# How Faithful Are Your Summaries? A Study of NLI-Based Verification in Portuguese

**Felipe S. F. Paula**[1], **Matheus Westhelle**[1], **Maria Cecília M. Corrêa**[1], **Luciana R. Bencke**[1], **and Viviane P. Moreira**[1]

[1]Institute of Informatics - UFRGS
Porto Alegre - Brazil

{fsfpaula,mwesthelle,mcmcorrea,lrbencke,viviane}@inf.ufrgs.br,

***Abstract.*** *Abstractive summarization systems often generate content that is not supported by the source text, making faithfulness verification a critical evaluation step. In this paper, we investigate the reliability of Natural Language Inference (NLI) methods for detecting summary faithfulness in Portuguese. Our contribution is two-fold: (i) we introduce* VERISUMM*, the first large-scale dataset for summary faithfulness detection in Portuguese, and (ii) we benchmark several NLI-based approaches applied to faithfulness detection. Our experiments revealed that zero-shot models exhibit low to moderate performance and that fine-tuning improves results. However, our error analysis showed that NLI models rely heavily on lexical overlap heuristics, limiting their effectiveness.*

## 1. Introduction

Abstractive summarization aims to mimic the way in which humans summarize text by not simply extracting verbatim portions of the input document. Instead, it generates new sentences that capture the key ideas of the original text via rephrasing, rewording, and synthesizing information [Nallapati et al. 2016]. Despite its superior results in comparison to extractive summarization [Feijo and Moreira 2019, El-Kassas et al. 2021, Sharma and Sharma 2022], abstractive summaries may not be faithful to the source text due to hallucinations – a known problem in generative models [Ji et al. 2022]. Faithfulness is crucial for developing reliable summarization systems. We define an unfaithful summary as a summary that presents information not supported by the source text.

Automatic summarization in Portuguese has come a long way. The development of datasets for this language paved the way for the establishment of new summarization models. Datasets such as Temário [Pardo and Rino 2003] and CST-News [Cardoso et al. 2011] were early landmarks in the history of pt-BR summarization. More recently, the release of large-scale datasets such as XL-Sum [Hasan et al. 2021] and RecognaSumm [Paiola et al. 2024] enabled the training of more powerful models with hundreds of millions of parameters, such as PTT5 [Piau et al. 2024]. However, the factual consistency of the abstractive summaries has largely been out of discussion. As we will discuss later, Natural Language Inference (NLI) is an important way to detect hallucinations in summaries. We are aware of at least one work that used NLI to evaluate faithfulness in summarization in pt-BR [Feijo and Moreira 2023]. Given the importance of the NLI-based summary consistency evaluation methods and the emerging models for summarization in pt-BR, we aim to answer the following question *To what extent can we rely on NLI-based summary faithfulness detection in the Portuguese language?*

In this paper, we developed VERISUMM, a dataset with documents, summaries, and their faithfulness labels. Then, we implemented several NLI models and tested them on VERISUMM. Our experimental results showed that zero-shot models present a low to moderate performance on the task. This highlights the difficulty of the task and the current state of pt-BR NLI models. We also find that fine-tuning a long premise model on the task data improves the performance. Additionally, through extensive error analysis, we show that models rely on lexical overlap heuristics to make predictions. Our data and prompts are available at `https://github.com/felipesfpaula/verisumm`.

## 2. Related Work

**Faithfulness Evaluation in Abstractive Summarization**. The ROUGE [Lin 2004] metric has been used for decades to evaluate the quality of summaries with respect to a reference. However, previous work showed that this metric correlates very poorly with human judgments of factuality [Yuan et al. 2021, Zhong et al. 2022] (Spearman's $\rho < 0.112$). This led to the creation of different families of new metrics. There were approaches based on information extraction that checked the presence of similar information pieces in summaries and source texts [Cao et al. 2018, Goodrich et al. 2019, Nan et al. 2021]. Similarly, but in a more general way, Question Answering (QA) methods matched information excerpts between the two texts. These QA-based methods [Durmus et al. 2020, Wang et al. 2020, Scialom et al. 2021] generate question-answer pairs from summaries, then the questions are answered using the source text as input, and the answers are compared. More recently, with the advent of Large Language Models (LLMs), the paradigm shifted towards using these bigger models for the task. In this line, it was shown that open-source models preferred factually correct summaries [Tam et al. 2023]. Also, there were works [Shen et al. 2023] that showed a high correlation between human judges and LLM judges while advising against using LLMs as a replacement for human annotators due to bias issues.

**NLI and NLI-based Faithfulness Detectors**. The NLI task involves determining a logical relationship between a premise and a hypothesis. A common way of approaching this problem is to classify the premise-hypothesis pair as entailment, neutral, or contradiction [Bowman et al. 2015, Williams et al. 2018], although different classification approaches exist [Fonseca et al. 2016, Real et al. 2020]. NLI can be used to the benefit of other NLP tasks. For example, in verifying the answers of QA systems [Chen et al. 2021], or detecting machine-generated text [Shastry et al. 2025]. Modeling summary faithfulness detection as an NLI task is a natural choice since the source text is entailed by faithful summaries. However, this conversion is not without its problems. Usually, NLI models are trained on sentence-level premises and hypotheses, which makes using standard models unfeasible. A line of research tries to solve this problem by proposing longer premise NLI models [Mishra et al. 2021, Utama et al. 2022]. In an orthogonal direction, there is a line of works that use sentence-level models in conjunction with techniques that decompose the longer texts into smaller segments [Laban et al. 2022, Schuster et al. 2022, Zhang et al. 2024]. In this work, we compare both types of NLI approaches applied to summary faithfulness detection. As far as we know, we are the first to propose a long premise NLI dataset and compare faithfulness detection methods for Portuguese.

**LLM as a judge**. The high quality of current state-of-the-art LLMs and the relatively cheap costs of use have driven the idea of using them as dataset annotators. Previ-

ous studies report that the agreement between LLMs such as GPT-4 and humans is higher than the agreement between humans in specific tasks [Zheng et al. 2023]. In other studies, the LLMs' agreement with humans is very high but still lower than between humans [Sottana et al. 2023, Thakur et al. 2025]. In related areas such as Information Retrieval, LLMs have been used to generate query-document relevance judgments [Faggioli et al. 2023, Thomas et al. 2024, Bueno et al. 2024], also reaching a high quality. The use of LLMs as evaluators is not without its limitations, as they have been shown to display their own hallucinations, systematic biases, and a lack of robustness [Gu et al. 2025]. In the absence of a human-evaluated summary faithfulness benchmark, we propose an LLM-annotated dataset as the best approximation. While it may not be optimal, it can help the community to uncover problems in the summarization and summarization evaluation models.

This research aims to close some important gaps in the NLP literature. First, it addresses the fact that there are no publicly available pt-BR faithfulness datasets. Second, as far as we know, there is currently no pt-BR long-premise NLI resource, which we also built. Third, we bring to the front the discussion of sentence-based vs. long premise-based NLI faithfulness detection. And finally, we perform a detailed error analysis and reveal that systems rely on lexical overlaps. This discussion was not found in the English language processing literature since the NLI models are more advanced.

## 3. VERISUMM– a dataset for faithfulness evaluation

In order to be able to evaluate faithfulness in summaries, we compiled VERISUMM, a dataset with documents, their summaries, and a binary label indicating whether the summary is faithful. A visual description of the dataset construction can be seen in Figure 1. Inspired by TrueTeacher [Gekhman et al. 2023], we generated summary data using different PTT5v2 [Piau et al. 2024] models. More specifically, we fine-tuned PTT5v2 models of different sizes (small, base, and large) in two news summarization datasets in Portuguese: XL-sum [Hasan et al. 2021] and RecognaSumm [Paiola et al. 2024]. Different from TrueTeacher, we also experimented with varying the decoding approach of the summary. For a stricter generation, we used beam search with five beams, and for a more creative generation, we used top-p sampling [Holtzman et al. 2020] with temperature of $1.0$ and $p = 0.85$. For test data, we selected $150$ source texts from each test set of XL-Sum and RecognaSumm, and generated summaries using each of the three models fine-tuned on the respective dataset and the two decoding strategies, resulting in a total of 2,100 pairs, as presented in Table 1[1]. For training data and validation, we did a similar procedure, selecting $600$ source texts from the validation sets from both datasets and generating the summaries using the six summarization models. From the resulting pairs, we sampled 5 000 for training and 1 000 pairs for validation. All splits include gold summaries.

To evaluate the faithfulness of the summaries, we employed an LLM as a judge approach. We built a prompt around the faithfulness error taxonomy proposed for the FRANK benchmark [Pagnoni et al. 2021] and submitted it to GPT-4.1 through OpenAI's API. The faithfulness violation typology in FRANK is grounded in linguistic theories. In their annotation experiment, crowdworkers and experts achieved a high level of agreement, showing that this taxonomy helps to reduce the subjectivity involved in the task. We

[1] 2 (datasets) $\times$ 150 (texts) $\times$ 3 (models) $\times$ 2 (decodings) + (150 $\times$ 2) gold = 2,100

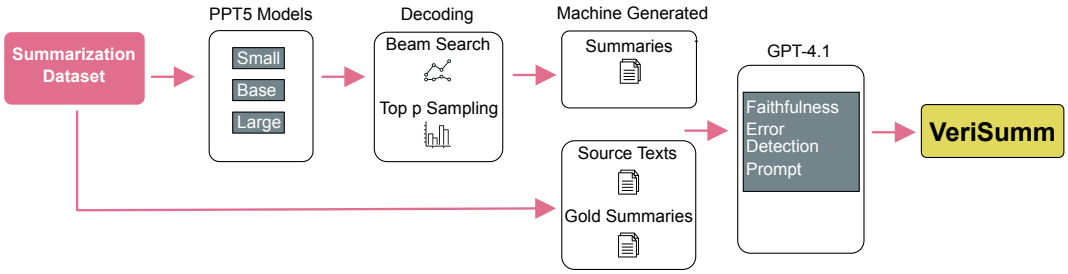


**Figure 1. VERISUMM construction diagram**

**Table 1. Dataset composition. Percentages are column-wise within each split. Numbers rounded to the second decimal.**

| Split | Pairs | Source (%) | | Generator Scale (%) | | | | Faithfulness (%) | |
|---|---|---|---|---|---|---|---|---|---|
| | | Recog | XL | Small | Base | Large | Gold | Faithful | Unfaith. |
| Train | 5 000 | 50 | 50 | 29 | 28 | 28 | 15 | 64 | 36 |
| Val | 1 000 | 47 | 53 | 30 | 26 | 28 | 16 | 64 | 36 |
| Test | 2 100 | 50 | 50 | 29 | 29 | 29 | 14 | 64 | 36 |
| All | 8 100 | 50 | 50 | 29 | 28 | 28 | 15 | 64 | 36 |

posit that having a well-grounded reference of which errors to look for can help steer the LLM in the direction of making the best decisions. Despite the resulting data being annotated with a fine-grained typology of errors, we binarize the results considering whether the summaries contain a faithfulness violation.

**Dataset Composition**. In Table 1 we can see the data statistics. The corpus is deliberately balanced across its main axes. Exactly half of the 8.1k instances come from RecognaSumm and half from XL-Sum, with the validation split showing only a slight 3-point drift. Automatic summaries generated by small, base, and large models each contribute roughly 28–30%, while an additional 15% of "gold" human summaries are included in every split, ensuring that no single scale dominates. Label distribution is stable, with 64% of summaries annotated as faithful and 36% unfaithful in train, validation, and test alike—providing a consistent, mildly imbalanced target that mirrors real-world error rates without skewing any single partition. Overall, the table shows that the dataset offers ample training material while preserving uniform composition across splits, making downstream comparisons straightforward and fair.

**Human Audit** To assess the quality of the labels produced by the LLM, we manually inspected a sample of the test set of VERISUMM. We sampled 75 instances labeled as unfaithful and 75 instances labeled as faithful. These 150 instances were analyzed by two human raters. They were asked if they agreed or disagreed with the label produced by the LLM. They had access to the fine-grained analysis and chain-of-thought explanation of the annotation. The inter-rater agreement between the two annotators was 0.33 (Cohen's $\kappa$), which suggests a fair to moderate agreement. Both humans agreed with the LLM on 122 of 150 cases (81%) and they both rejected the LLM label on 7 cases ($\approx$ 5%).

## 4. Models and Baselines for Faithfulness Detection

In order to evaluate how well NLI models can detect faithfulness, we implemented different models, which account for premises of varying lengths. Those are presented next.

### 4.1. Long Premise NLI model: PT-NLI-Long

**Corpus creation**. To construct a corpus suitable for training a long premise NLI model aimed at evaluating summary faithfulness, we transformed the SQuADv2 dataset [Rajpurkar et al. 2018], which was designed for QA, into an NLI format. SQuADv2 has both answerable and non-answerable questions (*i.e.,* questions that cannot be answered based on the context provided). Our methodology differs from prior approaches as follows: $(i)$ For unanswerable questions, we prompted an LLM to generate plausible yet incorrect span answers representative of flawed QA systems. Subsequently, we asked the LLM to convert these erroneous answer-question pairs into declarative sentences, labeling them as non-entailment. $(ii)$ For answerable questions, the LLM transformed correct question-answer pairs into declarative statements. To mitigate lexical overlap biases, each statement was paraphrased into five distinct sentences. $(iii)$ We then translated all instances into Brazilian Portuguese using the LLM, ensuring consistent translations between hypotheses and premises.

To further enrich the training data and enhance reasoning skills, we integrated additional NLI resources: a translated subset of ConTRoL [Liu et al. 2021], which emphasizes logical and textual comprehension challenges, and Portuguese NLI datasets ASSIN [Fonseca et al. 2016], ASSIN2 [Real et al. 2020], and InferBr [Bencke et al. 2024]. All datasets were standardized into binary entailment/non-entailment labels, resulting in a final training set comprising $169,963$ instances.

**Model Fine-tuning**. We fine-tuned the entire sequence-to-sequence model PTT5-large on this newly assembled dataset. As the model outputs tokens to perform classification, we mapped the entailment class to token "e" and the non-entailment class to token "n". The final prediction involves applying a softmax to the two-dimensional probability vector corresponding to these tokens. The model was fine-tuned for five epochs.

In Table 2, we can see the composition of the train/test corpora and the performance of the fine-tuned model. Overall, the performance is good, save for the ConTRoL test set, in which there is a significant drop. This is a clue that the model is not yet suitable to perform more sophisticated reasoning tasks.

**Table 2. PT-NLI-Long corpus statistics and model performance.**

| Dataset | Train Instances | Train Prop. | Test Instances | Test Prop. | Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|
| InferBR | 8,190 | 0.05 | 1,705 | 0.04 | 0.96 | 0.91 | 0.96 |
| ASSIN | 2,500 | 0.01 | 2,000 | 0.05 | 0.93 | 0.81 | 0.91 |
| ASSIN 2 | 6,500 | 0.04 | 2,448 | 0.06 | 0.91 | 0.86 | 0.96 |
| ConTRoL | 3,574 | 0.02 | 373 | 0.01 | 0.49 | 0.31 | 0.53 |
| SQuAD-NLI | 149,199 | 0.88 | 31,924 | 0.83 | 0.94 | 0.96 | 0.96 |
| **Total** | 169,963 | 1.00 | 38,450 | 1.00 | 0.94 | 0.94 | 0.96 |

### 4.2. Passage level NLI methods

**PT-NLI-Long Single Hypothesis** Although the premises can be long, this model was trained with a single sentence as hypothesis. However, many of the summaries are multi-sentence. To address this, we split the sentences of the summary and ran them individually through the model. The score of the passage is the minimum of the probabilities of the entailment class of each sentence. This strategy is in line with the fact that if a sentence is not supported by the source text, then the entire summary is unfaithful.

**PT-NLI-Long Single Hypothesis (fine-tuned)** To test whether supervision in the VeriSumm dataset helps the task of NLI-based faithfulness detection, we gather the sentences that the LLM judged as unfaithful to create non-entailment cases and the sentences the LLM judged faithful to create entailment cases. We further fine-tuned the PT-NLI-Long model on this data for 10 epochs.

### 4.3. Sentence level NLI methods

**SENTLI** [Schuster et al. 2022] handles zero-shot NLI over long premises by first splitting the document into individual sentences and scoring each sentence–hypothesis pair with a pre-trained NLI model. Since this technique needs an NLI model that predicts entailment, neutral, and contradiction, we fine-tune BERTimbau [Souza et al. 2020] on the InferBR [Bencke et al. 2024] dataset, which has annotations on the three classes. SENTLI's authors define a series of variants of this technique, and we report here the setup with the best performance.

**INFUSE** [Zhang et al. 2024] also uses an off-the-shelf NLI model to assess the faithfulness of summaries. However, it incrementally builds the supporting context for each hypothesis by greedily adding top-ranked sentences until the neutral-class probability reaches a local minimum, with optional reversed entailment and sub-sentence splitting to capture fragmentary evidence. We opted not to use sub-sentence splitting.

**SUMMAC** [Laban et al. 2022] assesses summary consistency by computing a matrix over all document–summary sentence pairs using a pre-trained NLI model. Two variants were considered: SUMMAC $_{ZS}$, a zero-shot approach that aggregates the matrix using a max-mean pooling strategy over entailment scores; and SUMMAC $_{CONV}$, a trained model that bins the matrix into score histograms and applies a 1-D convolution layer to better capture the distribution of evidence. SUMMAC $_{CONV}$ was trained on VERISUMM training split.

## 5. Faithfulness Detection Experiment

Our goal is to evaluate different NLI methods in detecting faithfulness in summaries. We applied the NLI models described in Section 4 on a binary classification task on the VERISUMM dataset described in Section 3.

### 5.1. Main Results

Table 3 reports balanced accuracy (BA), $F_1$, precision (P), recall (R), and ROC–AUC (AUC) for the six NLI-based faithfulness detectors evaluated on the VERISUMM test set. The positive class corresponds to *unfaithful* summaries. Overall, performance remains moderate across all systems, with none of the metrics exceeding 0.75, underscoring the difficulty of sentence-level faithfulness detection in Portuguese.

The best system, PT-NLI-Long (FT), improves over the strongest zero-shot baseline by +0.05AUC and +0.02$F_1$, but the absolute ceiling remains low (<0.8 on any metric). The heterogeneous precision–recall profiles highlight a crucial design choice: tasks prioritizing the detection of *any* unfaithfulness might adopt SENTLI-style thresholds, whereas automated evaluation pipelines that must avoid excessive false alarms should prefer INFUSE or PT-NLI-Long (FT).

**Ranking by discrimination ability.** PT-NLI-Long (FT) exhibits the highest AUC (**0.75**) and the best $F_1$ (**0.61**), suggesting that its probability scores separate faithful from unfaithful summaries more effectively than competing models. INFUSE yields the second-best AUC (0.71) but a lower $F_1$ (0.59), indicating worse performance when a decision threshold is applied.

**Threshold-dependent trade-offs.** A closer look at P and R reveals divergent operating characteristics. SENTLI and SUMMAC achieve the largest recall values (0.91 and 0.89, respectively) at the cost of very low precision (0.41 for both), confirming their tendency to over-flag summaries as unfaithful. Conversely, INFUSE adopts a more conservative policy: it attains the highest precision among baselines (0.50) but only middling recall (0.73), resulting in inflated plain accuracy (0.64) without gains in balanced accuracy.

**Balanced accuracy perspectives.** Balanced accuracy neutralizes the 3:2 class imbalance of the test set and therefore offers a fair comparison across systems. PT-NLI-Long (FT) and INFUSE are tied at the top (BA =0.66), followed closely by SUMMAC-Conv (0.63). Vanilla SUMMAC and SENTLI trail with BA= 0.60, reinforcing the observation that their high recall is offset by a surge in false positives.

**Sentence decomposition helps but is not sufficient.** Replacing document-level entailment with sentence decomposition (SUMMAC-Conv) yields consistent gains over vanilla SUMMAC on every metric (e.g. +0.03BA and +0.09AUC), yet it still underperforms the best T5 variant. This suggests that finer granularity mitigates, but does not eliminate, the reliance on superficial lexical cues.

**Table 3. Performance of summary faithfulness detectors on VeriSumm test split. Metrics include balanced accuracy, F1-score, precision, recall, and ROC-AUC.**

| Model | Balanced Acc. | F1 | Precision | Recall | ROC-AUC |
|---|---|---|---|---|---|
| INFUSE | **0.66** | 0.59 | **0.50** | 0.72 | 0.71 |
| SENTLI | 0.60 | 0.57 | 0.41 | **0.91** | 0.59 |
| SUMMAC | 0.60 | 0.56 | 0.41 | 0.89 | 0.58 |
| SUMMAC-Conv | 0.63 | 0.57 | 0.45 | 0.76 | 0.65 |
| PT-NLI-Long | 0.61 | 0.56 | 0.44 | 0.78 | 0.69 |
| PT-NLI-Long (FT) | **0.66** | **0.61** | 0.48 | 0.83 | **0.75** |

## 6. Error & Bias Analysis

In this section, we analyze the types of errors made by the NLI models in evaluating faithfulness and investigate the strategies that they use to make their predictions.

### 6.1. Model Error Analysis

**Error overlap analysis**. Before checking whether the models display systematic biases, we test if the faithfulness detectors make the same mistakes. To do this, we calculated

the Jaccard Index between instances that were incorrectly predicted by the models. The results for each pair of models are shown in a matrix in Figure 2(b). The minimum score is 0.3, which indicates that the detectors share a significant number of errors. Furthermore, sentence-based NLI detectors (INFUSE, SUMMAC, and SENTLI) have similar sets of errors. Long premise models present more dissimilar error patterns. These behaviors indicate that sentence decomposition and long premise models may be complementary.

**Lexical Overlap**. We now move our investigation to verify whether the models present systematic biases when making predictions. In particular, we are interested in checking if superficial lexical overlaps are a driving heuristic behind the predictions. To achieve our goal, we define a lexical overlap measure that aligns with the kind of processing the NLI models make. Algorithm 1 implements a simple *min–max* matching strategy to quantify how well each summary sentence is covered by the source text, using the ROUGE-1 metric. Concretely, for each summary sentence $t_i$, we compute its ROUGE-1 score against every source sentence $s_j$ and retain the maximum score $M_i = \max_{1 \le j \le n} \mathrm{ROUGE1}(t_i, s_j)$. This value represents the best lexical overlap between $t_i$ and any source sentence. Once all maximal match scores $M_1, \ldots, M_m$ are computed, the algorithm returns their minimum, *i.e.,* $\min_{1 \le i \le m} M_i$. This value reflects the weakest coverage among all summary sentences—if even one sentence in the summary has low lexical overlap with the source, the final score will be low.

**Algorithm 1** Minimum–Maximum ROUGE-1 Matching

**Require:** $\mathrm{SRC} = (s_1, \ldots, s_n)$ ▷ source text sentences
**Require:** $\mathrm{SUM} = (t_1, \ldots, t_m)$ ▷ summary sentences
**Ensure:** $\min_{i=1,\ldots,m} \max_{j=1,\ldots,n} \mathrm{ROUGE1}(t_i, s_j)$

```
for i ← 1 ... m do
    M_i ← 0
    for j ← 1 ... n do
        M_i ← max(M_i, ROUGE1(t_i, s_j))
    end for
end for
return min_{i=1,...,m} M_i
```

**Dataset Slicing**. To better understand the differences between the instances that the models predict correctly and incorrectly, we took two slices of the test data. In the first slice, which we call "Six Errors", we placed the instances that were incorrectly predicted by all models, and in the second slice, called "No errors", we placed the instances that were correctly predicted by all models. For each slice and ground truth label, we measured the lexical overlap based on the *min–max* strategy presented in Algorithm 1. The results are in Figure 2(a). We can see the predictions follow a structured pattern. In the "No errors" slice, for high overlap, models predicted *faithful* and were right, for low overlap, models predicted *unfaithful* and were right as well. However, in the "Six Errors" slide, when the ground truth is *faithful* and has a low overlap, the models misclassified these cases as *unfaithful*. Additionally, still in the "Six Errors" slice, when the ground truth is *unfaithful* and has a higher overlap, the models misclassify the instances as *faithful*. We can compute the difference $\Delta_{\text{slice}} = \overline{r}_{\text{faithful}} - \overline{r}_{\text{unfaithful}}$, where $\overline{r}_{\text{label}}$ is the mean ROUGE1-based measure for the label. We have $\Delta_{\text{No Error}} = 0.13$ and $\Delta_{\text{Six Errors}} = -0.06$. A positive $\Delta$


(a) Min-Max ROUGE-1 scores (Alg 1) (b) Jaccard similarity for inaccurate predictions

**Figure 2. Analysis of the prediction errors made by the NLI models**

means faithful summaries exhibit higher lexical overlap than unfaithful ones, and a negative otherwise. This is a further clue that the splits present opposing behaviors. While this analysis is promising, we couldn't reach statistical significance for the difference between the *faithful* and *unfaithful* groups in the "Six Errors" slice. We further explore the lexical overlap in the next section.

### 6.2. Surrogate model analysis

Surrogate analysis approximates the decision boundary of a complex or opaque system with a simpler, interpretable model. By fitting the black–box predictions against carefully chosen explanatory variables, we can quantify how much each variable contributes to the observed behaviour, without altering the original detector. In the context of faithfulness detection, this technique allows us to pinpoint whether a model is driven by *lexical shortcuts* (e.g. n-gram overlap) or by deeper semantic signals. Our objective is to test the hypothesis that lexical information constitutes the primary heuristic employed by the six NLI-based faithfulness detectors we are studying. We therefore build three logistic-regression surrogates per system:

$$\textbf{Lexical Model:} \quad \text{logit}(P(y=1)) = \beta_0 + \beta_1 \cdot \text{Len} + \beta_2 \cdot \text{Lex}$$

$$\textbf{Semantic Model:} \quad \text{logit}(P(y=1)) = \beta_0 + \beta_1 \cdot \text{Len} + \beta_2 \cdot \text{Lex} + \beta_3 \cdot \text{Sem}$$

$$\textbf{Gold Model:} \quad \text{logit}(P(y=1)) = \beta_0 + \beta_1 \cdot \text{Len} + \beta_2 \cdot \text{Lex} + \beta_3 \cdot \text{Sem} + \beta_4 \cdot \text{Gold}$$

where Len is the summary length, Lex is the lexical overlap (*e.g.,* ROUGE-1), Sem is the semantic overlap (*e.g.,* BERTScore), and Gold is the ground-truth label.

**Findings.** Surrogate performance is assessed on the binary prediction vectors of the detectors using the Area Under the ROC Curve (AUC) for *discrimination* and the Akaike Information Criterion (AIC) for *parsimony*. The results are in Table 4. Across all six detectors, the *Lex* model already achieves *AUC* 0.66–0.72, indicating that length and ROUGE-1 alone recover most of the black–box decision boundary. Adding BERTScore (*sem*) yields only a marginal median gain of $+0.7$ pp AUC, while raising AIC by $\approx 2$ points; the complexity penalty therefore outweighs the benefit. Injecting the oracle gold label (*all*) improves AUC by 3–6 pp, but again increases AIC ($\approx 4$ points), showing that even perfect task knowledge explains far less variance than lexical cues.

**Table 4. Surrogate model evaluation for summary faithfulness detectors showing AIC (lower is better) and AUC (higher is better) for different sets of features.**

| Model | Lexical | | Semantic | | Gold | |
|---|---|---|---|---|---|---|
| | AIC ↓ | AUC ↑ | AIC↓ | AUC ↑ | AIC ↓ | AUC ↑ |
| SENTLI | 4.968 | 0.715 | 6.966 | 0.719 | 8.940 | 0.738 |
| INFUSE | 5.285 | 0.684 | 7.280 | 0.689 | 9.229 | 0.724 |
| SUMMAC | 5.049 | 0.667 | 7.043 | 0.676 | 9.016 | 0.702 |
| SUMMAC-Conv | 5.272 | 0.662 | 7.259 | 0.677 | 9.229 | 0.703 |
| PT-NLI-Long (FT) | 5.254 | 0.655 | 7.253 | 0.657 | 9.177 | 0.719 |
| PT-NLI-Long | 5.354 | 0.608 | 7.347 | 0.619 | 9.300 | 0.667 |

**Model-specific observations.** SENTLI and INFUSE obtain the highest *lex* AUC ($0.71$ and $0.68$) yet gain almost nothing from BERTScore, corroborating their strong reliance on surface overlap. SUMMAC-Conv is the *most* sensitive to BERTScore ($+1.5$ pp AUC), though the AIC rise still deems the feature non-cost-effective. PT-NLI-Long variants benefit most from the gold label ($+6$ pp AUC), suggesting that, while occasionally correct, they often *guess right for the wrong reasons*.

## 7. Discussion and Conclusion

Despite the recent large-scale resources with reference summaries, there was no summary faithfulness detection dataset for Portuguese. The use of the LLM-as-a-judge framework allowed us to create a benchmark with scale and at a relatively low price. This new resource enabled the study of the limits of current NLI-based faithfulness detection for summarization in Brazilian Portuguese.

Our study found that the performance of zero-shot sentence-decomposition methods was in the moderate-low range, while in their English language counterpart, they range in the moderate-high performance levels. However, the backbone of our implementation of these sentence NLI methods was a BERTimbau model fine-tuned on the InferBR dataset (a dataset with fewer than 10k instances). The English language, on the other hand, can count on datasets such as MNLI [Williams et al. 2018] that feature different textual domains and have almost 0.5 million instances.

For better evaluation of faithfulness detection, there is a pressing need for NLI models that can handle longer, complex sentences and the three labels (entailment, contradiction, and neutral). We also found that all models use lexical overlap as a shortcut to make predictions. The use of this heuristic by NLI models has been well documented in previous literature [McCoy et al. 2019]. Remedies for this problem include adversarial data collection [Nie et al. 2020] and shortcut mitigation strategies [Korakakis and Vlachos 2023].

In summary, our analyses revealed that NLI models often rely on superficial lexical overlap heuristics rather than deeper semantic understanding, limiting their robustness. These findings underscore the need for better resources for training NLI models in Portuguese and more faithful evaluation techniques that go beyond lexical cues.

**Acknowledgments.** This work has been partially funded by CNPq-Brazil and Capes Finance Code 001.

## References


[Bencke et al. 2024] Bencke, L., Pereira, F. V., Santos, M. K., and Moreira, V. (2024). InferBR: A natural language inference dataset in Portuguese. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9050–9060, Torino, Italia. ELRA and ICCL.

[Bowman et al. 2015] Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

[Bueno et al. 2024] Bueno, M., de Oliveira, E. S., Nogueira, R., Lotufo, R., and Pereira, J. (2024). Quati: A brazilian portuguese information retrieval dataset from native speakers. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 236–246, Porto Alegre, RS, Brasil. SBC.

[Cao et al. 2018] Cao, Z., Wei, F., Li, W., and Li, S. (2018). Faithful to the original: Fact aware neural abstractive summarization. *Proc. Conf. AAAI Artif. Intell.*, 32(1).

[Cardoso et al. 2011] Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., Nunes, M. d. G. V., and Pardo, T. A. (2011). Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105. sn.

[Chen et al. 2021] Chen, J., Choi, E., and Durrett, G. (2021). Can NLI models verify QA systems' predictions? In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[Durmus et al. 2020] Durmus, E., He, H., and Diab, M. (2020). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

[El-Kassas et al. 2021] El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

[Faggioli et al. 2023] Faggioli, G., Dietz, L., Clarke, C. L. A., Demartini, G., Hagen, M., Hauff, C., Kando, N., Kanoulas, E., Potthast, M., Stein, B., and Wachsmuth, H. (2023). Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, page 39–50, New York, NY, USA. Association for Computing Machinery.

[Feijo and Moreira 2019] Feijo, D. and Moreira, V. (2019). Summarizing legal rulings: Comparative experiments. In *proceedings of the international conference on recent advances in natural language processing (RANLP 2019)*, pages 313–322.

[Feijo and Moreira 2023] Feijo, D. d. V. and Moreira, V. P. (2023). Improving abstractive summarization of legal rulings through textual entailment. *Artificial intelligence and law*, 31(1):91–113.

[Fonseca et al. 2016] Fonseca, E. R., Borges dos Santos, L., Criscuolo, M., and Aluísio, S. M. (2016). Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática*, 8(2):3–13.

[Gekhman et al. 2023] Gekhman, Z., Herzig, J., Aharoni, R., Elkind, C., and Szpektor, I. (2023). TrueTeacher: Learning factual consistency evaluation with large language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

[Goodrich et al. 2019] Goodrich, B., Rao, V., Liu, P. J., and Saleh, M. (2019). Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 166–175, New York, NY, USA. Association for Computing Machinery.

[Gu et al. 2025] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., and Guo, J. (2025). A survey on llm-as-a-judge.

[Hasan et al. 2021] Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R. (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.

[Holtzman et al. 2020] Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[Ji et al. 2022] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., and Fung, P. (2022). Survey of hallucination in natural language generation. *ACM Comput. Surv.*

[Korakakis and Vlachos 2023] Korakakis, M. and Vlachos, A. (2023). Improving the robustness of NLI models with minimax training. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14339, Toronto, Canada. Association for Computational Linguistics.

[Laban et al. 2022] Laban, P., Schnabel, T., Bennett, P. N., and Hearst, M. A. (2022). Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

[Lin 2004] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[Liu et al. 2021] Liu, H., Cui, L., Liu, J., and Zhang, Y. (2021). Natural language inference in context - investigating contextual reasoning over long texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13388–13396.

[McCoy et al. 2019] McCoy, R. T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

[Mishra et al. 2021] Mishra, A., Patel, D., Vijayakumar, A., Li, X. L., Kapanipathi, P., and Talamadupula, K. (2021). Looking beyond sentence-level natural language inference for question answering and text summarization. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1322–1336, Online. Association for Computational Linguistics.

[Nallapati et al. 2016] Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., and Xiang, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, page 280. Association for Computational Linguistics.

[Nan et al. 2021] Nan, F., Nallapati, R., Wang, Z., Nogueira dos Santos, C., Zhu, H., Zhang, D., McKeown, K., and Xiang, B. (2021). Entity-level factual consistency of abstractive text summarization. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

[Nie et al. 2020] Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

[Pagnoni et al. 2021] Pagnoni, A., Balachandran, V., and Tsvetkov, Y. (2021). Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

[Paiola et al. 2024] Paiola, P. H., Garcia, G. L., Jodas, D. S., Correia, J. V. M., Sugi, L. A., and Papa, J. P. (2024). RecognaSumm: A novel Brazilian summarization dataset. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Pro-*

*cessing of Portuguese - Vol. 1*, pages 575–579, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

[Pardo and Rino 2003] Pardo, T. A. S. and Rino, L. H. M. (2003). Temário: Um corpus para sumarização automática de textos. *São Carlos: Universidade de São Carlos, Relatório Técnico*.

[Piau et al. 2024] Piau, M., Lotufo, R., and Nogueira, R. (2024). ptt5-v2: A closer look at continued pretraining of t5 models for the portuguese language.

[Rajpurkar et al. 2018] Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

[Real et al. 2020] Real, L., Fonseca, E., and Oliveira, H. G. (2020). The assin 2 shared task: a quick overview. In *International Conference on Computational Processing of the Portuguese Language*, pages 406–412. Springer.

[Schuster et al. 2022] Schuster, T., Chen, S., Buthpitiya, S., Fabrikant, A., and Metzler, D. (2022). Stretching sentence-pair NLI models to reason over long documents and clusters. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[Scialom et al. 2021] Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., Staiano, J., Wang, A., and Gallinari, P. (2021). QuestEval: Summarization asks for fact-based evaluation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *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.

[Sharma and Sharma 2022] Sharma, G. and Sharma, D. (2022). Automatic text summarization methods: A comprehensive review. *SN Computer Science*, 4(1).

[Shastry et al. 2025] Shastry, R., Chiril, P., Charney, J., and Uminsky, D. (2025). Entailment progressions: A robust approach to evaluating reasoning within larger discourse. In Johansson, R. and Stymne, S., editors, *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 651–660, Tallinn, Estonia. University of Tartu Library.

[Shen et al. 2023] Shen, C., Cheng, L., Nguyen, X.-P., You, Y., and Bing, L. (2023). Large language models are not yet human-level evaluators for abstractive summarization. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

[Sottana et al. 2023] Sottana, A., Liang, B., Zou, K., and Yuan, Z. (2023). Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

[Souza et al. 2020] Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*, pages 403–417. Springer.

[Tam et al. 2023] Tam, D., Mascarenhas, A., Zhang, S., Kwan, S., Bansal, M., and Raffel, C. (2023). Evaluating the factual consistency of large language models through news summarization. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

[Thakur et al. 2025] Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., and Hupkes, D. (2025). Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges.

[Thomas et al. 2024] Thomas, P., Spielman, S., Craswell, N., and Mitra, B. (2024). Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1930–1940, New York, NY, USA. Association for Computing Machinery.

[Utama et al. 2022] Utama, P., Bambrick, J., Moosavi, N., and Gurevych, I. (2022). Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

[Wang et al. 2020] Wang, A., Cho, K., and Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

[Williams et al. 2018] Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

[Yuan et al. 2021] Yuan, W., Neubig, G., and Liu, P. (2021). Bartscore: evaluating generated text as text generation. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

[Zhang et al. 2024] Zhang, H., Xu, Y., and Perez-Beltrachini, L. (2024). Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

[Zheng et al. 2023] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

[Zhong et al. 2022] Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., and Han, J. (2022). Towards a unified multi-dimensional evaluator for text generation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.