From Bag-of-Words to Reasoning: Comparing Traditional Supervised ML and Zero-Shot LLMs for Sexual Predator Identification in Brazilian Portuguese

Leonardo Ferreira dos Santos[1], **Gustavo Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

leonardo.santos@eic.cefet-rj.br, gustavo.guedes@cefet-rj.br

Abstract. *Automated detection of online sexual predators has traditionally relied mostly on supervised classifiers using both textual and engineered features and annotation labels, simplifying the complexities of sexual predatory behavior. With the emergence of LLMs and the scarcity of real-world data, it is important to explore their potential in this research domain. In this work, four commercial LLMs with reasoning capabilities are evaluated in zero-shot mode on the PREDADORES-BR dataset for binary classification of predatory conversations in Brazilian Portuguese. The best-performing model achieved F_1 = 96% with 100% precision and zero false positives, with recall statistically indistinguishable from the best supervised baseline (SVM, F_1 = 89.87%).*

1. Introduction

Online sexual predation against children and adolescents has been a concern worldwide for decades. The proliferation of social applications, online games, and messaging platforms has favored predators, who exploit these channels to contact, groom, and ultimately abuse minors [Kloess et al., 2019]. In Brazil, 92% of children and adolescents aged 9-17 years access the internet, with 96% using mobile phones as their primary device [Comitê Gestor da Internet no Brasil, 2025], creating a large pool of potential victims. In this context, researchers have developed computational models to identify sexual predators (SPI) in text conversations [Villatoro-Tello et al., 2012; Ebrahimi et al., 2016] typically framing the task as a supervised text classification problem.

The supervised approaches share two characteristics. First, they require labeled training data, either from sting operations such as Perverted Justice[1] or, very rarely, from real legal records. Second, while post-hoc interpretability techniques exist for traditional classifiers, such as feature ablation, chi-squared tests, and model-agnostic methods like LIME [Ribeiro et al., 2016] and SHAP [Lundberg and Lee, 2017], these methods identify *which textual features* drove a classification decision, not *which theoretical constructs of predatory behavior* are present in a conversation.

Due to limited publicly available real-world research data, such as legal research, large language models (LLMs) are emerging as valuable tools for modeling SPI beyond basic textual features. They can analyze higher-level, conversation-derived features and improve the interpretability of chat context through reasoning. Combining these reasoning abilities with established frameworks like the Luring Communication

[1]http://www.perverted-justice.com/

Theory (LCT) [Olson et al., 2007], which describes predatory communication in terms of deceptive trust, desensitization, and isolation, highlights a disconnect with traditional bag-of-words features and TF-IDF weights.

In this study, LCT provides conceptual motivation rather than an evaluated coding framework. We evaluate four commercial LLMs with reasoning capabilities in zero-shot mode: Claude Opus 4.6[2] (Anthropic), GPT-5.2[3] (OpenAI), Gemini 3.1 Pro Preview[4] (Google), and Grok 4.20[5] (xAI). Using the PREDADORES-BR dataset [Santos and Guedes, 2020], we compare their zero-shot binary classification performance against the traditional supervised ML baselines reported in Santos and Guedes [2020].

2. Related work

Since the SPI domain was first defined as a standard text classification task [Pendar, 2007], it gained momentum with the PAN-2012 competition at CLEF [Inches and Crestani, 2012]. Since then, researchers have explored various methods, including different feature representations within traditional ML models, and examined the domain from multiple perspectives. Recently, detecting predatory behavior using machine learning has advanced from traditional techniques to transformer architectures [Borj et al., 2023].

Brazilian researchers have also investigated the early detection of sexual grooming. Milon-Flores and Cordeiro [2022] identified seven behavioral traits from the NRC emotion lexicon and introduced the BF-PSR framework. Using the unfiltered PAN-2012 dataset — renamed as SGD (Sexual Groomer Detection) — his framework achieved F_1 = 57% with 10% of the conversation content, increasing to 72% with 100%. Additionally, two new publicly available datasets, PJZ and PJZC, combine Perverted Justice conversations with IRC logs and chat data. Panzariello [2022] employed different strategies with the PAN-2012 dataset, including SVM, Random Forest, kNN, MLP, and BERT. The top strategy achieved $F_{0.5}$ = 85.96% without data balancing and $F_{0.5}$ = 99.89% with subsampling, both using only the first 10 messages of each conversation.

A related research line has already addressed SPI in Brazilian Portuguese from complementary perspectives. In dos Santos and Guedes [2018], the focus was on psycholinguistic interpretation, with LIWC-based analyses used to investigate narcissistic traits in predator communication. In dos Santos and Guedes [2019], the emphasis shifted to automatic detection, using conversations extracted from legal records and Convolutional Neural Networks for preliminary binary classification in Portuguese. This agenda was consolidated in Santos and Guedes [2020], which introduced the PREDADORES-BR dataset and evaluated multiple traditional supervised ML classifiers for Brazilian Portuguese SPI, with SVM achieving the best reported F_1 score of 89.87%.

An alternative line of research has explored fine-tuning open-source LLMs for predatory chat classification rather than using them in zero-shot mode. Nguyen et al. [2024] fine-tuned Llama 2 7B via LoRA/PEFT on the PAN-2012 dataset as a one-size-fits-all classifier, achieving *Acc* = 100%, F_1 = 98%, and $F_{0.5}$ = 98%, surpassing all prior methods including SimCSE-based fusion models, while also demonstrating cross-lingual

[2]https://platform.claude.com/docs/en/about-claude/models/overview
[3]https://developers.openai.com/api/docs/models/gpt-5.2
[4]https://deepmind.google/models/gemini/pro/
[5]https://docs.x.ai/developers/models

capability on Roman Urdu and Urdu abusive language datasets. Recently, Hamm and McKeever [2025] fine-tuned the smaller LLaMA 3.2 1B model on the PAN-2012 dataset and achieved F_1 = 99%, *P* = 99%, and *R* = 99% for SPI, while additionally investigating the role of tone — finding that predators use a positive tone in 46.85% of messages compared to 25.86% for non-predators.

Despite the growing amount of research on SPI, most studies mainly focus on English data, with the PAN-2012 dataset serving as the standard benchmark for nearly all research. The closest alternative in Brazilian Portuguese language is the publicly available dataset, PREDADORES-BR. Additionally, no studies published after 2020 have examined SPI using Brazilian Portuguese data, and no previous work has applied commercial LLMs to Portuguese-language predatory conversations — whether in zero-shot or fine-tuned settings. This work aims to fill that gap.

3. Methodology

3.1. Dataset

We use the PREDADORES-BR dataset, which contains 39 predatory conversations obtained from Brazilian legal records and a larger set of non-predatory conversations collected from public Discord servers across different thematic categories. From the non-predatory pool, we randomly sampled 39 conversations to form a balanced evaluation set of 78 dyadic conversations (i.e., involving two participants). The dyadic restriction is theoretically motivated: LCT models predatory communication as an inherently dyadic process [Olson et al., 2007]. We define a single task (binary classification): each model receives a full conversation and must classify it as predatory or non-predatory, and justify its classification.

3.2. Models and Configuration

The current work selected four commercial LLMs with reasoning capabilities for the experiments in zero-shot mode: Claude Opus 4.6 (`claude-opus-4-6`, extended thinking), GPT-5.2 (`gpt-5.2`, high reasoning effort), Gemini 3.1 Pro Preview (`gemini-3.1-pro-preview`, thinking budget), and Grok 4.20 (`grok-4-0709`, always-on reasoning). All of them offer commercial APIs and libraries to support experimentation. No fine-tuning, few-shot examples, or task-specific adaptation was applied. The prompt instructs each model to read a conversation, classify it as predatory or non-predatory, and justify its classification. The prompt text, raw results, and analysis notebooks are available in the work repository[6].

3.3. Evaluation

For the classification task, we report accuracy, precision, recall, and F_1 for each model. Since all four models classify the same 78 conversations, the predictions are matched — a structure that Cochran's Q exploits to test whether the proportion of correct classifications differs across the four models simultaneously, and that McNemar exploits in pairwise follow-ups to identify which specific pairs differ. We therefore apply Cochran's Q as the omnibus test; if significant, we follow up with pairwise McNemar exact tests (exact variant, given $n = 39$). Both tests require per-instance binary outcomes, which maps

[6] `https://github.com/LaCAfe/BraSNAM2026-Predadores`

naturally onto recall on the predatory class: each predatory conversation is either a true positive (1) or a false negative (0), yielding a matched binary vector of length $n = 39$ per model. Inter-model agreement is quantified by Fleiss' κ.

To compare the LLM results with the traditional supervised ML baselines, we computed the exact Clopper-Pearson confidence intervals of 95% for each model. Because the supervised models were evaluated under 5-fold stratified cross-validation and their per-conversation predictions are not available, paired tests between training regimes cannot be applied; therefore, the comparison is descriptive, based on the overlap of confidence intervals, which is known to be conservative relative to a standard difference test [Schenker and Gentleman, 2001].

4. Results and Discussion

4.1. Classification Performance Across LLMs

Tab. 1 summarizes the zero-shot classification performance of the four LLMs on the PREDADORES-BR dataset. We first report standard classification metrics per model, then examine whether the observed differences in recall are statistically significant.

Tabela 1. Zero-shot classification results on PREDADORES-BR ($n = 78$; 39 predatory, 39 negative).

Model	Acc.	P	R	$\mathbf{F_1}$	κ	FP	FN
Claude Opus 4.6	96.15%	100%	92.31%	96.00%	0.923	0	3
Grok 4.20[†]	89.61%	100%	78.95%	88.24%	0.792	0	8
Gemini 3.1 Pro	87.18%	100%	74.36%	85.29%	0.744	0	10
GPT-5.2	80.77%	100%	61.54%	76.19%	0.615	0	15

[†] $n = 77$; 1 conversation blocked by CSAM safety filter.

All four models achieved 100% precision: every conversation flagged as predatory was indeed predatory. The models differed substantially in recall, which ranged from 61.54% (GPT-5.2) to 92.31% (Claude), yielding F_1 scores between 76.19% and 96.00%. Notably, all classification errors were false negatives; no model produced a false positive. Grok's safety filter blocked one conversation (conv_id=37); this instance was excluded from Grok's evaluation, yielding $n = 77$ for that model. Three predatory conversations (conv_ids 8, 21, 33) were misclassified by all four models. These conversations contain between 4 and 8 messages and lack explicit sexual content or age-related markers; all four models cited insufficient evidence of grooming indicators in their justifications.

To determine whether the observed differences in recall reflect genuine performance gaps rather than sampling variability, Cochran's Q test was applied to the predatory subset, yielding $Q = 19.40$, $df = 3$, $p < 0.001$. This rejects the null hypothesis that all four models have equal recall. Tab. 2 reports the subsequent pairwise McNemar tests.

The pairwise comparisons reveal that Claude significantly outperformed GPT-5.2 ($p < 0.001$) and Gemini ($p = 0.016$) while the difference between Claude and Grok was not significant at $\alpha = 0.05$ ($p = 0.063$). With regards to the remaining pairs, only Grok versus GPT-5.2 reached significance ($p = 0.016$); the other two pairs (Grok–Gemini and Gemini–GPT-5.2) did not differ significantly. It is worth noting that in all pairs involving Claude, $c = 0$, meaning that Claude's error set is a strict subset of every other model's.

Tabela 2. Pairwise McNemar exact tests on predatory conversations ($n = 38$).

	Claude	**Grok**	**Gemini**	**GPT-5.2**
Claude	36/39			
Grok	0/5 (.063)	30/38		
Gemini	0/7 (.016*)	2/3 (1.00)	29/39	
GPT-5.2	0/12 (<.001***)	0/7 (.016*)	3/8 (.227)	24/39

Diagonal: TP/n. Off-diagonal: b/c (p);
b = correct by row only, c = correct by column only.
*$p < .05$; ***$p < .001$.

One plausible interpretation is that these missed conversations exhibit little nuance in their communication patterns, which aligns with methodological decisions (e.g., setting a minimum number or percentage of messages) reported in the literature, and is a challenge inherent to the research domain rather than a model weakness.

Fleiss' κ on the full dataset ($n = 77$) was 0.79, indicating substantial agreement [Landis and Koch, 1977]. In contrast, when focusing solely on predatory conversations, κ dropped to 0.44 (moderate agreement). These results corroborate the finding that the four LLMs generally recognize non-predatory conversation patterns well; they differ considerably in their evaluations of which conversations are actually predatory.

4.2. LLMs versus Supervised Machine Learning

Tab. 3 places the zero-shot LLM results alongside the traditional supervised ML baselines. Since all LLMs achieved 100% precision, recall is the only metric that varies across models and admits meaningful comparison with the baseline setting. Since per-conversation predictions from the supervised models are not available, paired tests across training regimes cannot be applied; the comparison relies on confidence interval overlap, which is conservative relative to a standard difference test [Schenker and Gentleman, 2001].

Tabela 3. Supervised models (5-fold stratified CV) versus zero-shot LLMs.

	Model	**P**	**R**	**R 95% CI**	**F_1**
i-Sys 2020	SVM	90.10%	92.50%	[.79, .98]	89.87%
	MNB	87.32%	92.50%	[.79, .98]	88.98%
	CNN-RMSPROP	100%	76.42%	[.60, .89]	85.92%
	RF	91.28%	82.50%	[.67, .93]	85.77%
	CNN-SGD	86.76%	87.50%	[.73, .96]	86.36%
	DT	79.81%	58.93%	[.42, .74]	62.19%
LLMs	Claude	100%	92.31%	[.79, .98]	96.00%
	Grok	100%	78.95%	[.63, .90]	88.24%
	Gemini	100%	74.36%	[.58, .87]	85.29%
	GPT-5.2	100%	61.54%	[.45, .77]	76.19%

Recall 95% CIs are Clopper-Pearson exact intervals.

Two patterns stand out. First, regarding precision: all four LLMs achieved 100%, meaning every flagged conversation was indeed predatory. Among the supervised models, only CNN-RMSPROP matched this; the remaining models ranged from 79.81% to 91.28%, indicating that the supervised paradigm distributes errors across both false positives and false negatives, whereas the LLMs err exclusively toward false negatives. Under zero-shot conditions, reasoning-capable LLMs appear to adopt a conservative classification posture: they do not flag a conversation unless the textual evidence is unambiguous.

For deployment in a triage setting, this means every flagged conversation is actionable — at the cost of missing some predatory conversations.

Second, regarding recall: the confidence intervals reveal three tiers of comparability with the best supervised model (SVM, R = 92.50%, CI [.79, .98]). Claude's CI [.79, .98] fully overlaps with the SVM's, with point estimates differing by less than 0.2%. Grok and Gemini partially overlap, suggesting their recall may or may not match the SVM depending on the true underlying proportion. GPT-5.2's CI [.45, .77] does not overlap with the SVM's, indicating a recall deficit that persists even under the conservative overlap criterion.

Despite the methodological asymmetry between paradigms — traditional ML supervised models benefit from cross-validation with hyperparameter tuning, while the LLMs operate in a single pass with no task-specific adaptation — the central observation holds: the best-performing LLM (Claude, F_1 = 96.00%) achieved recall comparable to that of the best supervised model (SVM, F_1 = 89.87%) with no training data, no feature engineering, and zero false positives.

5. Conclusion

The current work evaluated four commercial LLMs with reasoning capabilities in zero-shot mode for the identification of sexual predators in Brazilian Portuguese conversations. To achieve this goal, we addressed binary classification of predatory conversations relative to traditional ML supervised baselines, using a balanced subset of 78 conversations from the PREDADORES-BR dataset.

The results show that zero-shot LLMs can reach classification performance comparable to supervised baselines trained on the same data. Claude Opus 4.6 achieved F_1 = 96.00% with 100% precision and recall statistically indistinguishable from the best supervised model (SVM, F_1 = 89.87%) All four models exhibited a conservative error regime, producing exclusively false negatives concentrated on short conversations with few grooming markers.

Some limitations should be acknowledged: (i) the PREDADORES-BR dataset contains only 39 predatory conversations, yielding wide confidence intervals that impact fine-grained differentiation between models; (ii) the LLM predictions were single-pass, leaving intra-model variance across repeated runs unquantified; (iii) the comparison across training regimes is descriptive rather than inferential, since per-conversation predictions from the supervised baselines are unavailable.

Future work should address these limitations along four directions:

1. **Output variability.** Repeating the experiments would quantify intra-model variance and yield tighter performance bounds, especially since deterministic reasoning was used but not checked via repeated sampling.
2. **Dataset expansion.** Expanding PREDADORES-BR with additional conversations from legal records or synthetic augmentation would strengthen statistical power and enable more robust cross-model comparisons.
3. **Task-specific adaptation.** Exploring few-shot prompting and parameter-efficient fine-tuning of open-weight models on Portuguese data would clarify whether task-specific adaptation can recover the recall lost by lower-performing models, particularly on the short, ambiguous conversations missed by all four models.

Referências

Borj, P. R., Raja, K., and Bours, P. (2023). Online grooming detection: A comprehensive survey of child exploitation in chat logs. *Knowledge-Based Systems*, 259:110039.

Comitê Gestor da Internet no Brasil (2025). *Pesquisa sobre o uso da Internet por crianças e adolescentes no Brasil: TIC Kids Online Brasil 2025*. CGI.br, São Paulo.

dos Santos, L. and Guedes, G. (2019). Identificação de predadores sexuais brasileiros por meio de análise de conversas realizadas na internet. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*, pages 143–154, Porto Alegre, RS, Brasil. SBC.

dos Santos, L. F. and Guedes, G. P. (2018). Detecção de traços de narcisismo em conversas com predadores sexuais. In *Anais do VII Brazilian Workshop on Social Network Analysis and Mining*, pages 217–222, Porto Alegre, RS, Brasil. SBC.

Ebrahimi, M., Suen, C. Y., and Ormandjieva, O. (2016). Detecting predatory conversations in social media by deep convolutional neural networks. *Digital Investigation*, 18:33–49.

Hamm, L. and McKeever, S. (2025). Comparing machine learning models with a focus on tone in grooming chat logs. *Frontiers in Pediatrics*, 13:1591828.

Inches, G. and Crestani, F. (2012). Overview of the international sexual predator identification competition at pan-2012. *CLEF (Online working notes/labs/workshop)*, 30.

Kloess, J. A., Hamilton-Giachritsis, C. E., and Beech, A. R. (2019). Offense processes of online sexual grooming and abuse of children via internet communication platforms. *Sexual Abuse*, 31(1):73–96.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774.

Milon-Flores, D. F. and Cordeiro, R. L. F. (2022). How to take advantage of behavioral features for the early detection of grooming in online conversations. *Knowledge-Based Systems*, 240:108017.

Nguyen, T. T., Wilson, C., and Dalins, J. (2024). Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. In *ESANN 2024 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 613–618, Bruges, Belgium.

Olson, L. N., Daggs, J. L., Ellevold, B. L., and Rogers, T. K. (2007). Entrapping the innocent: Toward a theory of child sexual predators' luring communication. *Communication Theory*, 17(3):231–251.

Panzariello, M. R. (2022). Estratégias para detecção precoce de predadores sexuais em conversas realizadas na internet. Dissertação de mestrado, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. *International Conference on Semantic Computing (ICSC 2007)*, 1:235–241.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Santos, L. F. and Guedes, G. P. (2020). Identificação de predadores sexuais brasileiros em conversas textuais na internet por meio de aprendizagem de máquina. *iSys: Revista Brasileira de Sistemas de Informação*, 13(2):26–53.

Schenker, N. and Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186.

Villatoro-Tello, E., Juárez-González, A., Escalante, H. J., Montes-y Gómez, M., and Pineda, L. V. (2012). A two-step approach for effective detection of misbehaving users in chats. In *CLEF (Online Working Notes/Labs/Workshop)*, volume 1178.