NegNLI-BR: A Brazilian Portuguese Benchmark for Negation in Natural Language Inference

Matheus Westhelle, Viviane Moreira
Institute of Informatics
Universidade Federal do Rio Grande do Sul
Porto Alegre, Rio Grande do Sul, Brazil
{matheus.westhelle, viviane}@inf.ufrgs.br

Abstract

Recent studies have questioned the ability of Large Language Models (LLMs) to handle logical negation. We revisit this issue within the Natural Language Inference (NLI) task, specifically investigating whether modern LLMs can distinguish negations that alter logical entailment ("important") from those that do not ("unimportant"). For this purpose, we introduce NegNLI-BR, a new benchmark dataset in Portuguese designed to exercise this distinction. We evaluate a range of recent open-source LLMs, comparing the performance of their base and post-trained versions. Furthermore, we employ a causal probe to measure the Average Treatment Effect of negation interventions on the internal representations of LLMs. Our findings show that many recent LLMs, including smaller variants, perform well on explicit negation in this controlled Portuguese NLI benchmark. The causal analysis reveals that important negations induce a stable and significant effect on model representations, distinct from unimportant negations or neutral filler words. We also observe that post-training generally enhances this representational sensitivity, suggesting it refines the models' ability to encode the logical impact of negation.

Keywords: Large Language Models, Natural Language Inference

1. Introduction

In the recent past, large neural language models have dominated the Natural Language Processing (NLP) landscape, having shown remarkable success in many downstream tasks (Brown et al., 2020). Until recently, however, language models have had limited logical reasoning, an example of which is negation. Horn (2001) analyzes the study of negation throughout history and details its role in semantics and pragmatics. The first application of automating the processing of negation has reportedly originated in the medical domain, where it is crucial for the proper processing of clinical reports and discharge summaries (Morante and Daelemans, 2012; Vincze et al., 2008). Other notable applications include sentiment analysis (Moore and Barnes, 2021) and information retrieval (Weller et al., 2024).

Truong et al. (2023) found that larger models were associated with a reduced sensitivity to negation, and the performance in Natural Language Inference (NLI) datasets was below random chance. However, their study was conducted with early Large Language Models (LLMs) such as GPT-3 and InstructGPT, and the question remains as to whether more recent LLMs still face limitations in dealing with negation. We investigate negation in the context of Natural Language Inference (NLI), as that task naturally represents logical reasoning, where reasoning under negation is crucial. More specifically, we aim to determine whether more recent LLMs can distinguish between negations that affect the label of a task instance and those that do not. In particular, we are interested in examining this phenomenon in Portuguese, which, to the best of our knowledge, has not received attention in this regard.

Hossain et al. (2020) made a distinction between the negations that are important and unimportant in the context of the NLI task: a negation is **important** if, when dropped, it would change the label of the premise-hypothesis sentence pair for an NLI task instance, and **unimportant** otherwise. This categorization of important and unimportant negations is logical and is something an intelligent system should be able to capture, so it is naturally something we wish our dataset to contain.

In order to conduct our investigation, we constructed a dataset that stresses the property of negation, both important and unimportant, and evaluated the performance of a range of open-source models on it. We also investigated the effect of post-training for the purpose of negation understanding. Post-training refers to an umbrella of different techniques used to align models to user preferences, as well as to hone their capabilities (Lai et al., 2025). Instruction Fine-Tuning boosts a model's ability to follow instructions (Chung et al., 2024). Reinforcement learning techniques have been used to align a model with user preferences (Ouyang et al., 2022), but with the advent of reasoning models such as OpenAI o1 (Jaech et al., 2024) and Deepseek R1 (Guo, 2025), attention has been drawn to reinforcement learning as a way to imbue models with reason-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1226–1235
11–16 May 2026.

ing abilities through what has become known as test-time compute (Snell et al., 2024).

Another goal of this work is to investigate the causal effect of negations on the model's representations. If a model is capable of encoding a logical negation in its representations, this suggests that it has internalized the capability to reason in that scenario.

This work aims to answer the following research questions:

RQ1. How does post-training affect an LLM's ability to perform the task proposed in the dataset?

RQ2. Are important negations encoded in the representations of LLMs?

To answer Item RQ1., we evaluated whether post-training improves the ability of an LLM to handle negation by comparing base and post-trained models on NegNLI-BR, a dataset focused on negation that we constructed for this investigation. For answering Item RQ2., we designed a causal probe to verify whether LLMs encode negations that are important for an NLI task by establishing negation as an intervention that is carried out on a premise in a premise-hypothesis pair. We measured the magnitude of the effect of a negation on the representations of LLMs and compared it against a baseline effect of inserting a filler word. We found that the Average Treatment Effect (ROSENBAUM and RUBIN, 1983) of important negation is stable for most models we tested, which suggests that they capture this property in their representations. We also observed that, on average, the representations of post-trained LLMs are more sensitive to important negations than for their pre-trained counterparts.

The contributions of this work include (i) a dataset for negation in NLI, (ii) a study on how post-training impacts the treatment of negation, (iii) an investigation on how recent LLMs represent negations, and (iv) a case study in Portuguese, a language that, despite being widely spoken, is underrepresented in terms of linguistic resources.

2. Background

The Natural Language Inference (NLI) task, which we explore in this work, consists of establishing an entailment relationship between two fragments of text, a premise P and a hypothesis H. There are three possible scenarios in this task:

1. **Entailment:** $P \models H$, or H can be logically inferred from P.
2. **Contradiction:** $P \models \neg H$, or H contradicts P.
3. **Neutral:** $P \not\models H \wedge P \not\models \neg H$, no relation can be inferred between P and H.

An example of each case can be seen in Table 1.

In the literature, NLI is also commonly referred to as Recognizing Textual Entailment (Dagan et al., 2005). In our work, we consider a simpler, binary version of NLI where we are only interested in determining whether a premise entails a hypothesis or not. Real et al. (2020) use that formulation, for example. This simplifying assumption allows us to handle negation uniformly because, when applying negation to a premise, the shift of its relation to the hypothesis would depend on the previous relation present in the premise-hypothesis pair. Consider the following premise-hypothesis pairs in Table 2. In this case, instead of *Contradiction*, we have *Non-entailment*, which suffices for our goals.

Model probing, in general, is the practice of investigating whether a model encodes a property of interest. A notable example is the work of Rogers et al. (2020) in probing the BERT model to understand how its representations map to linguistic phenomena. A causal probe, as defined by Amini et al. (2023), attempts to discover the *causal* relationship between a property and the internal representations of a model. *Causality*, for our purposes, is expressed in the framework defined by Pearl (2009), which is a probabilistic graphical model expressed by causal diagrams. Within that framework, probing is interested in interventions, where a variable of interest is set to a fixed value without altering any of the parents of the variable in the diagram. Applying an intervention creates what is called a **counterfactual** or, in other words, *What is the effect on variable Y if we change a variable X, but keep everything else the same?*. Interventions can happen directly on the representations of a model (Ravfogel et al., 2021), or at the level of inputs (Vig et al., 2020; Amini et al., 2023), which is what we explore in this work.

3. Related Work

Despite the rapid progress of deep neural models for language modeling, simple distributional semantics have been shown to be insufficient for capturing the meaning of negation. This was found to be true by Kassner and Schütze (2020) for Pre-trained Language Models (PLMs) such as BERT (Devlin et al., 2019), which uses an encoder-only transformer architecture, and ELMo, which is based on a bidirectional LSTM (Peters et al., 2018). This has also been demonstrated by citet truong-2023-naysayers for large decoder-only transformer models (or LLMs, as they have become notoriously known).

Kassner and Schütze (2020) demonstrated the failure of PLMs in grasping the concept of negation by formulating a *cloze* task in which the negation

Premise	Hypothesis	Label
Uma criança alegre está brincando no parque. *(A cheerful child is playing in the park.)*	Uma criança está no parque. *(A child is in the park.)*	Entailment
Um indivíduo está tocando piano. *(An individual is playing the piano.)*	Ninguém está tocando piano. *(No one is playing the piano.)*	Contradiction
O gato está dormindo. *(The cat is sleeping.)*	O gato tem pelo preto. *(The cat has black fur.)*	Neutral

Table 1: Example of instances in an NLI task

Premise	Hypothesis	Relation
O homem está dançando. *(The man is dancing.)*	O homem está fazendo uma dança. *(The man is doing a dance.)*	Entailment
O homem não está dançando. *(The man is not dancing.)*	O homem está fazendo uma dança. *(The man is doing a dance.)*	Non-entailment

Table 2: Example of NLI instances with binary (*Entailment* and *Non-entailment*) labels.

cue *not* is inserted in factual sentences, thus creating positive/negative sentence pairs. They found that the predicted filler words have a high overlap. An example of the kinds of sentence pairs seen in their experiment is:

Positive: *The theory of relativity was developed by [MASK].*

Negative: *The theory of relativity was not developed by [MASK].*

Truong et al. (2023) investigated the capabilities of LLMs with regard to negation and also found them lacking. In addition to *cloze* tasks, they also assessed lexical semantics of negation through antonymy classification and the ability to reason with negation in the NLI task. They found that LLMs faced difficulties in all three scenarios. Furthermore, they found that larger models were more insensitive to negation, a finding consistent with that of Zhang et al. (2023).

Vrabcová et al. (2025) investigated NLI in the context of negation. They experimented with several English datasets that they also translated into Czech, German, and Ukrainian. They tested post-trained variants of LLama 3 (Grattafiori et al., 2024), Qwen 2.5 (Yang et al., 2025a), and Mistral (Mistral AI, 2024). Their goal was to understand whether these models are robust to negation, where robustness is defined as a measure of performance degradation in the presence of negation. Previously, Hossain et al. (2020) also studied negation in NLI, proposing NegNLI, a benchmark for reasoning under negation, in which they also found that LLMs struggle. Our work differs in that we aim to (i) understand how important post-training is to that task, for which we use Portuguese, a language from the Romance family, which was not represented in that work, and (ii) find whether the internal representations of LLMs encode the ability to distinguish between negations that are necessary to perform logical entailment.

4. NegNLI-BR – Dataset Construction

We constructed our dataset using instances from Portuguese NLI datasets, namely ASSIN 2 (Real et al., 2020) and InferBR (Bencke et al., 2024). Examples of each dataset can be seen in Table 3.

In order to answer our research questions, we need a dataset that fulfills a set of desiderata, namely:

1. It should be easy to apply an intervention on any instance in the form of a simple negation.
2. It should be simple to verify the effect of a negation intervention on the label of an instance.
3. The dataset should be robust to any shortcuts that a model may take to correctly classify an instance.

Since Infer-BR used the three-label variant of NLI, we processed it to convert all instances that are labeled as either *Contradiction* or *Neutral* into *Non-entailment*. The dataset construction process is summarized in Figure 1.

We proceeded to filter the datasets for instances where the premise is a simple *Noun Phrase + Verb Phrase* sentence, with the help of a constituent parser (Qi et al., 2020). We removed instances that already have a simple negation in the premise, and then select the remaining instances that have

	Instances	Labels	Example
ASSIN 2	10,000	Entailment, None	**Premise:** Uma criança risonha está segurando uma pistola de água e sendo espirrada com água. (*A cheerful child is holding a water gun and being sprayed with water.*) **Hypothesis:** Uma criança está segurando uma pistola de água. (*A child is holding a water gun.*) **Entailment judgment:** Entailment.
InferBR	8,767	Contradiction, Entailment, Neutral	**Premise:** O homem presenteia sua esposa com um colar de pérolas. (*The man gives his wife a pearl necklace.* **Hypothesis:** O homem presenteia sua esposa com um colar de pérolas no aniversário dela. *The man gives his wife a pearl necklace on her birthday.* **Label:** Neutral.

Table 3: A comparison of the ASSIN 2 and InferBR datasets for Natural Language Inference. Both datasets have an even class balance, which is why we omit label distribution. Portuguese examples are accompanied by English translations for clarity.

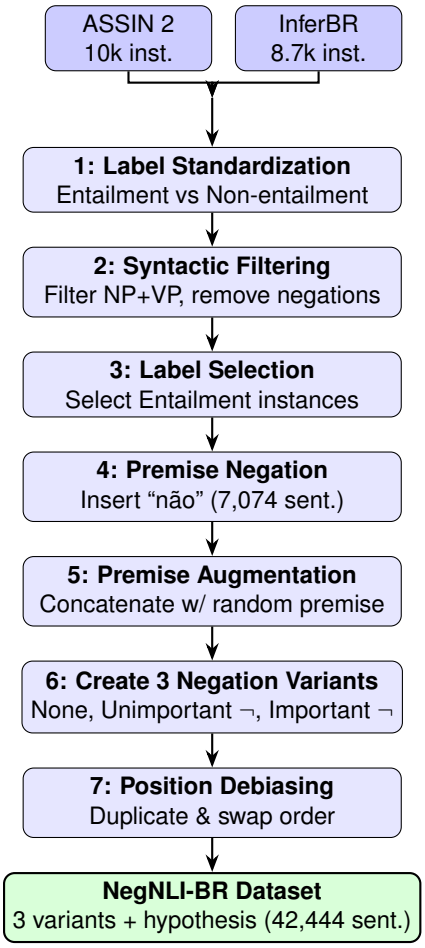


Figure 1: Dataset construction pipeline from ASSIN 2 and InferBR.

the *Entailment* label such that when the premise is negated, that label should flip to *Non-entailment*. To negate the premises in our filtered dataset, we used a constituent parser, inserting the Portuguese negation adverb "não" modifying the verb in the sentence, which works because we have filtered them so that they are always `NP + VP`). At this point, the dataset had $7{,}074$ instances. Then, we would like to fulfill item 3 of our desiderata, which is to make the instances robust to shortcuts a model could take to solve the task, such as automatically marking an instance as *Non-entailment* due to the existence of a negation cue.

We also needed to account for when a negation is important to the outcome of the classification of a premise-hypothesis pair. For that purpose, we augmented each of our instances by randomly sampling other unrelated premises from our same dataset, connecting both premises with an additive connector, which looks like the following:

> **Premise 1:** Batatas estão sendo fatiadas por um homem.
> *Potatoes are being sliced by a man.*
>
> **Premise 2:** Um caminhão está descendo rapidamente um morro.
> *A truck is quickly going down a hill.*
>
> **New premise:** Batatas estão sendo fatiadas por um homem, e um caminhão está descendo rapidamente um morro.
> *Potatoes are being sliced by a man, and a truck is quickly going down a hill.*

The resulting augmentation is intentionally synthetic and may yield sentences with low discourse coherence, such as the one in the example. Therefore, the evidence we obtain reflects primarily internal validity within a controlled setting.

At the same time, this construction enables us to create minimal interventions by adding *important* and *unimportant* negations to each of our instances: a negation is important when added to the original premise and unimportant when added to the sampled premise.

Category	Number of sentences
Without negation	14,148
With important negation	14,148
With unimportant negation	14,148
Total	**42,444**

Table 4: Distribution of sentences across negation categories in the dataset.

We would also like to ensure that the dataset is robust against position bias, so we included each instance twice, with the order of premises swapped. This doubles the number of instances to 14,148. An example row of our dataset can be seen in Table 5, and the sentence counts across negation categories can be found in Table 4. The NegNLI-BR dataset is available at https://huggingface.co/datasets/hapaxlegomenon/NegNLI-BR. Likewise, our research code is published on Github[1].

5. Materials and Methods

We conducted our experiments on a collection of open-source models, namely Qwen3 (Yang et al., 2025b), OLMo 2 (OLMo et al., 2024), and Gemma 3 (Kamath et al., 2025). We also experimented with Tucano (Corrêa et al., 2024), an LLM trained exclusively in Portuguese. For each model, the tests were conducted with both its pre-trained and post-trained versions to understand the role of instruction fine-tuning in the performance of a model.

In order to verify whether the models can infer if there is entailment between a premise and a hypothesis, we formulate the task as a two-shot prompt, shown in Figure 2. The translation into English, shown in blue, is not part of the prompt; it is included only for clarity.

5.1. RQ1 - Model performance

To answer **RQ1** *How does post-training affect an LLM's ability to perform the task proposed in the dataset?*, we measured the prediction accuracy of each model in our dataset for each version of our instances: without negation, with an important negation, and with an unimportant negation. In order to obtain the predictions, we use greedy decoding to obtain the most likely token after the prompt. We report the harmonic mean of the accuracy values to heavily penalize models that underperform on any version of the task.

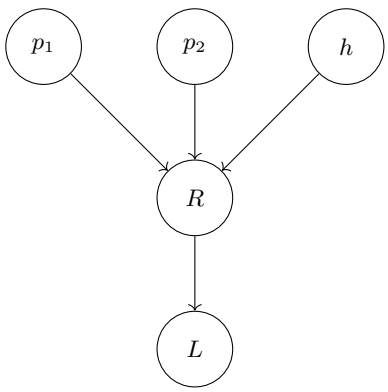
Our results in Table 6 show that about half (8 out of 13) of the tested LLMs are capable of understanding the role of an important negation in the NLI task in their pre-trained variants, showing results above a random baseline for the harmonic mean of all variants of the task. The base versions of Olmo 2.1 B and Tucano 2.4 B show very poor performance on important negations, indicating that these models struggle with negation. Olmo 2 13B appears to show excellent performance on the versions without negation and with a unimportant negation, but fails on important negations, which essentially means that it guesses "yes" for all or most instances most of the time, a behavior that the harmonic mean duly penalizes.

Post-training shows a considerable effect on important negation for the smaller variants of the tested LLMs. Interestingly, in the case of Qwen3 1.7B, the performance is considerably worsened (-37.2%), but the other small models benefit from it: $+37.2\%$ for Olmo 2, $+14.3\%$ for Tucano 2B4 (from 0.002%, a very large relative increase), and most impressively $+80.7\%$ from 0% in the case of Gemma 2. In contrast, post-training has a relatively small effect on the larger model variants. This is a more nuanced perspective on the findings of Truong et al. (2023), who found that instruction fine-tuning strictly improves reasoning under negation for classification tasks.

5.2. RQ2 - Causal probe

We framed important and unimportant negations as interventions that are applied on a premise. To have a baseline of comparison, we created versions of our premise instances where the important and unimportant negations are respectively replaced by a filler word that has no impact on the label of the premise-hypothesis pair. We obtained the representations from the last hidden layer of the model, at the last token of each instance, before the prediction. From the model representations of each version of our premise-hypothesis pairs, we compute Individual Treatment Effects (ITEs), or the difference between the representations of a premise-hypothesis pair and the representations of that pair when an intervention is applied.

The following is a causal diagram (Pearl, 2009) that describes our probe:



[1] https://github.com/mwesthelle/negnli-br

Premise without negation	Premise with unimportant negation	Premise with important negation	Hypothesis
Uma mulher está pintando um quadro em seu estúdio, e uma colcha rosa com estampa floral cobre a cama de casal. *(A woman is painting a picture in her studio, and a pink floral bedspread covers the double bed.)*	Uma mulher não está pintando um quadro em seu estúdio, e uma colcha rosa com estampa floral cobre a cama de casal. *(A woman is not painting a picture in her studio, and a pink floral bedspread covers the double bed.)*	Uma mulher está pintando um quadro em seu estúdio, e uma colcha rosa com estampa floral não cobre a cama de casal. *(A woman is painting a picture in her studio, and a pink floral bedspread does not cover the double bed.)*	A cama de casal está revestida com uma colcha de cor rosa e com flores estampadas. *(The double bed is covered with a pink bedspread with floral patterns.)*

Table 5: NegNLI-BR example row with English translations.

Prompt

```
Premissa: Sócrates é homem e todo homem é mortal. (Premise: Socrates is a man and every man is mortal.)
Hipótese: Sócrates é mortal. (Hypothesis: Socrates is mortal.)
A premissa implica a hipótese (Sim/Não)? Sim (Does the premise entail the hypothesis (Yes/No)? Yes)

Premissa: Pedro tem dois irmãos. (Premise: Peter has two siblings.)
Hipótese: Pedro é alto. (Hypothesis: Peter is tall.)
A premissa implica a hipótese (Sim/Não)? Não (Does the premise entail the hypothesis (Yes/No)? No)

Premissa: $premise
Hipótese: $hypothesis
A premissa implica a hipótese (Sim/Não)? (Does the premise entail the hypothesis (Yes/No)?)
```

Figure 2: Two-shot prompt used in NegNLI-BR; blue glosses are explanatory only.

where p_1 and p_2 are the coordinate clauses in the premise, h is the hypothesis, R is the internal representation of the last token of the task instance before the label, and L is the label itself.

Our interventions were applied on p_1 or p_2, depending on the position of the negation, as described in Section 4. There are four kinds of interventions: *important negation*, *unimportant negation*, *important filler* (when a filler word takes the place of an important negation), and *unimportant filler* (when a filler word takes the place of an unimportant negation). The former two are necessary for us to obtain baselines of comparison, namely the difference in the representations of a model when we insert a filler word or when we insert a negation cue.

Our choice of filler word is non-trivial, as we need to account for the effect of changing the length of a sequence of tokens. As such, the chosen filler word must be the same length in tokens as the negation cue "não" for all tokenizers of the chosen models. For this purpose, the Portuguese word "pois" fulfills those requirements.

Formally, the Average Treatment Effect (ATE) is defined as the expected difference between two treatments (Rosenbaum and Rubin, 1983), as shown in Eq. 1

$$\text{ATE} = \mathbb{E}[R(f_{neg}(X))] - \mathbb{E}[R(f_{filler}(X))] \quad (1)$$

where R is a variable that stands for the representations of the model, X is our dataset, f_{neg} is an intervention that adds a negation cue to the instances in the dataset, and f_{filler} is a baseline intervention that adds the filler word to the instances in the dataset.

As in Amini et al. (2023), our experimental design naturally allows us to compute an approximation of the ATE using a paired estimator, where a pair is given by the same instance affected by our two interventions (2). We compute 95% confidence intervals for the ATE values by bootstrapping.

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N} R(f_{neg}(X)) - R(f_{filler}(X)) \quad (2)$$

Model	No negation		Important negation		Unimportant negation		Harmonic mean	
	Base	Post-trained	Base	Post-trained	Base	Post-trained	Base	Post-trained
Qwen								
Qwen 3 1.7B	0.910	0.944	0.880	0.508	0.880	0.945	0.890	0.734
Qwen 3 8B	0.943	0.956	0.962	0.926	0.899	0.946	0.934	0.943
Qwen 3 14B	0.979	0.979	0.930	0.823	0.980	0.980	0.962	0.921
Qwen 3 30B A3B	0.977	0.989	0.914	0.862	0.978	0.987	0.955	0.942
OLMo								
OLMo 2 0425 1B	0.930	0.629	0.031	0.403	0.964	0.607	0.087	0.524
OLMo 2 1124 7B	0.047	0.587	**0.999**	0.892	0.032	0.326	0.056	0.509
OLMo 2 1124 13B	**1.000**	**0.999**	0.002	0.002	**1.000**	**1.000**	0.006	0.006
OLMo 2 0325 32B	0.971	0.927	0.484	0.783	0.978	0.937	0.729	0.876
Tucano								
Tucano 2B4	0.998	0.931	0.002	0.145	0.999	0.884	0.006	0.329
Gemma								
Gemma 3 1B	0.983	0.625	0.000	0.807	0.990	0.452	0.000	0.594
Gemma 3 4B	0.761	0.938	0.950	0.752	0.596	0.965	0.742	0.874
Gemma 3 12B	0.987	0.960	0.609	0.934	0.997	0.972	0.820	0.955
Gemma 3 27B	0.984	0.971	0.935	**0.968**	0.983	0.975	**0.967**	**0.971**

Table 6: Accuracy of base and post-trained models on NegNLI-BR.

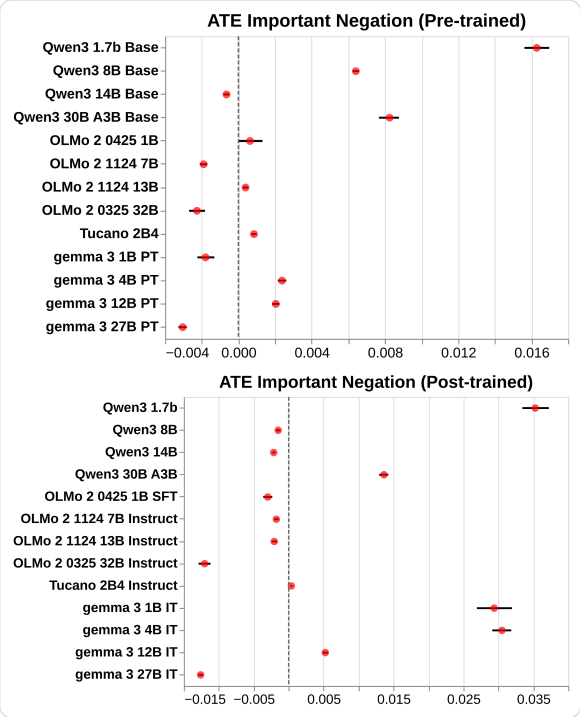


Figure 3: ATE for important negation in base (top) and post-trained (bottom) models.

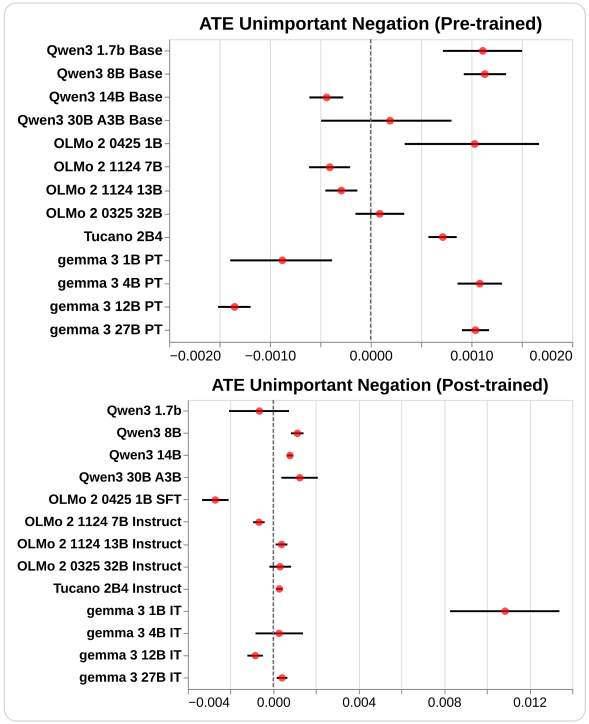


Figure 4: ATE for unimportant negation in base (top) and post-trained (bottom) models.

The pre-trained Qwen 3 models showed less sensitivity to important negation the larger the model (Figure 3), considering that the 30B A3B variant is a Mixture-of-Experts model that has only 3B parameters active during inference (Yang et al., 2025b). Most models show a marked increase in sensitivity to important negation after post-training. An interesting exception is OLMo 2 7B, for which the effect does not change much. The pre-trained version of this model has an issue where it guessed "no" most of the time, showing very low performance in the *No negation* version of the task. This has a marked improvement of 54% with post-training, although performance actually worsens by 10.7% on the *Important negation* task. We also note that our estimates were, for the most part, very stable, which indicates that the models interpret important negation in a very specific and almost

deterministic way.

Unimportant negations, on the other hand, presented unstable estimates on the pre-trained models Figure 4, and much lower values than important negations, which points to the fact that the negation cue in an irrelevant position is interpreted as noise by the LLMs. After post-training, LLMs become more sensitive to important negations, indicating a more general behavior of post-trained models. It is essential to note that, although there is an increased sensitivity, the values are either significantly lower than in the case of important negation or much noisier, as is the case for Gemma 3B IT.

6. Conclusion

This work focused on determining whether LLMs can distinguish between negations that affect the label of an instance and those that do not. The first step was to create NegNLI-BR, a dataset to evaluate the capabilities of models on identifying simple negations for the NLI task. NegNLI-BR is in Portuguese and was derived from two existing NLI datasets (ASSIN 2 and InferBR). We then set up a two-shot prompt to steer models toward providing appropriate completions for the NLI task, given the instances in our dataset. Our results were evaluated in two dimensions: First, we verified the ability of the chosen models to perform the proposed task by measuring the accuracy they obtained on three categories (“No negation”, “Important negation”, and “Unimportant negation”). Second, we probed the representations of the models to check whether negation has a significant impact on them.

Our findings suggest that LLMs can handle explicit clause-level negation in Portuguese NLI. Many of them are capable of extracting the underlying logical meaning of a negation cue in relation to the context in which they exist. Furthermore, LLMs are capable of precisely encoding whether a negation cue is important for an NLI task and when it is not, and this effect is more noticeable for smaller models. In addition, this capability arises even on models that are very small for the current standards, suggesting that the refinement of training techniques and amount of training data play a much more important role in negation understanding than model size. For future work, we intend to investigate exactly what that role is, or in other words, what is the training regimen required for the emergence of negation understanding.

Ethics

Our dataset, NegNLI-BR, carries over the same ethical considerations of InferBR (Bencke et al., 2024), which contains instances automatically generated with GPT-4 (OpenAI et al., 2024), potentially reflecting societal biases in its closed training data. The other dataset it is derived from, ASSIN 2, comes from image captions that, to the best of our knowledge, do not carry any significant ethical considerations.

Limitations

Our work is limited to a controlled binary version of NLI in the Portuguese language. We also limit our scope to simple negation with não ("*not*") and do not explore other negation cues, or even more complex forms of negation, such as morphological negation through the use of affixes, lexical negation, and antonymy. The synthetic nature of our dataset, where we concatenate the original premise with a randomly sampled unrelated premise, isolates whether negation impacts entailment, but can also produce sentences that are not reflective of naturalistic discourse.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and CNPq.

7. Bibliographical References

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. Naturalistic Causal Probing for Morpho-Syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403.

Luciana Bencke, Francielle Vasconcellos Pereira, Moniele Kunrath Santos, and Viviane Moreira. 2024. InferBR: A natural language inference dataset in Portuguese. In *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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language

models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 1877–1901. Curran Associates Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

Nicholas Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2024. Tucano: Advancing neural text generation for portuguese.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*.

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

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models.

Daya others Guo. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Laurence R. Horn. 2001. *A Natural History of Negation*. The David Hume Series. CSLI.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. 2025. Gemma 3 technical report. *ArXiv*, abs/2503.19786.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, Xin Lv, Minlie Huang, Yuxiao Dong, and Jie Tang. 2025. A survey of post-training scaling in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2771–2791, Vienna, Austria. Association for Computational Linguistics.

Mistral AI. 2024. Mistral AI Models on Hugging Face. Models: 12B Nemo, 22B Small, 123B Large.

Andrew Moore and Jeremy Barnes. 2021. Multi-task learning of negation and speculation for targeted sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2838–2869, Online. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2012. Conandoyle-neg: Annotation of negation cues and their scope in conan doyle stories. In *International Conference on Language Resources and Evaluation*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious.

OpenAI, Josh Achiam, Steven Adler, et al. 2024. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv e-prints*, page arXiv:2203.02155.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The assin 2 shared task: A quick overview. In *Computational Processing of the Portuguese Language: 14th International Conference, PROPOR 2020, Evora, Portugal, March 2–4, 2020, Proceedings*, page 406–412, Berlin, Heidelberg. Springer-Verlag.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

PAUL R. ROSENBAUM and DONALD B. RUBIN. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Paul R. Rosenbaum and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *ArXiv*, abs/2408.03314.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(11):S9.

Tereza Vrabcová, Marek Kadlčík, Petr Sojka, Michal Štefánik, and Michal Spiegel. 2025. Negation: A Pink Elephant in the Large Language Models' Room?

Orion Weller, Dawn Lawrie, and Benjamin Van Durme. 2024. NevIR: Negation in neural information retrieval. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2274–2287, St. Julian's, Malta. Association for Computational Linguistics.

An Yang et al. 2025a. Qwen2.5 technical report.

An Yang et al. 2025b. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z. HaoChen, James Zou, Percy Liang, and Serena Yeung. 2023. Beyond positive scaling: How negation impacts scaling trends of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7479–7498, Toronto, Canada. Association for Computational Linguistics.