our model makes incorrect predictions, we perform an error analysis on the results obtained using the fine-tuned RoBERTa-large model (Table 4) for 3-class classification. We show the confusion matrices for the 3-class classification done using the fine-tuned RoBERTa-large model in Appendix C. We took 50 random samples of each error case and analyzed them. We find the following three types of error scenarios which are shown with examples in Figure 5.

1. Incorrect classification of summaries as non-hallucinated is seen when the summary contains contradictory or unverifiable content but there is still a major information overlap between the summary and the document. In the first row of Figure 5, according to the document, the first two competitors are correctly mentioned in the summary, but ***Corrine Hammond*** should have been mentioned in place of ***Jamie Hughes***.

2. The instances where the intrinsically hallucinated summary is predicted as extrinsically hallucinated or vice-versa are seen when the

generated summary contains evidence of both intrinsic and extrinsic hallucination. In the second row of Figure 5, according to the document, *bronze sculpture of a* ***Hindu temple*** is an intrinsic hallucination. Whereas, ***Indian museum*** is extrinsically hallucinated.

3. The instances where non-hallucinated summaries are incorrectly classified as intrinsic or extrinsic happen when information in the summary is difficult to infer from the document. In the third row of Figure 5, the non-hallucinated summary is incorrectly marked as intrinsic. In the corresponding document, the word ***house*** is used. The house is referred to as ***castle*** very late in the document which makes it difficult to resolve.

Further analysis comparing the performance of our approach with the best baseline, QAFactEval (Fabbri et al., 2022), is shown in Appendix C.

8 Conclusion and Future Work

We propose an entailment-grounded approach to detect types of hallucination in summaries generated by abstractive summarizers. The models fine-tuned on NLI datasets show a **12%** increase in accuracy compared to the existing consistency-checking models. We also release a new dataset, **XSumFaith++**, containing a balanced number of hallucinated and non-hallucinated instances. We do this by augmenting the XSumFaith dataset with $7,282$ non-hallucinated data instances. XSumFaith++ dataset can be used to aid further research in hallucination detection.

The takeaway from our work is that zero-shot entailment-grounded classification of hallucination works better than the existing factuality detection models for abstractive summarization. In this work, we do not tackle scenarios where both intrinsic and extrinsic hallucinations are present in the same instance. This is a potential future direction of our work. We also plan to explore approaches to generalize hallucination detection for all NLG tasks.

References

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. corr abs/1901.08079 (2019). *arXiv preprint arXiv:1901.08079*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):1–23.

Jan Dirk Blom. 2010. *A dictionary of hallucinations*. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Stéphane Clinchant, Cyril Goutte, and Eric Gaussier. 2006. Lexical entailment for information retrieval. In *Advances in Information Retrieval: 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006. Proceedings 28*, pages 217–228. Springer.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. 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.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. 2008. The fourth pascal recognizing textual entailment challenge. In *TAC*.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. 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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551.

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.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Annie Louis and Joshua Maynez. 2022. Opinesum: Entailment-based self-training for abstractive opinion summarization. *arXiv preprint arXiv:2212.10791*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas 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.

Khalil Mrini, Franck Dernoncourt, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021. Joint summarization-entailment optimization for consumer health question understanding. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 58–65, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

A Nenkova and K McKeown. 2011. Automatic summarization, foundations and trends in information retrieval.

OpenAI. 2023. Gpt-4 technical report.

Amarnath Pathak, Riyanka Manna, Partha Pakray, Dipankar Das, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2021. Scientific text entailment and a textual-entailment-based framework for cooking domain question answering. *Sādhanā*, 46:1–19.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

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.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2022. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, 36(1):318–354.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Sabine Wehnert, Sayed Anisul Hoque, Wolfram Fenske, and Gunter Saake. 2019. Threshold-based retrieval and textual entailment detection on legal bar exam questions. *arXiv preprint arXiv:1905.13350*.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404.

A Dataset

A.1 Manual evaluation of paraphrased summaries

The statistics of the manual evaluation of 500 random samples from the XSumFaith++ dataset are shown in Table 1. The common misaligned instances are not removed from the dataset as their IAA scores indicate substantial agreement.

Annotator	Aligned (%)	Misaligned (%)
Annotator A	446 (89.2%)	54 (10.8%)
Annotator B	442 (88.4%)	58 (11.6%)
Annotator C	433 (86.6%)	67 (13.4%)
Average (%)	**88.07**	**11.93**

Table 1: Overview of manual annotations for the analysis of 500 random samples of paraphrased data.

B Experimentation Details

B.1 Experimental Setup

For experiments, we use the NVIDIA A100-SXM4-80GB GPU. Table 2 contains further details of the number of parameters and run time for 10 epochs.

Model	#Parameters	Run time
DeBERTa-base	$\sim 184M$	$\sim$ 26 mins
RoBERTa-large	$\sim 355M$	$\sim$ 45mins

Table 2: Additional details of the models along with their number of parameters and run time.

B.2 Hyperparameters

For results on the XSumFaith++ dataset, we did the hyperparameter search manually as follows: number of epochs = [5, 7, 10, 15, 20, 25], learning rate = [1e-5, 5e-5, 1e-4, 1e-6, 2e-4, 5e-4], and batch size = [4, 8, 16, 32, 64]. The following hyperparameter values were chosen based on the best-performing model: Number of epochs: 20, Learning rate: 1e-6, Batch size: 32.

C Analysis

The confusion matrix and normalized confusion matrix for 3-class classification using the RoBERTa model fine-tuned on the XSumFaith++ dataset are shown in Figure 1 and Figure 2 respectively.

Table 3 shows few examples from XSumFaith dataset comparing the performance of RoBERTa model finetuned for NLI task and the performance of the best SOTA model QAFactEval (Fabbri et al., 2022).

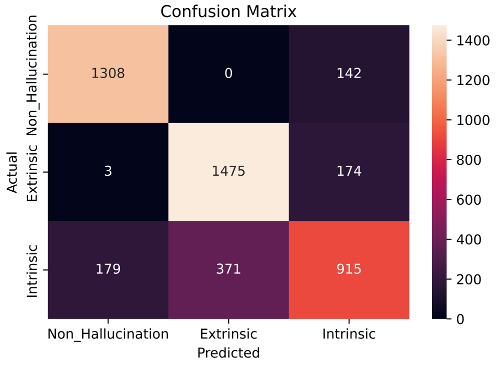


Figure 1: Confusion matrix of fine-tuned RoBERTa-large model for 3-class hallucination detection.

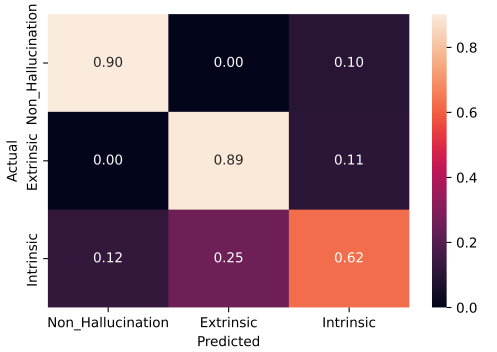


Figure 2: Normalized Confusion matrix of fine-tuned RoBERTa-large model for 3-class hallucination detection.

References

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

bbcid	Summary	QAFactEval's Prediction	Our model's Prediction
38324244	A Sydney police officer and huge star wars fan has become a local hit after creating a Darth Vader costume painted with the Australian flag.	Hallucinated ✗	Non-Hallucinated ✓
35360960	The family of a man killed in a crash with an 87-year-old who was travelling the wrong way on the m1 have called for older drivers to be retested.	Non-hallucinated ✗	Hallucinated ✓
40965536	BBC sport is showing live coverage of the Eurobasket warm-up game between Great Britain and Greece at the copper box in London on Saturday 19 august.	Hallucinated ✓	Non-Hallucinated ✗
36207647	A man has died in a collision between a tractor and a motorcycle in Lincolnshire.	Non-Hallucinated ✓	Non-Hallucinated ✓

Table 3: Few examples from the XSumFaith dataset showing a comparison between the predictions made by QAFactEval (Fabbri et al., 2022) model and our RoBERTa-large model. We show the bbcid of the document, it's summary and the 2-class classification made by the QAFactEval model and our model. The green tick (✓) indicates correct prediction and the red cross (✗) indicates wrong prediction.