Getting Logic From LLMs
Annotating Natural Language Inference with Sabiá

Fabiana Avais[1], **Marcos Carreira**[2], **Livy Real**[1]

[1]Federal University of Paraná
Curitiba, PR – Brazil

[2]State University of Ponta Grossa
Ponta Grossa, PR – Brazil

avaisfabiana@gmail.com, marcoscarreira@uepg.br, livyreal@gmail.com

Abstract. *We discuss the difficulties of annotation for Natural Language Inference in Portuguese, comparing human and Large Language Model annotations. We used 200 sentence pairs from the ASSIN2 dataset and re-annotated them for the inference task. A semanticist conducted the first annotation, and a second round was conducted using Sabiá-3, a large language model trained on Brazilian Portuguese data. We found that Sabiá-3 has the same agreement score as human annotators, but the LLM and human annotators disagree in cases involving different linguistic phenomena. While humans tend to disagree on pairs involving pragmatics or cultural knowledge, Sabiá-3 tends to mislabel sentences that share context but with no clear, logical relations among them. It shows that although LLMs are now statistically as effective as humans, LLMs and humans have different patterns for disagreement or mistaken annotations for Natural Language Inference.*

1. Introduction

Since the advent of classical machine learning methods, reliable annotated data has been the bottleneck of Natural Language Processing (NLP). Data augmentation techniques have been fruitful in developing and enhancing models for specific tasks, mainly when dealing with real-world scenarios.

Currently, the first wave of Large Language Models (LLMs) takes place. Considering the effort and time involved, LLMs are highly resource-intensive and expensive. In this scenario, using LLMs to label new data to train a classical model is feasible. However, evaluating the generated data is still an open challenge.

Here, we investigate the possibility of using Sabiá-3, a Brazilian LLM, for the Natural Language Inference (NLI) task. NLI is considered the initial step for semantic reasoning, and although it is elementary for humans, it remains a significant hurdle for machine reasoning.

Natural Language Inference is a task from Natural Language Understanding or Semantic Reasoning. It is, briefly, the task of inferring the validity of a given piece of text from another, i.e., '[s]uch inference can be defined as the process of concluding the truth of a textual statement based on (the truth of) another given piece of text' [Dagan et al. 2013]. It is different from Classical Linguistics or Formal Semantics, in which inferences are seen as a strictly semantic phenomenon. It is also different from

implicatures, in which pragmatics and world knowledge are responsible for the relations derived between two informational pieces. Therefore, NLI is commonly simplified as the task of finding the relations between two sentences or two short texts, no matter which linguistic phenomena are involved in getting those relations [Condoravdi et al. 2003].

The first works on NLI date from the early 2000s with the Pascal Challenges [Dagan et al. 2006], when the task posed was Recognizing Textual Entailment (RTE). With the success of machine learning approaches that require large volumes of annotated data, larger datasets became available in the 2010s, such as SICK [Marelli et al. 2014], SNLI [Bowman et al. 2015] and MultiNLI [Williams et al. 2017]. In Brazilian Portuguese, we also have some work done on NLI. ASSIN shared task [Fonseca et al. 2016] and the SICK-BR dataset [Real et al. 2018] introduced the discussion on Portuguese, and some other works, such as ASSIN 2 [Real et al. 2020] and [Salvatore 2020] continued to discuss and advance the Brazilian Portuguese state-of-the-art. Nowadays, the theme is expected to be revisited with the advent of LLMs. [Bencke et al. 2024] showed that GPT-4 could classify simple entailment relations, highly agreeing with humans.

Thus, we intend to profoundly understand what LLMs can do for NLI in Portuguese. The primordial question we want to answer is how well LLMs, particularly Sabiá-3, can deal with Brazilian Portuguese inference. Since NLI is a subtle task that poses challenges even for humans, we also want to investigate how different the processing of corner cases for humans and LLMs is.

Finally, it is relevant to highlight that several works have pointed out inherent disagreements on NLI interpretation [Kalouli et al. 2017, Pavlick and Kwiatkowski 2019, Zhang and de Marneffe 2021, Kalouli et al. 2023]. It is also relevant to understand if this problem arises when using LLMs for NLI. This study also attempts to deeply understand inherent disagreements in Brazilian Portuguese and contribute positively to achieving noiseless NLI corpora in the future.

2. Related Works

In this section, we focus on works describing difficulties in NLI annotation, and also in NLI resources for Brazilian Portuguese.

[Kalouli et al. 2017] describes a manual investigation of the SICK Corpus [Marelli et al. 2014], reannotating part of the SICK data that were considered logically wrong as sentence pairs that were labeled as entailment in one direction and contradiction in another. This work pointed out the necessity of deeply understanding the linguistic phenomena involved in those cases. It led to [Kalouli et al. 2019], in which inference pairs were labeled by a group of graduate students on semantics that also provided explanations of their reasoning for each label. This work proposes a categorization of linguistics phenomena inherently difficult for humans to annotate, as *direcionality* (a sentence can be inferred by another, but the opposite it is not true) and *loose definition* (a lexical item is polysemous, vague or ambiguous making it hard to ground the sentence meaning).

Finally, [Pavlick and Kwiatkowski 2019] discusses cases in which inherently human annotators would disagree, proposing that NLI systems should predict the distributions over human judgments and not categorical labels. [Zhang and de Marneffe 2021]

trained an Artificial Annotator to detect those inherent disagreements, simulating the uncertainty in the annotation process.

Considering the Brazilian Portuguese context, the ASSIN corpus [Fonseca et al. 2016] was the first dataset for NLI to appear. It was used in the *Avaliação de Similaridade Semântica e Inferência Textual*[1] shared task. ASSIN2 was the second edition of this challenge [Real et al. 2020]. Although both editions tested NLI and Semantic Similarity, we only focus on NLI here. The ASSIN2 dataset comprises 10k pairs of sentences that are human-labeled for inference and semantic similarity. It was thought to be as simple as possible: it has no named entities, and all the sentence pairs are in the present tense. Inferences labels are simply 'entailment' or 'none', leaving aside other relevant labels for inference as 'contradictions'. It makes it a good candidate for our investigation since this is the first attempt at using Sabiá-3 for NLI.

[Bencke et al. 2024] is a recent work describing the InferBR dataset, a NLI corpus created semi-automatically. It parts from previous datasets, as SICK-BR [Real et al. 2018], and it uses classical techniques to redesign the premises of the sentence pairs and GPT-4 to generate new hypotheses and label new pairs. The human evaluation shows that 99.9% of the assigned labels by GPT-4 are useful. Authors state that the 'errors found are related to neutral boundaries with entailment and contradictions' [Bencke et al. 2024, p.9056], which, we believe, are potentially related to inherent disagreements. These impressive results also inspired this work: we want to see if a Brazilian LLM would have similar results and which kind of pairs would be the most challenging.

Commonly, NLI is one of the tasks used to evaluate an LLM [Rodrigues et al. 2023, Chaves Rodrigues et al. 2023], but, to the best of our knowledge, there is no work digging into the applications of LLMs for NLI considering linguistic phenomena in Portuguese.

3. Methodology

We propose to evaluate how good the Sabiá-3 model, from the Sabiá family [Almeida et al. 2024], is for NLI. To do so, we took a sample of sentence pairs of ASSIN2, considered as our baseline. Then, two re-annotations were conducted, one by a semanticist and another by the Sabiá-3 model. Finally, we compare the results and discuss linguistic questions that arise.

Since LLMs are trained on vast amounts of public data and Sabiá-3 was not fine-tuned for the NLI task, our intention is not to describe the 'LLM reasoning' but to check the feasibility of using the model as one NLI annotator and describe the patterns we find in the annotated data.

Thus, our first step was blindly re-annotating the inference labels of 200 sentence pairs randomly selected from the ASSIN2. In the second step, we reannotated the same sample with LLM Sabiá-3 through MariTalk[2] platform. MariTalk is a free-of-charge chatbot that serves Sabiá models. Afterwards, we statistically analyzed the results of the three annotations, categorizing all disagreements by the semantic phenomena prominent in each

[1]Evaluating Semantic Similarity and Textual Entailment.

[2]https://chat.maritaca.ai/, on August 28th, 2024.

sentence pair.

The hypothesis was that by re-annotating part of the corpus ASSIN2 we would find different inference labels in some pairs. Each pair of ASSIN2 was annotated from 3 to 5 people and the final labels were the ones with the majority agreement. One semanticist exclusively did our re-annotation to be grounded on logical relations and to leave aside pragmatical influence whenever possible.

Another investigation concerning the Sabiá-3 outputs took place, in which we compared its reannotation labels to ours and ASSIN2's first annotation. Below there are some examples[3] of ASSIN2 pairs and labels agreed upon by all three processes.

Premise	**Hypothesis**	**Label**
A senhora está mexendo ovos em uma tigela.	A mulher está mexendo ovos em uma tigela.	Entailment
Um homem está tocando violão.	Um homem está tocando o instrumento.	Entailment
Não tem água sendo bebida por um gato.	Um caminhão está descendo rapidamente um morro.	None
Não tem muitas pessoas no parque de patinação no gelo.	Muitas pessoas estão em um parque de patinação no gelo.	None

Table 1. Examples of ASSIN2.

For the Sabiá-3 classification, we created four different prompts and tested them on 10 pairs with different semantic phenomena. We tried out zero-, one- and few-shot prompting [Dang et al. 2022], using ASSIN2 samples as examples. Since the four prompts had the same results for the 10 analyzed pairs, we opted for the simplest one, the zero-shot prompt. The selected prompt was the following:

'Você é um anotador da tarefa de acarretamento entre pares de sentenças. A relação de acarretamento acontece quando a partir de uma sentença [A] podemos concluir que uma outra sentença [B] também é verdadeira. Ou seja, de [A] podemos concluir [B]. Para cada par de sentenças, estabeleça a relação classificando-o como ENTAILMENT quando há acarretamento entre as sentenças e NONE quando não há acarretamento.'[4]

The prompt was also tested on both ChatGPT (GPT-4o mini) and Maritalk (Sabiá-3), classifying one, five, or ten pairs per request. Since we obtained the same results for these ten examples, we opted to go only with Sabiá-3 and to classify ten samples per request.

4. Results

Comparing the three outputs, the total number of sentence pairs with single disagreements, in which only one annotator disagrees with the other two, was 30 pairs out of 200. Within this set of differences, Sabiá-3 had 36.6% of the single disagreements; ASSIN2 had 13.3% of single disagreements; and finally, the semantic expert annotated 50% of the single disagreements.

[3]Translations of the examples: (A) The lady is stirring eggs in a bowl. (B) The woman is stirring eggs in a bowl. (A) A man is playing the guitar. (B) A man is playing the instrument. (A) There is no water being drunk by a cat. (B) A truck is going down a hill quickly. (A) There are not many people at the ice skating park. (B) Many people are at an ice skating park.

[4]'You are an annotator of the entailment task between pairs of sentences. The entailment relation happens when from a sentence [A] we can conclude that another sentence [B] is also true. In other words, from [A] we can conclude [B]. For each pair of sentences, establish the relationship classifying it as ENTAILMENT when there is entailment between the sentences and NONE when there is no entailment.'

Considering inter-agreement annotations (ASSIN2 against LLM, ASSIN2 against semanticist, semanticist against LLM), the sum of mismatches within the groups was 60 disagreements. Sometimes, more than one annotator disagreed with the same pair, so the pair was listed in more than one group. In this scenario, 19 pairs formed the mismatches between the semanticist and the ASSIN2; 15 composed the disagreement between the LLM and the ASSIN2; and 26 pairs formed disagreement between the semanticist and the LLM.

Therefore, there were more annotation discrepancies between the semanticist and the ASSIN2, rather than the LLM and the ASSIN2. Consequently, it shows the LLM tends to agree more with the general labels from ASSIN2 than with the logically grounded analysis. This might show that the LLM captures a more general understanding of the task. However, [Davani et al. 2022] and [Uma et al. 2022] pointed out that aggregated labels, such as the ones considered golden in ASSIN2, often lead to an oversimplification of a given task, making the evaluation dataset less reliable.

The table above displaces group mismatches on annotations.

Group Disagreements	**Quantity**
ASSIN2 and Semanticist	19 pairs
LLM and Semanticist	26 pairs
LLM and ASSIN2	15 pairs
Sum	60 pairs

Table 2. Disagreement pairs

Considering ASSIN2 as a baseline, we can say Sabiá-3 performed the same as our specialist. The results also suggest the LLM is more aligned with the general understanding of the task seen in ASSIN2 labels than with a more logically grounded analysis. Since the disagreements did not occur in the same samples, we discuss these cases in depth in the next section.

5. Qualitative Analysis

Here we consider the linguistic aspects (such as semantics, pragmatics, lexical semantics, and syntax) more prominent on each pair with disagreement, considering the three annotations. Based on [Kalouli et al. 2019], we found the following categories: *loose definition, subevent, directionality, annotation error, interpretation of preposition.*

We take the label *loose definition* to categorize "concepts that are 'loose', subjective or vague to define" [Kalouli et al. 2017]. See one example from the subset below:

(1) (A) Dois meninos no sofá estão jogando vídeo games.
(B) Dois meninos estão no sofá jogando jogos na televisão.[5]

The example above is not a clear entailment (from A to B), yet there are relations between A and B. In most instances of playing video games, the event conceptually involves a screen exhibition, which can be thought of as a television. However, this is not

[5](A)Two boys on the couch are playing videogames – (B)Two boys are on the couch playing games on the television.

a logical entailment, considering that nowadays video games can be played on other devices such as PCs, mobile phones, or handheld game consoles. Therefore, we understand that the boundaries between the definitions of 'videogames' and 'games on the television' are loose.

The category *subevent* was used to label pairs of sentences that could describe the same event, yet each one focuses on different moments, or subevents, of a wider event.

(2) (A) Uma árvore está sendo apanhada por um homem.
(B) Um homem está carregando uma árvore.[6]

In this case, the verb in A, 'apanhar' (to pick), is considered part of the event 'carregar' (to carry), in the sense that the act of carrying only happens after picking up the object being carried, following [Parsons 1990]. This intuition may not align exactly with what a theory of event semantics would support, but it can be considered within the boundaries of semantics and pragmatics.

We categorized *directionality* for pairs in which B was more specific than A, therefore, by definition, one could not say that the pair had an entailment relation. It is well illustrated below:

(3) (A) A mulher está tocando a flauta.
(B) Uma mulher está habilmente tocando uma flauta.[7]

In this case, 'habilmente' (skillfully) is a subset of the set in which women play the flute. For B being more specific than A, we do not consider an entailment from A to B, but there is an entailment from B to A.

We found some *Annotation errors*. This happens when information is probably misunderstood by the annotator. In the case of human annotators, it tends to happen in pairs with long lengths, in which only one element changes.

(4) (A) Três meninos estão pulando nas folhas.
(B) Crianças em camisas vermelhas estão brincando nas folhas.[8]

It would be a clear mistake for the annotator to indicate that A entails B, as 'camisas vermelhas' (red shirts) in B adds a specific detail not present in A. While both sentences could describe children playing on the leaves, nothing in A implies or restricts the color of their shirts, as introduced in B.

We considered an *Interpretation of preposition* if the major difference between the pair's sentences was due to a change in prepositions.

(5) (A) Duas equipes estão jogando futebol **de** campo.
(B) Diferentes times estão jogando futebol **no** campo.[9]

[6](A) A tree is being picked up by a man. – (B) A man is carrying a tree.
[7](A) The woman is playing the flute. – (B) A woman is skillfully playing a flute.
[8](A) Three boys are jumping on the leaves. – (B) Children in red shirts are playing on the leaves.
[9](A) Two teams are playing soccer (football of field). – (B) Different teams are playing football on the field.

On one hand, in A the preposition ‘de’ (of) introduces a football type, a subset of the football sport played on a grass field (soccer). On the other hand, in B, the preposition ‘no’ (‘em’ + ‘o’ – ‘in the’) introduces the idea of place, but does not entail soccer.

From this categorization, the ‘loose definition’ category falls under the scope of the inherent disagreements [Pavlick and Kwiatkowski 2019]. The category ‘Interpretation of a preposition’ is hard to define. Logically, sentences of an entailed pair have different meanings, but in the context of NLI annotation tasks, these minor details (such as the interpretation of a single preposition) may be seen more as a prank than real data. It might have happened because part of the ASSIN2 corpus is a translation from the SICK corpus[Marelli et al. 2014] and some translations are just odd in Portuguese.

Most disagreements between the semanticist and the other annotations fall under the ‘loose definition’ category (10 of 15 cases). It suggests that ASSIN2 and Maritalk are more aligned with some pragmatic/contextual understanding of the language than a logician. An example of this is the pair:

(6) (A)Um homem negro está andando perto de uma loja em uma cidade grande.
(B) Um homem negro está andando próximo a um prédio em uma grande cidade.[10]

This pair was labeled both by ASSIN2 and by Sabiá-3 as entailed, although a store is not necessarily a building.

Other disagreements between the semantic expert and other annotations fall under the ‘interpretation of preposition’ category, suggesting that these refined meanings were not covered by the LLM or the general Portuguese-speaking annotators. Example (5) was annotated as an entailment both by ASSIN2 and by Sabiá-3.

Concerning Sabiá-3 disagreements, we mostly found the ‘directionality’ category (6 out 15), and the ‘subevent’ category (4 out 15) taking place. First, we argue that there is no single pattern of disagreements on the LLM, but a plethora of phenomena. More specifically, the disagreements are not motivated by the same linguistic phenomena found in the logical grounded annotation. Considering the ‘directionality’ category, it may suggest that the model gets confused by the nature of the task, labeling a relation in a specific direction only and not considering the context of the pair. It also may explain the ‘subevent’ cases, since all the sentences in those cases are somehow related. The following examples were labeled ‘Entailment’ only by Sabiá-3.

(7) (A) O cavalo está sendo montado por um homem.
(B) O homem está no passeio com o cavalo.[11]

(8) (A) Alguns homens estão jogando críquete.
(B) Um pequeno grupo de homens está alegremente jogando críquete.[12]

Finally, ASSIN2 labels were different from the two other annotations in cases of ‘loose definition’ (2 out of 4) and ‘annotation mistakes’ (2 out of 4).

[10](A) A Black man is walking near a store in a big city. – (B) A Black man is walking near a building in a big city.

[11](A) The horse is being ridden by a man. – (B) The man is on a ride with the horse.

[12](A) Some men are playing cricket. – (B) A small group of men is cheerfully playing cricket.

(9) (A) Uma mulher está pegando uma lata.
(B) Uma mulher está agarrando uma lata.[13]

(10) (A) Um homem está usando uma camisa azul e andando com os pés descalços em uma quadra de tênis.
(B) Uma pessoa está usando uma saia azul e andando descalça na quadra de tênis.[14]

Comparing the performance of Sabiá-3 and humans, Sabiá-3 got a human performance statistically. However, in the automated annotation, the mislabeled cases are not part of the 'inherent disagreements' cases that are expected to be problematic for humans. Sabiá-3 gets confused more often by the constraints of the task itself. It is also reasonable to say that, for the model, if the two sentences already shared enough contextual background, the model would find a relation between them. So, there is still room for Sabiá-3 models to get qualitative human performance.

6. Conclusions

To summarize, our study highlights annotation challenges in Natural Language Inference by comparing human and Large Language Model (LLM) annotations. Using 200 sentence pairs from the ASSIN2 dataset, we re-annotated the data with a semanticist and Sabiá-3, an LLM trained in Brazilian Portuguese. We then analyzed all the disagreements between the annotations and categorized them by the most prominent linguistic phenomenon involved in the pair.

While Sabiá-3 achieved the same agreement score as human annotators, the patterns of disagreement differed. Human annotators often diverged on pairs involving pragmatics or cultural knowledge, which are considered 'inherent disagreements'. Sabiá-3 tended to mislabel pairs that shared some context but were not logical entailments. These findings suggest that although LLMs are now statistically comparable to human performance, they exhibit distinct patterns of error and disagreement, particularly in handling specific linguistic phenomena.

References

Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabiá-2: A new generation of portuguese large language models.

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, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Chaves Rodrigues, R., Tanti, M., and Agerri, R. (2023). Natural Portuguese Language Benchmark (Napolab).

[13](A) A woman is picking up a can. – (B) A woman is grabbing a can.

[14](A) A man is wearing a blue shirt and walking barefoot on a tennis court. – (B) A person is wearing a blue skirt and walking barefoot on the tennis court.

Condoravdi, C., Crouch, D., De Paiva, V., Stolle, R., and Bobrow, D. (2003). Entailment, intensionality and text understanding. In *HLT-NAACL 2003 workshop on Text meaning*.

Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In Quiñonero-Candela, J., Dagan, I., Magnini, B., and d'Alché Buc, F., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. (2013). Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–222. Publisher Copyright: © Morgan and Claypool Publishers. All rights reserved.

Dang, H., Mecke, L., Lehmann, F., Goller, S., and Buschek, D. (2022). How to prompt? opportunities and challenges of zero- and few-shot learning for human-ai interaction in creative applications of generative models.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Fonseca, E., Borges dos Santos, L., Criscuolo, M., and Aluisio, S. (2016). Visao geral da avaliacao de similaridade semantica e inferencia textual. *Linguamatica*, 8(2).

Kalouli, A., Real, L., and de Paiva, V. (2017). Textual inference: getting logic from humans. *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*.

Kalouli, A.-L., Hu, H., Webb, A. F., Moss, L. S., and de Paiva, V. (2023). Curing the SICK and Other NLI Maladies. *Computational Linguistics*, 49(1):199–243.

Kalouli, A.-L., Real, A. B. L., Palmer, M., and de Paiva, V. (2019). Explaining simple natural language inference. *Proceedings of the 13th Linguistic Annotation Workshop (LAW 2019), ACL*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*.

Parsons, T. (1990). *Events in the semantics of English: A study in Subatomic Semantics*. MIT Press/Cambrige, London.

Pavlick, E. and Kwiatkowski, T. (2019). Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Real, L., Fonseca, E., and Oliveira, H. G. (2020). Organizing the assin 2 shared task. *ASSIN 2 Shared Task: Evaluating Semantic Textual Similarity and Textual Entailment in Portuguese*.

Real, L., Rodrigues, A., Vieira e Silva, A., Albiero, B., Thalenberg, B., Guide, B., Silva, C., de Oliveira Lima, G., Câmara, I. C. S., Stanojević, M., Souza, R., and de Paiva, V. (2018). Sick-br: A portuguese corpus for inference. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR*

2018, Canela, Brazil, September 24–26, 2018, Proceedings, page 303–312, Berlin, Heidelberg. Springer-Verlag.

Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of portuguese with transformer albertina pt-*. In Moniz, N., Vale, Z., Cascalho, J., Silva, C., and Sebastião, R., editors, *Progress in Artificial Intelligence*, pages 441–453, Cham. Springer Nature Switzerland.

Salvatore, F. d. S. (2020). *Analyzing natural language inference from a rigorous point of view*. PhD thesis, Universidade de Sao Paulo.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2022). Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Williams, A., Nangia, N., and Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv*.

Zhang, X. F. and de Marneffe, M.-C. (2021). Identifying inherent disagreement in natural language inference. 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 4908–4915, Online. Association for Computational Linguistics.