Textual Inference in Portuguese: Comparing Language Models

Fabiana Avais[1,2], **Valeria de Paiva**[3], **and Livy Real**[4,5]
[1] Universidade Federal do Paraná
[2] Universidade Estadual de Ponta Grossa
[3] Topos Institute, Berkeley, USA
[4] Universidade Federal do Amazonas
[5] Instituto Kunumi
avaisfabiana@gmail.com valeria@topos.institute livy@kunumi.com

Abstract

Large language models (LLMs) are increasingly used for Natural Language Inference (NLI), yet their ability to perform logic-sensitive semantic reasoning, especially outside English, remains underexplored. This paper presents a preliminary investigation into the feasibility and usefulness of developing FraCaS-BR, a Portuguese adaptation of the FraCaS benchmark for semantic inference. Using a small diagnostic subset of seven FraCaS problems focusing on generalized quantifiers, plurals, and nominal anaphora, we evaluate the behavior of three LLMs (ChatGPT, Maritalk, and Evaristo) on Brazilian Portuguese translations. Each problem is submitted multiple times to assess correctness, variance, and consistency relative to the original FraCaS gold labels. The results reveal systematic differences across models. While ChatGPT shows higher overall correctness and stability, all models exhibit limitations that undermine their reliability on logic-controlled inference tasks. The extent of manual correction required during translation further underscores the necessity of human-in-the-loop evaluation. Taken together, these findings support and motivate the development of FraCaS-BR as a controlled evaluation resource for assessing semantic reasoning in Portuguese.

1 Introduction

The European Commission project *Framework for Computational Semantics*, abbreviated as FraCaS (Cooper et al., 1996)[1], ran from 1993 to 1996. Its goal was to develop an informal framework for comparing semantic approaches to language, both in terms of their theoretical claims and their suitability for implementation. The project was highly successful: the collection of examples devised to compare different formalisms is still in use today, more than thirty years later, as a benchmark for a wide range of reasoning and semantic tasks.

One of the main semantic tasks is still Natural Language Inference (NLI). There are plenty of resources, corpora, scripts, and competitions to evaluate NLI in English. However, there are fewer resources for NLI in Portuguese. The corpus ASSIN (for *Avaliação de Similaridade Semântica e Inferência Textual*) (Fonseca et al., 2016) is "a corpus annotated with pairs of sentences written in Portuguese that is suitable for the exploration of textual entailment and paraphrasing classifiers"[2]. The corpus does not consider contradictions, a serious issue if one thinks logically about inference.

A second corpus for NLI in Portuguese is SICK-BR [3], a careful translation of the SICK corpus of Marelli et al. (2014). Unlike ASSIN, the corpus SICK was meant to have simplified sentences, as far as linguistic phenomena are concerned. Thus, named entities, complicated time expressions, and world knowledge are kept at a minimum, as is the size of the vocabulary. The task, the corpus used for the task, and the evaluation results are described in Real et al. (2020).

InferBR (Bencke et al., 2024) is a third Portuguese NLI resource in which premises are semi-automatically generated, and hypotheses are subsequently generated automatically. Premises are constructed from two source datasets: PraCegoVer (dos Santos et al., 2021) and SICK-BR (Real et al., 2018). These datasets are processed using GPT-4 and transformed into premise sentences. Hypotheses are then generated via few-shot prompt engineering, where each premise serves as input for producing three hypotheses corresponding to the labels entailment, contradiction, and neutral. The overall generation and evaluation process was reviewed by three human annotators.

[1] https://cordis.europa.eu/project/id/LRE62051

[2] https://huggingface.co/datasets/assin

[3] https://github.com/livyreal/SICK-BR

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 201–209
April 13-16, 2026

Our long-term goal is to produce a corpus called FraCas-BR (de Paiva and Real, 2024), a resource that goes back to the original goal of benchmarking a large selection of semantic phenomena. We hope to obtain manually checked translations of FraCaS sentences into Portuguese. We plan to verify, through careful annotation work, that the (mostly logical) phenomena described in English remain the focus of the new dataset and that the semantic phenomena 'behave' in Portuguese the same way they do in English. This is similar to the work in Amblard et al. (2020) for French and in the MultiFraCaS project[4] for Farsi, German, Greek, and Mandarin.

The original FraCas corpus consists of 346 'problems', each problem contains one or more premises and one question. There are a total of 536 premises, or an average of 1.55 premises per problem. This work presents a preliminary study on the feasibility and usefulness of developing FraCaS-BR. We conduct a small-scale experiment with seven examples, distributed across the three inference labels, entailment, contradiction and neutral. After translating these examples into Portuguese, we examine (i) whether the translations preserve the original logical relations and (ii) whether Portuguese-language LLMs can interpret them correctly. For this, we analyzed how three different LLMs (ChatGPT [5], Maritalk [6], and Evaristo [7]) deal with the FraCas problems in Portuguese. We conclude by summarizing our findings, which indicate a strong need for an evaluation resource of this kind for Portuguese.

Related Work

Haruta et al. (2020) proposed an end-to-end logic-based inference system for labeling both comparatives and generalized quantifiers, and evaluated it with FraCas. The system is successful in the task, and has five modules: implementation of a Combinatory Categorical Grammar parser, transformation to syntactic trees, semantic parsing, conversion to formal logics, and automatic inference.

Bernardy and Chatzikyriakidis (2021) created a similar NLI system, which transforms syntax trees into logical formulas (using the Rocq proof assistant[8]). Their goal was to handle temporal semantics in the whole FraCas dataset, and they obtained an overall accuracy of 81%, and 73% on temporal reference problems.

Amanaki et al. (2022) translated FraCas to Greek with human validation. They also added 428 new problems to the dataset, specifically to deal with Greek syntax.

Taken together, these studies underscore both the continued relevance of FraCaS as a benchmark for logical inference and the practicality of adapting it to new languages. However, little is known about how contemporary large language models perform on such tightly controlled, logic-oriented inference problems in Portuguese. In this preliminary study, we therefore examine seven FraCaS examples translated into Portuguese and analyze the responses produced by three large language models, offering a focused exploration of the challenges involved.

2 Evaluating Large Language Models

Evaluating the logical reasoning capabilities of LLMs is increasingly important. Although LLMs are now widely deployed and can generate fluent, persuasive text, evidence shows that they continue to struggle with basic logical reasoning tasks that are straightforward for humans (Suzgun et al., 2024). These models excel at predicting context and recognizing linguistic patterns, but this often comes at the expense of sound reasoning, with convincing chains of thought sometimes masking logical errors.

Recent work on legal applications (Trautmann et al., 2024) shows that combining LLMs with classical NLI frameworks yields strong performance in legal question answering, in part because NLI supports auditable claim verification. In this context, we argue that a general-purpose semantic resource for Portuguese would significantly strengthen the current evaluation landscape. Our preliminary study takes a first step in this direction by using a small, semantically and logically representative dataset to explore both the challenges of translating the FraCaS corpus into Portuguese and the logical reasoning behavior of LLMs in that language.

3 Our project

Given the current capabilities and widespread use of LLMs, a natural research question arises: is it still relevant to fully translate the FraCaS corpus into Portuguese, or can contemporary LLMs already handle its inference problems reliably? An-

[4] https://gu-clasp.github.io/multifracas/
[5] https://chat.openai.com
[6] https://www.maritaca.ai
[7] https://evaristo.ai
[8] https://rocq-prover.org/

swering this question requires evaluating not only whether LLMs assign the correct inference labels (correctness), but also how stable and confident their predictions are across runs (variance and consistency).

This leads to a secondary question: which LLM performs most effectively on semantic inference tasks in Brazilian Portuguese, and therefore produces the most reliable labels for FraCaS-style problems? By addressing these questions, we aim to determine whether a complete Portuguese translation of the FraCaS framework remains necessary, or whether its evaluative role is already subsumed by state-of-the-art LLMs.

FraCaS problems are organized by linguistic phenomenon, including generalized quantifiers, plurals, (nominal) anaphora, ellipsis, adjectives, comparatives, temporal reference, verbs, and attitudes. This experiment is intended both to probe LLM performance on logically controlled inference tasks in Portuguese and to inform best practices for translating the full FraCaS corpus based on observed model behavior.

3.1 Translating FraCas to Portuguese

The first task was to validate both the dataset's translation and its logical interpretation in Portuguese.

The FraCas sections selected for this work were generalized quantifiers, plurals, and nominal anaphora. More specifically, we chose seven problems listed in Table 1. A first experiment was conducted using the free versions of three large language models: ChatGPT, Maritalk, and Evaristo. These models were selected for complementary reasons. ChatGPT was included due to its widespread use and accessibility. Maritalk because it is a Brazilian LLM developed by Maritaca AI and trained specifically on Brazilian Portuguese data, with support for integration into platforms such as LangChain and Langflow. Evaristo, by contrast, is a recently released European Portuguese chatbot, designed around principles of open AI and user privacy, and built on open-source LLMs to promote transparency and community involvement. We use both variants of Portuguese (European and Brazilian) because criteria such as 'naturality' of the translation into Portuguese depends on the variant of Portuguese used by the human annotator.

The experimental runs took place in August 2025. Each of the seven problems was translated once by each model. Following a linguistic analysis of the outputs, we chose to manually correct the version produced by ChatGPT, as its initial translations more consistently preserved the intended semantic relations and exhibited a more natural syntactic structure, particularly in cases involving monotonicity. The final, curated translations are presented in Table 2.

3.2 Predicting NLI labels

After validating the translations and their logical adequacy, we turn to the second step of our investigation: evaluating how LLMs perform on the NLI task itself. Our experiment examines whether different models can correctly and consistently assign FraCaS-style inference labels to the translated problems, shifting the focus from translation quality to semantic reasoning behavior.

Here, we test the seven Portuguese problems across the three LLMs under investigation. Each problem instance was run ten times in each LLM, and the same prompt was used each time. The prompt was built with meta-prompting, which relies on writing a prompt and asking an LLM to improve it (Schulhoff et al., 2025). The prompt is in Portuguese below.

Analise se a HIPÓTESE decorre logicamente da PREMISSA. Responda apenas com uma das seguintes opções: YES, NO, ou UNKNOWN. Responda YES se a hipótese for uma consequência lógica da premissa (ou seja, sempre for verdadeira quando a premissa for verdadeira). Responda NO se a hipótese contradiz a premissa ou não pode ser verdadeira ao mesmo tempo que a premissa. Responda UNKNOWN se a premissa não fornece informação suficiente para determinar a veracidade da hipótese.[9]

This analysis focuses on a small subset of FraCaS problems involving generalized quantifiers (e.g. *todo, algum, poucos*). The subset is not intended to be statistically representative of the full FraCaS corpus; rather, it serves as an exploratory diagnostic sample. Because the FraCaS dataset spans a wide range of semantic phenomena, each section places different semantic demands on language models. Working with a limited subset therefore allows for an initial assessment of LLM be-

[9]Answer with only one of the following options: YES, NO, or UNKNOWN. Answer YES if the hypothesis is a logical consequence of the premise (that is, it is always true when the premise is true). Answer NO if the hypothesis contradicts the premise or cannot be true at the same time as the premise. Answer UNKNOWN if the premise does not provide enough information to determine the truth of the hypothesis.

ID	Premises	Hypothesis	Label
6	No really great tenors are modest.	Are there really great tenors who are modest?	NO
35	All Europeans can travel freely within Europe. Every European is a person. Every person who has the right to live in Europe can travel freely within Europe.	Do all Europeans have the right to live in Europe?	UNKNOWN
50	Every Canadian resident can travel freely within Europe. Every Canadian resident is a resident of the North American continent.	Can every resident of the North American continent travel freely within Europe?	UNKNOWN
96	The Ancient Greeks were all noted philosophers.	Was every Ancient Greek a noted philosopher?	YES
137	There are 100 companies. ICM is one of the companies and owns 150 computers. It does not have service contracts for any of its computers. Each of the other 99 companies owns one computer. They have service contracts for them.	Do most companies that own a computer have a service contract for it?	YES
211	All elephants are large animals. Dumbo is a small elephant.	Is Dumbo a small animal?	NO
223	The PC-6082 is faster than the ITEL-XZ. The PC-6082 is slow.	Is the ITEL-XZ fast?	NO

Table 1: Inference examples with premises, hypotheses, and labels

havior, making it possible to identify patterns in their predictions and gain insight into their general semantic competence.

Even with only seven problems, this diagnostic sample can reveal systematic behaviors, such as biases in label distribution or recurring inference errors. Moreover, failure to correctly handle these basic quantifier-related cases makes it unlikely that a model would perform well on the full FraCaS dataset. For this reason, a small but carefully chosen subset already provides meaningful evidence about model capabilities and limitations.

Using this subset also allows us to examine the stability of LLM responses and their ability to consistently capture semantic relations. Each problem was submitted ten times to each model, enabling analysis along three dimensions: correctness, variance, and consistency.

3.3 Comparing to Gold

In the first stage of the analysis, we evaluate the correctness of each model's predictions by aggregating results across the ten runs. The corresponding results are reported in Table 3.

For each problem, we determine the majority label across the ten runs and compare this aggregated prediction with the original FraCaS gold label. In this context, correctness is defined strictly as agreement with the FraCaS annotation. Table 3 summarizes these comparisons.

Overall, ChatGPT correctly labeled 6 out of the 7 problems, whereas Evaristo and Maritalk correctly labeled 4. When performance is broken down by label, ChatGPT correctly predicted 2 of the 2 'Yes' cases (IDs 96, 137), while Maritalk correctly predicted 1 and Evaristo predicted 1. For the 'No' label, ChatGPT and Evaristo achieved perfect performance (3/3), whereas Maritalk again labeled only 1 correctly. For the 'Unknown' label, Chat-

ID	Premises	Hypothesis	Label
6	Não há grandes tenores modestos.	Há grandes tenores modestos?	NO
35	Todos os europeus podem viajar livremente pela Europa. Todo europeu é uma pessoa. Toda pessoa que tem o direito de viver na Europa pode viajar livremente pela Europa.	Todos os europeus têm o direito de viver na Europa?	UNKNOWN
50	Todo residente canadense pode viajar livremente pela Europa. Todo residente canadense é um residente do continente norte-americano.	Todo residente do continente norte-americano pode viajar livremente pela Europa?	UNKNOWN
96	Os gregos antigos eram todos filósofos notáveis.	Todo grego antigo era filósofo notável?	YES
137	Existem 100 empresas. ICM é uma empresa e possui 150 computadores. Ela não tem contratos de serviço para nenhum dos seus computadores. Cada um das outras 99 empresas possui um computador. Elas têm contratos de serviço para eles.	A maioria das empresas que têm computador tem contratos de serviços?	YES
211	Todos os elefantes são animais grandes. Dumbo é um pequeno elefante.	Dumbo é um animal pequeno?	NO
223	O PC-6082 é mais rápido do que o ITEL-XZ. O PC-6082 é lento.	O ITEL-XZ é rápido?	NO

Table 2: Translated inference examples with premises, hypotheses, and labels

ID	Gold	ChatGPT	Maritalk	Evaristo
6	No	No	Unknown	No
96	Yes	Yes	Yes	No
35	Unknown	Unknown	Unknown	Yes
50	Unknown	No	Unknown	Yes
137	Yes	Yes	Unknown	Yes
211	No	No	No	No
223	No	No	Unknown	No

Table 3: Comparison between gold labels and model predictions.

GPT correctly predicted 1 case, Maritalk correctly predicted 2, and Evaristo did not predict this label correctly in any instance.

Taken together, the results in Table 3 indicate that ChatGPT exhibits higher overall correctness than both Evaristo and Maritalk, as well as a more balanced distribution of predicted labels. Maritalk shows a tendency to over-predict the Unknown label, while Evaristo fails to predict this label altogether.

3.4 Consistency of answers

The second stage of the analysis focuses on consistency across the ten runs for each model. Specifically, we measure how frequently a model assigns the same label to the same problem, which serves as an indicator of confidence in its predictions.

A first inspection of Table 4 shows that, for every problem, the models vary their predictions between at most two labels; no problem is assigned all three labels by any model. This suggests a baseline level of internal consistency in the models' responses.

A further pattern emerges from the distribution of label variation. Models rarely alternate between the logically contradictory labels Yes and No for the same problem. Instead, most variability involves pairs such as Yes/Unknown or No/Unknown. Instances of direct Yes/No alternation are rare: ChatGPT exhibits a single such case (ID 211), while Evaristo shows two (IDs 35 and 96). Moreover, the only case in which the aggregated major-

Table 4: ChatGPT, Maritalk and Evaristo predictions

ChatGPT											
ID	**Gold**	1	2	3	4	5	6	7	8	9	10
6	No	No	No	No	No	No	No	No	No	No	No
35	Unk	Unk	Yes	Unk	Unk	Yes	Unk	Yes	Unk	Unk	Yes
50	Unk	No	No	No	No	No	No	No	No	No	No
96	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
137	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
211	No	No	No	Yes	No	No	No	No	No	No	No
223	No	No	No	No	No	No	No	No	No	No	No
Maritalk											
ID	**Gold**	1	2	3	4	5	6	7	8	9	10
6	No	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
35	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
50	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk
96	Yes	Yes	Yes	Unk	Unk	Yes	Yes	Unk	Yes	Unk	Unk
137	Yes	No	No	No	No	No	Unk	No	Unk	No	Unk
211	No	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	Unk	No
223	No	Unk	Unk	Unk	Unk	No	No	Unk	Unk	Unk	No
Evaristo											
ID	**Gold**	1	2	3	4	5	6	7	8	9	10
6	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
35	Unk	Yes	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
50	Unk	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
96	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	Yes
137	Yes	No	No	No	No	No	No	No	No	No	No
211	No	No	No	No	No	No	No	No	No	No	No
223	No	No	No	No	No	No	No	No	No	No	No

ity label conflicts as Yes/No arises with Evaristo (Table 1, ID 96). Overall, mislabeling involving Unknown is far more frequent than mislabeling between Yes and No, suggesting that the semantic interpretation of the Unknown label poses a particular challenge for LLMs. This is further reflected in systematic tendencies across models: Maritalk tends to over-predict Unknown, whereas Evaristo systematically under-predicts it.

Beyond label distribution, we analyze consistency across the ten runs for each model as an indicator of confidence in its predictions.

Consistency was evaluated at ID level, measuring the stability of the assigned labels across multiple model runs. For each label, only IDs that matched entirely the ten runs were considered. Under this metric, ChatGPT showed the highest consistency for the No label (0.75), moderate consistency for Yes (0.50), and no consistency for Unknown (0.00). Applying the same metric to Maritalk, a lack of consistency was observed for the Yes and No labels, since no ID that received these labels maintained the same label across the ten runs. In contrast, the Unknown label showed the moderate consistency of 0.43. Finally, Evaristo achieved moderate consistency with Yes (0.50) and No (0.60), and absence of the label Unknown. On this measure, ChatGPT emerges as the most consistent model.

Table 4 illustrates the variance in label distributions across models. ChatGPT shows the most balanced profile. Maritalk, by contrast, displays a strong bias toward 'Unknown', selecting this label too many times. This skewed distribution helps explain its lower overall correctness and highlights differing strategies adopted by LLMs when faced with semantic uncertainty.

Finally, we are at a loss to explain why Evaristo did not find a single 'Unknown' in the whole set. At the time of this research, Evaristo was in beta, so we expected this issue to evolve in subsequent versions. Also, Evaristo is designed to process Eu-

ropean Portuguese, but this focus alone is unlikely to explain the observed difficulty, since logical reasoning in both Portuguese variants is largely the same.

A natural follow-up question is whether LLMs appear more confident when their predicted label matches the gold standard. For ChatGPT, this does not seem to be the case: the model shows very little variation across runs, regardless of whether its aggregated prediction is correct or incorrect. More precisely, ChatGPT's predictions vary in only two (ID #35 and #211) of the seven problems, and it produces fewer mislabels overall than the other models.

For instance, in ID 35, where the correct label is Unknown, ChatGPT alternates between 'Unknown' (6) and 'Yes' (4) times. Despite this variance, the aggregated prediction is correct. A similar pattern appears in ID 211, where the correct label is 'No' and ChatGPT predicts 'No' in 9 out of 10 runs, with a single 'Yes'. In both cases, correctness is preserved at the aggregate level despite some instability. ChatGPT mislabels only one problem: ID 50 (predicted 'No' instead of Unknown) and it might be said that 'world knowledge' could have played a role.

Maritalk mislabels three problems (IDs 6, 137, and 223). In one of these cases (ID 6), there is no variation across runs: the model consistently assigns an incorrect label, yielding 100% incorrect predictions. In the remaining cases, variance is present. For ID 137, where the correct label is 'Yes', Maritalk predicts No (7 times) and Unknown (3 times). For ID 223, where the correct label is 'No', it predicts 'Unknown' in 7 runs and 'No' in 3. Even when Maritalk produces correct predictions, its answers often vary, but the more salient pattern is its systematic tendency to over-predict the 'Unknown' label, including in cases where 'Yes' or 'No' is warranted.

Evaristo mislabels the same number of problems as Maritalk (IDs 35, 50, and 96), but exhibits a different failure mode. The model never predicts the 'Unknown' label, and its outputs across the seven problems vary only between Yes and No. As a result, problems whose correct label is 'Unknown' are systematically misclassified, indicating a fundamental limitation for NLI tasks that rely on three-way inference distinctions.

In summary, ChatGPT outperforms the other models in both correctness and consistency, exhibiting the lowest variance and the most balanced label distribution. Maritalk's tendency to overuse 'Unknown' degrades its performance, while Evaristo's complete avoidance of this label makes it unsuitable for the NLI task considered here.

Taken together, these results highlight systematic and model-specific limitations in handling logic-sensitive semantic distinctions, particularly with respect to the Unknown label. The divergent behaviors observed suggest that current LLMs do not yet provide a reliable substitute for carefully curated inference Portuguese benchmarks. Rather than eliminating the need for resources such as FraCaS-BR, these findings reinforce their importance as controlled evaluation tools for diagnosing semantic competence in Portuguese. Moreover, the extent of manual correction required even in this small pilot study underscores the necessity of a human-in-the-loop approach to corpus construction and validation, especially for logic-based resources where subtle semantic errors can invalidate entire inference patterns. These considerations motivate the broader discussion of FraCaS-BR as both an evaluation benchmark and a methodological safeguard for assessing LLM reasoning in Portuguese.

4 Discussion

The results of this preliminary study point to clear and systematic limitations in current LLMs when confronted with logic-sensitive semantic inference tasks in Portuguese. Although all three models exhibit a baseline level of internal consistency, their behavior diverges sharply with respect to correctness, label balance, and the treatment of semantic underspecification. In particular, the 'Unknown' label emerges as a persistent source of difficulty, either being overused (Maritalk) or entirely avoided (Evaristo), with only ChatGPT showing a more balanced, though still imperfect, handling of three-way inference distinctions.

These findings suggest that LLMs do not uniformly internalize the semantic conditions underlying FraCaS-style inference. While models rarely oscillate between the logically contradictory labels 'Yes' and 'No', they frequently collapse uncertainty into 'Unknown' or eliminate it altogether. This pattern indicates that LLMs may rely on surface-level heuristics or distributional cues rather than robust semantic representations capable of sustaining underspecified or indeterminate interpretations.

Crucially, higher consistency does not always correlate with correctness. ChatGPT's performance

illustrates that stable predictions can still be wrong, while Maritalk's high consistency on incorrect labels reveals systematic semantic bias rather than random noise. These observations reinforce the need to evaluate LLMs along multiple dimensions (correctness, variance, and consistency) rather than accuracy alone. Logic-based benchmarks such as FraCaS are particularly well suited to exposing these distinctions, as small semantic errors can be clearly traced to specific linguistic phenomena.

Taken together, these results argue strongly for the relevance of FraCaS-BR as a dedicated evaluation resource for Portuguese. Rather than being obviated by state-of-the-art LLMs, the need for a carefully translated and validated FraCaS corpus becomes more pressing in light of their uneven performance. FraCaS-BR can serve both as a diagnostic benchmark for assessing LLM reasoning and as a methodological anchor for future work on semantic inference, hybrid symbolic–neural systems, and language-specific evaluation in Portuguese.

Finally, while this study is necessarily extremely limited in scale, its findings provide concrete guidance for future work. Expanding FraCaS-BR to cover additional semantic phenomena, increasing the number of annotated examples, and incorporating formal consistency metrics will allow for more robust comparisons across models and languages. More broadly, the results suggest that progress in LLM reasoning should be measured not only by fluency or task performance, but by the ability to respect the logical structure of meaning—a goal for which FraCaS remains a uniquely valuable benchmark.

5 Conclusions and Future Work

This work presented a preliminary evaluation of Large Language Models on a small subset of FraCaS problems translated into Brazilian Portuguese. Rather than aiming at broad generalization, the experiment was designed as a diagnostic analysis, intended to reveal whether systematic patterns of success and failure already emerge when LLMs are confronted with logically grounded inference problems in Portuguese.

The results indicate that, even in a limited setting, LLMs display difficulties with semantic inference as in the FraCaS framework. While all models performed relatively well on problems labeled 'No', performance varied considerably for 'Yes' and, most notably, for 'Unknown'. The latter emerged as the most challenging label for Maritalk and Evaristo models. This may suggest that some current LLMs do not reliably encode the distinction between the absence of information and logical contradiction, a distinction that is central to formal semantic reasoning.

Differences between models were not limited to overall correctness, but also involved systematic biases in label usage. ChatGPT achieved the highest correctness with relatively low variance across runs, indicating stable behavior across repeated queries. Maritalk showed a strong tendency to overlabel 'Unknown', which resulted in confident but often incorrect classifications for problems whose correct label were 'Yes' or 'No'. Evaristo[10], in contrast, seems to ignore the 'Unknown' label, reducing the task to a binary classification problem. These patterns suggest that model-specific strategies or training biases strongly influence how inference categories are interpreted.

From the perspective of resource development, these findings support the relevance of translating and adapting FraCaS to Portuguese. The observed errors cannot be attributed solely to translation issues or surface-level linguistic ambiguity, but instead reflect deeper challenges in modeling logical inference. As such, relying exclusively on state-of-the-art LLMs does not eliminate the need for carefully constructed semantic benchmarks. On the contrary, benchmarks like FraCaS-BR remain essential for diagnosing specific reasoning failures that are not easily captured by large-scale NLI datasets.

Finally, this study highlights the importance of evaluation frameworks grounded in semantics, especially in multilingual contexts. While LLMs have demonstrated impressive linguistic capabilities, their performance on logically controlled inference tasks remains uneven. Expanding FraCaS-BR and applying it to a broader set of models and semantic phenomena constitutes a natural next step toward a more principled evaluation of logical reasoning in Portuguese-language LLMs.

References

Eirini Amanaki, Jean-Philippe Bernardy, Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik,

[10]We ran the beta version of Evaristo, which may explain the limitations found in it. We reported the problematic cases to the developers, and that's why we want to test Evaristo in the future to see if it has improved, believing in the power of the community surrounding an open-source LLM.

Aram Karimi, Adam Ek, Eirini Chrysovalantou Giannikouri, Vasiliki Katsouli, Ilias Kolokousis, Eirini Chrysovalantou Mamatzaki, Dimitrios Papadakis, Olga Petrova, Erofili Psaltaki, Charikleia Soupiona, Effrosyni Skoulataki, and Christina Stefanidou. 2022. Fine-grained entailment: Resources for Greek NLI and precise entailment. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 44–52, Marseille, France. European Language Resources Association.

Maxime Amblard, Clément Beysson, Philippe de Groote, Bruno Guillaume, and Sylvain Pogodalla. 2020. A french version of the fracas test suite. In *LREC 2020-Language Resources and Evaluation Conference*, page 9.

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.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2021. Applied temporal analysis: A complete run of the FraCaS test suite. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 11–20, Groningen, The Netherlands (online). Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan van Eijck, Chris Fox, Josef van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. Using the framework. the FraCaS Consortium.

Valeria de Paiva and Livy Real. 2024. Towards FraCas-BR. *OpenCor Workshop*.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2021. #pracegover: A large dataset for image captioning in portuguese. *Preprint*, arXiv:2103.11474.

Erick Rocha Fonseca, Leandro Borges dos Santos, Marcelo Criscuolo, and Sandra Maria Aluísio. 2016. ASSIN: Avaliação de similaridade semântica e inferência textual. In *PROPOR*, pages 1–8.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2020. Logical inferences with comparatives and generalized quantifiers. *Preprint*, arXiv:2005.07954.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC 2014*.

Livy Real, Erick Fonseca, and Hugo Gonçalo Oliveira. 2020. The ASSIN 2 shared task: A quick overview. In *PROPOR 2020, Evora, Portugal*, volume 12037 of *LNCS*, pages 406–412. Springer.

Livy Real, Ana Rodrigues, Andressa Vieira e Silva, Beatriz Albiero, Bruno Guide, Bruna Thalenberg, Cindy Silva, Igor C. S. Câmara, Guilherme de Oliveira Lima, Rodrigo Souza, Milos Stanojevic, and Valeria de Paiva. 2018. SICK-BR: a portuguese corpus for inference. In *PROPOR 2018*.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2025. The prompt report: A systematic survey of prompt engineering techniques. *Preprint*, arXiv:2406.06608.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E. Ho, Thomas Icard, Dan Jurafsky, and James Zou. 2024. Belief in the machine: Investigating epistemological blind spots of language models. *Preprint*, arXiv:2410.21195.

Dietrich Trautmann, Natalia Ostapuk, Quentin Grail, Adrian Pol, Guglielmo Bonifazi, Shang Gao, and Martin Gajek. 2024. Measuring the groundedness of legal question-answering systems. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 176–186, Miami, FL, USA. Association for Computational Linguistics.