Prompt Engineering for Small Language Models: Evaluating ICL for Portuguese Sentiment Analysis

André da F. Schuck[1], Gabriel L. Garcia[1], João Renato R. Manesco[1], Pedro Henrique Paiola[1], Leandro A. Passos [1], João Paulo Papa[1]

[1]São Paulo State University (UNESP) / Av. Eng. Luís Edmundo Carrijo Coube, 14-01 – Bauru – SP – Brazil

{andre.schuck, gabriel.lino, joao.r.manesco, pedro.paiola, leandro.passos, joao.papa}@unesp.br

Abstract. *The In-Context Learning (ICL) paradigm enables adapting LLMs without parameter tuning. This work evaluates the impact of prompt engineering on 9B models for Brazilian Portuguese sentiment classification, comparing six prompting formats (few- and zero-shot) across three linguistically diverse models (Gemma2-9B-it, Boto-9B-IT, Qwen3.5-9B) and four public datasets, anchored by a fine-tuned encoder (MSA-DistilBERT, ∼0.1B) and a 685B MoE model (DeepSeek-V3.2) under a unified protocol. Results show that all 9B models outperform the encoder and recover 71–101% of the weak-to-strong gap, with Qwen3.5-9B achieving the highest mean coverage (94.5%). Statistical tests indicate that well-defined instructions are the main driver of ICL performance, capturing most of the achievable accuracy even in zero-shot settings, while elaborate formats (roleplay, JSON schema, meta-prompting) add no consistent gain over a minimal instruction-plus-demonstrations prompt, offering practical guidance for deploying SLMs in resource-constrained PT-BR NLP scenarios.*

1. Introduction

The In-Context Learning (ICL) paradigm [Brown et al. 2020] enables Large Language Models (LLMs) to adapt to new tasks via natural language alone: rather than updating parameters, the model is conditioned on a small set of instructions and/or demonstrations supplied in its input context [Dong et al. 2024]. Prompt engineering, defined as the systematic design and optimization of such inputs [Schulhoff et al. 2024, Reynolds and McDonell 2021, White et al. 2023], has consolidated a flexible paradigm for task execution across numerous applied scenarios [Liu et al. 2021b, Zhao et al. 2024], including sentiment analysis (SA) [Xu et al. 2024]. ICL remains, however, notoriously brittle, since its performance can shift substantially with prompt format [Sclar et al. 2024] and demonstration ordering [Lu et al. 2022], motivating rigorous empirical study, particularly for underrepresented languages.

SA concerns the automatic identification of subjective orientation in text, typically as positive, negative, or neutral [Liu 2012, Pang and Lee 2008], and underpins applications ranging from market intelligence to public-health surveillance [Birjali et al. 2021]. For Brazilian Portuguese (PT-BR), SA remains challenging: opinion texts exhibit informality, slang, and culturally-specific idioms, while labeled corpora are scarce when compared to English [Pereira 2021].

Despite the potential of ICL, systematic studies applying it to PT-BR remain scarce [Peres 2023, Assis et al. 2024], especially in SA [Marreira et al. 2025, Piorino et al. 2025, Lima et al. 2026].

This work addresses that gap by focusing on open-source Small Language Models (SLMs), defined here as Transformer models with at most 10 billion parameters [Oliveira et al. 2025], which can be deployed on modest local infrastructure while remaining competitive with massive proprietary LLMs [Lu et al. 2025, Nguyen et al. 2025, Wang et al. 2025]. Their availability raises a concrete deployment dilemma for PT-BR practitioners: under modest local infrastructure, should one fine-tune a small encoder, prompt an open 9B model, or pay for a frontier-scale API? Answering this requires more than leaderboard numbers, as it demands knowing how much performance depends on prompt design, how fragile each option is to formatting choices, and whether language-targeted adaptation helps or hurts.

Recent evidence suggests that SLMs are competitive for PT-BR SA [Lima et al. 2026], but also highly sensitive to prompt format and complexity [Schuck et al. 2025]; unlike those evaluations, which compare heterogeneous model families, we hold scale constant and isolate linguistic profile, including a controlled base→fine-tuned pair. We thus evaluate three open-source 9B SLMs: (i) Gemma2-9B-it [Gemma Team et al. 2024], representing English-dominant pretraining; (ii) Boto-9B-IT [Brígida 2024], a PT-BR-specialized LoRA fine-tune of Gemma2-9B-it; and (iii) Qwen3.5-9B [Qwen Team 2026], a natively multilingual model, anchored between a fine-tuned encoder floor (∼0.1B) and a 685B ceiling under a single unified protocol. The study is organized around three research questions:

- **RQ1:** Where do 9B models under ICL sit between the ∼0.1B encoder floor and the 685B ceiling, and does prompt sensitivity persist across scales?
- **RQ2:** Do instructions and demonstrations contribute distinctly to ICLperformance for PT-BR SA, and does combining them outperform either in isolation?
- **RQ3:** Do more elaborate combined formats yield statistically significant gains over the minimal one, or do they exhibit practical equivalence?

Our contributions are threefold: (i) a controlled PT-BR SA design that fixes scale at 9B and varies linguistic profile, including a base→fine-tuned pair (Gemma2/Boto) that isolates PT-BR adaptation at constant architecture and initialization, with the 685B baseline run under the same six prompts, demonstrations, and orderings so that prompt sensitivity, not only accuracy, becomes comparable across scales; (ii) evidence that 9B SLMs recover 71–101% of the weak-to-strong gap, with a natively multilingual model matching the 685B model's robustness; and (iii) evidence that well-defined instructions are the main driver of ICL performance, capturing most of the achievable accuracy even zero-shot, while demonstrations alone may degrade it [Min et al. 2022].

The remainder of this paper is organized as follows: Section 2 covers theoretical foundations; Section 3 details the experimental design; Section 4 reports results; Section 5 discusses findings; and Section 6 presents conclusions.

2. Background

Formally, ICL conditions the model's forward pass on a prompt p that concatenates input–output pairs (x_i, y_i), so prediction for a new x corresponds to the most probable continuation of $p \oplus x$ under the pre-training distribution f_θ [Brown et al. 2020, Wei et al. 2023, Dong et al. 2024]. As a research field, prompt engineering spans zero- and few-shot prompting [Brown et al. 2020], automated prompt generation [Shin et al. 2020, Zhou et al. 2023], demonstration selection [Liu et al. 2021a] and ordering [Lu et al. 2022], and reasoning-enhancing techniques [Wang et al. 2023].

3. Experimental Design

3.1. Tasks and Datasets

The proposed approaches were evaluated on two tasks, i.e., a binary sentiment classification (BSC: Positive/Negative) and a multiclass sentiment classification (MSC: Positive/Negative/Neutral) tasks, using four PT-BR datasets, detailed in Table 1. Datasets exceeding 1,500 test instances were subsampled via stratified random sampling with a fixed seed. Few-shot demonstrations were drawn exclusively from the training split; evaluation used the test split. Original class labels were mapped to numerical codes: -1 (Negative), 0 (Neutral), 1 (Positive).

Table 1. Datasets employed in the experimental evaluation. 🔴: `negative`; 🟡: `neutral`; and 🟢: `positive`. Distribution refers to the training split (source of demonstrations). Train and test splits share the same distribution.

Task	Dataset	Test Size	Class Distribution
BSC	IMDB_PT (IMDB)	1,500	🔴 50% ; 🟢 50%
	OPCovid-BR (OPC)	123	🔴 50% ; 🟢 50%
MSC	ReLI	1,500	🔴 4.82% ; 🟡 72.72% ; 🟢 22.46%
	TweetSentBr (TSBr)	1,500	🔴 33.3% ; 🟡 33.3% ; 🟢 33.3%

3.2. Models and Baselines

Table 2 summarizes the five evaluated systems, including three 9B-parameter SLMs to control for model size: Boto, a PT-BR-specialized variant of Gemma2 fine-tuned on the instruction-following dataset *cetacean-ptbr*[1], and Qwen3.5, which provides a multilingual contrast with support for 201 languages through native architectural design. The weak baseline (MSA-DistilBERT) represents the fine-tuning paradigm [Sanh et al. 2020, Devlin et al. 2018] and was used strictly off-the-shelf, with no dataset-specific adaptation: a multilingual encoder already fine-tuned for SA via supervised learning and evaluated without prompting. Its five-class output was consolidated by summing polarity extremes, i.e., $P(\text{very neg.}) + P(\text{neg.})$ and $P(\text{pos.}) + P(\text{very pos.})$, retaining Neutral for MSC and discarding it for BSC. Two asymmetries therefore constrain its interpretation: the model is tuned for the task but not for the target domains, and the BSC consolidation forces a polarity decision on instances whose modal class was Neutral. Both act in the same direction, so this

[1] `https://huggingface.co/datasets/lucianosb/cetacean-ptbr`

Table 2. Models and Baselines.

Role	Model Short	Parameters	Linguistic Profile
SLM	Gemma2[a]	9B	English-dominant
	Boto[b]	9B	PT-BR fine-tuned
	Qwen3.5[c]	9B	Multilingual (201 langs.)
Weak BL	MSA-DistilBERT[d]	0.1B	Multilingual (23 langs.)
Strong BL	DeepSeek[e]	685B	Multilingual

[a] `google/gemma-2-9b-it` [Gemma Team et al. 2024]
[b] `lucianosb/boto-9B-it` [Brígida 2024]
[c] `Qwen/Qwen3.5-9B` [Qwen Team 2026]
[d] `tabularisai/multilingual-sentiment-analysis` [TabularisAI et al. 2025]
[e] `deepseek-ai/DeepSeek-V3.2 (Non-thinking Mode)` [DeepSeek-AI 2025]

baseline should be read as a conservative floor and the gap-coverage percentages of Section 4.1 as correspondingly optimistic. The strong baseline (DeepSeek), in contrast, was evaluated under the same ICL protocol as the SLMs, ensuring comparability across tasks, datasets, prompt formats, demonstrations, and orderings.

3.3. Prompt Engineering

Table 3 presents the six evaluated prompt formats. *Demos-only* and the zero-shot *Inst-only* act as ablation anchors isolating the contributions of demonstrations and instructions (RQ2), and are competitive baselines in prior work [Ajith et al. 2023]; *Inst+Demos* is the simplest combined format, providing explicit classification guidance, against which three more elaborate variants are compared (RQ3): *JSON-Schema*, which enforces a predefined output schema for consistent parsing [Simmering and Huoviala 2023]; *Roleplay*, which assigns a domain-relevant persona [White et al. 2023, Zheng et al. 2024]; and *Meta-prompt*, an LLM-generated instruction (Claude 3.5 Sonnet, two refinement iterations) [Schulhoff et al. 2024].

All few-shot settings use 6 demonstrations (3 per class for BSC, 2 per class for MSC) in a "Q:{text}, A:{label}" format with English structural tokens and Portuguese content [Fu et al. 2022], selected via k-means clustering over *jina-embeddings-v3* representations to maximize semantic diversity [Sturua et al. 2024, Gao et al. 2023, Liu et al. 2021a], with three independent random orderings per configuration; full Jinja2 templates are available in the supplementary repository[2].

Table 3. Prompt format variants. All formats except *Inst-only* use 6 balanced demonstrations (3 per class for BSC, 2 per class for MSC) in the Q:A format. Full Jinja2 prompt templates are provided in the supplementary repository.

Variant	Components
Demos-only	Demonstrations only
Inst+Demos	Task instruction + demonstrations
JSON-Schema	Task instruction + JSON schema + demonstrations
Roleplay	Role specification + task instruction + demonstrations
Meta-prompt	Meta-prompt optimized instruction + demonstrations
Inst-only	Task instruction only (zero-shot)

[2] `https://github.com/AndreSchuck/slm_prompt_analysis`

3.4. Evaluation Methodology

Output parsing. Model outputs are parsed via format-specific regular expressions: for *JSON-Schema*, a single pattern targets the *sentimento* field via `re.search()`, tolerating common wrappers (e.g., Markdown fences, integer/float labels); for the remaining formats, a three-layer cascaded `re.match()` strategy in descending priority captures (i) a label at the start of the output, (ii) *"A resposta ... é {label}"*, and (iii) *"A classificação ... é {label}"*. Outputs unmatched by any layer are assigned code 2, denoting Format Non-Compliance (FNC). Full regex patterns and validation cases are available in the supplementary repository[2].

Metrics. Balanced Accuracy (BAcc), the mean of per-class Recall [Grandini et al. 2020], is the primary metric, essential given the 72.7% Neutral skew in ReLI. FNC outputs are retained rather than discarded and count as misclassifications, penalizing unparseable outputs identically to incorrect ones; their rate is reported separately as a diagnostic of output-structure adherence. Sensitivity to format is measured by the *Amplitude* (Δ), the BAcc range across the six prompt formats, and to ordering by the *Coefficient of Variation* (CV) [Brown 1998], σ/μ of BAcc across the three orderings, averaged over the five few-shot formats.

Prediction consolidation. Since the three iterations per configuration share the same inputs and differ only in the demonstration ordering, their predictions are not independent, and the consolidation is performed via majority voting [Dietterich 2000], yielding one prediction per instance that reflects the model's modal behavior. Tripartite ties, where each iteration predicts a different class (possible only in MSC), are assigned code 2 (FNC), following the same conservative principle applied to unparseable outputs.

Statistical testing. Pairwise differences between prompt formats are assessed with the paired bootstrap test [Koehn 2004] ($B = 10{,}000$ resamples, $\alpha = 0.05$, two-sided), following recommendations for NLP evaluation [Dror et al. 2018, Berg-Kirkpatrick et al. 2012], resampling instance-level prediction pairs from the majority-voted outputs. Comparisons target RQ2 (*Inst-only* vs. *Demos-only*, and *Inst+Demos* vs. each isolated format) and RQ3 (pairwise among combined formats). Beyond significance, differences are read against a practical-equivalence band of ± 0.02 *BAcc*, a margin no larger than the mean ordering-induced *BAcc* range observed under *Inst+Demos* (2.5 pp, Section 4.2), so that differences inside the band are of the same order as the variation the protocol itself introduces.

3.5. Implementation

SLMs were evaluated in BFloat16 precision with greedy decoding (`do_sample=False`) and `max_new_tokens=10` (20 for JSON-Schema), using Hugging Face pipelines[3] for both generation and classification tasks, while DeepSeek-V3.2 was accessed via its official API with `temperature=0` to approximate deterministic behavior; all

[3] `https://huggingface.co/docs/transformers/en/main_classes/pipelines`

models operated in standard (non-reasoning) modes. Reproducibility was ensured through fixed random seeds for dataset subsampling and demonstration selection ($s = 42$) and three consistent demonstration orderings ($s \in \{42, 2, 3\}$) applied across configurations. The evaluation comprised 337,479 inference calls spanning all models, datasets, prompt formats, and permutations, with local runs on NVIDIA A100 (40 GB) GPUs via Google Colab; code and full prompt templates are provided in the supplementary repository[2].

4. Experimental Results

4.1. RQ1 – ICL performance across model scales

SLMs surpass the fine-tuned encoder and approach the 685B ceiling. Under their best prompt, all three 9B models exceed MSA-DistilBERT on every dataset by +0.12 to +0.20 *BAcc*, covering 71–101% of the weak-to-strong gap, with 9 of 12 model–dataset cells above 85% (Table 4). Qwen3.5 attains the highest mean coverage (94.5%) and is the only 9B model to reach the 685B ceiling on any dataset (ReLI, 101%). The lower bound falls on the smallest test set (Gemma2 on OPC, N=123), where a single instance shifts gap coverage by roughly 5 pp.

Prompt sensitivity persists across scales but varies by model. Sensitivity does not vanish at 685B: DeepSeek's best and worst prompts still differ by 8.5 pp *BAcc* on average, and by 12.6 pp on TSBr. What scale does not predict is its magnitude. Qwen3.5 matches DeepSeek on both dimensions (mean Δ 0.090 vs. 0.085; CV 0.022 vs. 0.019), while Boto and Gemma2, identical in architecture and initialization, differ by 1.8× in format sensitivity and 3.2× in ordering sensitivity, with Boto's extreme case on TSBr ($\Delta = 0.614$) coinciding with 18.1% FNC. The two dimensions rank consistently together across the 16 model–dataset cells (Spearman $\rho = 0.76$, $p < 0.01$). Sensitivity is thus tied less to scale than to the model itself, motivating the analysis of prompt components in RQ2.

4.2. RQ2 – Disentangling instruction and demonstration contributions

Instructions outperform demonstrations in isolation on most datasets. *Inst-only* achieves significantly higher *BAcc* than *Demos-only* in 7 of 12 model–dataset comparisons (Table 5, top; Figure 1a), with a mean $\Delta BAcc$ of +0.122. The effect is largest for Boto (+0.213), driven by a 27.0 pp FNC reduction, and smallest for Qwen3.5 (+0.040), whose negligible FNC difference (+0.6 pp) indicates a gain in classification rather than in format adherence; Gemma2 lies in between (12.9 pp). ReLI (72.7% Neutral) is the consistent exception: *Demos-only* numerically outperforms *Inst-only* for all three models (significantly for Gemma2 and Qwen3.5), and adding instructions to demonstrations yields no gain.

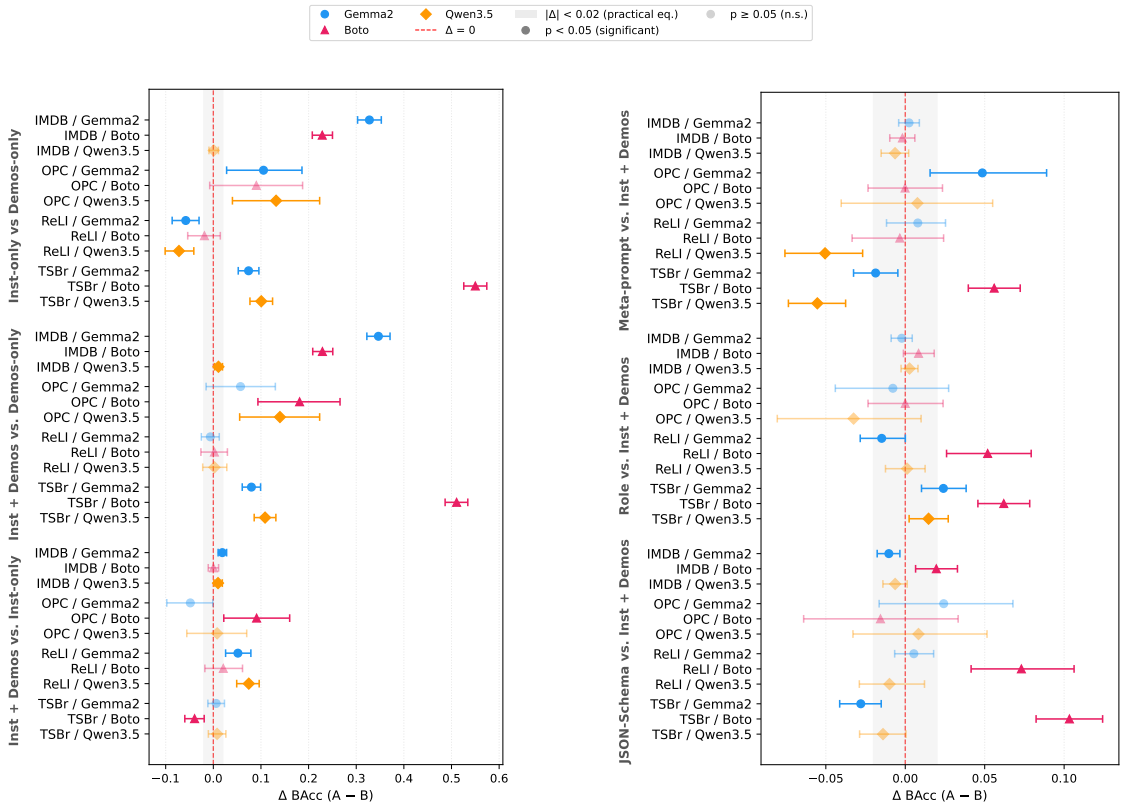
Demonstrations add limited value once instructions are present. *Inst+Demos* significantly outperforms *Demos-only* (mean $\Delta = +0.138$, 8 ↑), but the advantage over *Inst-only* is only +0.017 (5 ↑ 1 ↓), with virtually unchanged FNC rates (+0.3 pp). The forest plot (Figure 1a) confirms that most confidence intervals are narrow and

Table 4. Performance and sensitivity by model × dataset. *Best BAcc*: highest across the six prompts (majority-voted). *Gap*: percentage of the weak-to-strong gap covered, $(\text{BAcc} - \text{MSA})/(\text{DeepSeek} - \text{MSA}) \times 100$, computed from unrounded values; the two baselines anchor the scale at 0 and 100 by construction. Δ: max–min BAcc across prompts (format sensitivity). *CV*: σ/μ of BAcc across the three orderings, averaged over the five few-shot formats, with the mean BAcc range (Δpp) in parentheses. *FNC %*: mean across formats. Bold/underline: best/second-best per dataset (the Gap ranking is identical by construction).

				Format	Ordering	
Model	Dataset	Best BAcc	Gap %	Δ	CV (Δpp)	FNC %
MSA-DistilBERT	IMDB (BSC)	0.755	0		—	
	OPC (BSC)	0.652			—	
	ReLI (MSC)	0.572			—	
	TSBr (MSC)	0.507			—	
Gemma2	IMDB (BSC)	0.941	95	0.349	0.030 (3.8)	7.5
	OPC (BSC)	0.773	71	0.106	0.040 (5.2)	2.8
	ReLI (MSC)	0.711	79	0.060	0.012 (1.7)	1.3
	TSBr (MSC)	<u>0.710</u>	96	0.104	0.028 (3.4)	0.3
	Mean	—	*85.3*	*0.155*	*0.027 (3.5)*	*3.0*
Boto	IMDB (BSC)	0.938	94	0.249	0.051 (7.6)	2.8
	OPC (BSC)	<u>0.813</u>	95	0.181	0.038 (5.2)	0.4
	ReLI (MSC)	0.730	90	0.094	0.051 (6.4)	2.7
	TSBr (MSC)	0.679	81	0.614	0.208 (6.9)	18.1
	Mean	—	*90.0*	*0.284*	*0.087 (6.6)*	*6.0*
Qwen3.5	IMDB (BSC)	<u>0.948</u>	99	0.013	0.006 (1.1)	0.9
	OPC (BSC)	0.797	86	0.148	0.032 (4.5)	2.3
	ReLI (MSC)	**0.749**	101	0.076	0.019 (2.5)	0.1
	TSBr (MSC)	0.703	92	0.123	0.033 (3.9)	1.4
	Mean	—	*94.5*	*0.090*	*0.022 (3.0)*	*1.2*
DeepSeek	IMDB (BSC)	**0.951**	100	0.029	0.005 (0.9)	0.2
	OPC (BSC)	**0.822**		0.113	0.023 (3.2)	0.5
	ReLI (MSC)	<u>0.747</u>		0.072	0.020 (2.8)	0.4
	TSBr (MSC)	**0.719**		0.126	0.028 (3.5)	0.7
	Mean	—	—	*0.085*	*0.019 (2.6)*	*0.5*

centered near zero. In practice, zero-shot instructions capture most of the combined format's performance on balanced or moderately skewed datasets, whereas demonstration-only prompts risk substantial degradation.

Instructions also stabilize predictions against ordering perturbations. Across the three orderings, the mean *BAcc* range drops from 9.6 pp (CV = 0.122) under *Demos-only* to 2.5 pp (CV = 0.018) under *Inst+Demos*, a 3.8× narrowing of the range that holds for every model, DeepSeek at 685B included (4.4×); the stabilizing role of instructions is therefore not scale-dependent (cf. Section 4.1). Gemma2 on IMDB is the clearest case: *Demos-only* yields 0.522, 0.586 and 0.678 across orderings (15.6 pp), converging to 0.938–0.940 (0.2 pp) once instructions are added. Per-ordering trajectories for all six formats, disaggregated by model × dataset,



(a) Decomposing instruction and demonstration contributions: isolated formats and their combination.

(b) Elaborated vs. minimal combined formats (practical-equivalence band).

Figure 1. Paired Bootstrap Test Results ($B = 10,000$, $\alpha = 0.05$). Predictions consolidated via majority voting across 3 demonstration orderings.

are provided in the supplementary repository[4].

4.3. RQ3 – Elaborated vs. minimal combined formats

No elaborated format consistently outperforms the simplest combined baseline. Six of the nine model × format mean differences fall inside the ±0.02 practical-equivalence band (Table 5, bottom; Figure 1b), and the significant comparisons that do occur alternate in sign: Gemma2 gains and loses in equal measure under *Meta-prompt* and *Roleplay*, and is significantly worse under *JSON-Schema* (2 ↓), while Qwen3.5 shows no net change under *Roleplay* or *JSON-Schema* (−0.004, −0.005) and degrades under *Meta-prompt* (−0.026, 2 ↓). Elaboration is thus not a reliable route to accuracy once a task instruction is present.

Boto is the only model that benefits. Elaboration helps precisely the model with the weakest baseline compliance (Section 4.1): *JSON-Schema* (+0.045, 3 ↑), *Roleplay* (+0.031, 2 ↑) and *Meta-prompt* (+0.013, 1 ↑). All three cut Boto's FNC

[4]Direct link: `https://github.com/AndreSchuck/slm_prompt_analysis/blob/master/Article/Images/rq2_ordering_sensitivity_6fmt.png`

Table 5. Summary of paired bootstrap tests ($B = 10{,}000$, $\alpha = 0.05$): each cell reports the mean $\Delta BAcc$ and the number of significant comparisons (↑/↓ favoring the first/second system, – none), out of 4 per model and 12 overall. Top: RQ2; bottom: RQ3.

Comparison	Gemma2	Boto	Qwen3.5	All
Inst-only vs. Demos-only	+0.112 (3 ↑ 1 ↓)	+0.213 (2 ↑)	+0.040 (2 ↑ 1 ↓)	+0.122 (7 ↑ 2 ↓)
Inst+Demos vs. Demos-only	+0.119 (2 ↑)	+0.231 (3 ↑)	+0.065 (3 ↑)	+0.138 (8 ↑)
Inst+Demos vs. Inst-only	+0.007 (2 ↑)	+0.018 (1 ↑ 1 ↓)	+0.025 (2 ↑)	+0.017 (5 ↑ 1 ↓)
Meta vs. Inst+Demos	+0.010 (1 ↑ 1 ↓)	+0.013 (1 ↑)	−0.026 (2 ↓)	−0.001 (2 ↑ 3 ↓)
Role vs. Inst+Demos	−0.000 (1 ↑ 1 ↓)	+0.031 (2 ↑)	−0.004 (1 ↑)	+0.009 (4 ↑ 1 ↓)
JSON vs. Inst+Demos	−0.002 (2 ↓)	+0.045 (3 ↑)	−0.005 (–)	+0.013 (3 ↑ 2 ↓)

rate by a similar margin (≈3.8–4.0 pp) despite differing in whether they constrain output structure, so the benefit tracks instructional specificity rather than format constraints per se.

5. Discussion

The findings across RQ1–RQ3 converge on a coherent narrative: effective ICL-based SA in PT-BR with 9B instruction-tuned SLMs depends less on prompt sophistication than on the presence of a well-formed instruction.

5.1. Scale vs. Performance: 9B SLM Competitiveness

The RQ1 results show that 9B SLMs under ICL can substantially reduce the performance gap to a 685B state-of-the-art model, reinforcing recent findings for PT-BR sentiment classification [Schuck et al. 2025, Lima et al. 2026]. This competitiveness, however, coexists with the known volatility of ICL, where performance depends strongly on prompt design and demonstration ordering [Sclar et al. 2024, Schoch and Ji 2025]. Within this context, results indicate that prompt sensitivity at the 9B scale can match that of 685B models, as Qwen3.5 achieves similar robustness to DeepSeek, possibly reflecting multilingual pre-training and more recent model generation. This suggests that sensitivity is driven more by training regime than scale, implying that, in resource-constrained settings, well-chosen 9B models with effective prompting can recover most of the performance of large proprietary LLMs.

5.2. Instructions as the Primary Driver of ICL Performance

The convergence of RQ2 and RQ3 highlights a hierarchy of prompt components in which instructions dominate over demonstrations, and additional prompt elaboration yields minimal gains. On balanced or moderately skewed datasets, zero-shot *Inst-only* captures most of the achievable performance, while removing instructions (*Demos-only*) significantly harms accuracy and output compliance. This aligns with prior work showing that instruction-tuning enables generalizable instruction-following [Wei et al. 2022], whereas demonstrations mainly convey label space, input distribution, and format [Min et al. 2022]; since these are already encoded in the instructions, their contribution is marginal, explaining the small improvement from *Inst-only* to *Inst+Demos* (mean $\Delta BAcc = +0.017$) and negligible FNC reduction.

5.3. Targeted Fine-tuning and Increased Prompt Sensitivity: A Hypothesis

A counterintuitive result arises in the sensitivity analysis: Boto, the only PT-BR–specific fine-tuned model, shows the highest sensitivity in both format and ordering, whereas Qwen3.5, a natively multilingual model of the same scale, matches the stability of the 685B baseline. One plausible account is that LoRA adaptation to a narrow formatting distribution (*cetacean-ptbr*) weakened the base model's instruction-following behavior and its generalization to unseen prompt structures. This reading rests on a single base→fine-tuned pair, however, and should be taken as a hypothesis rather than a demonstrated causal effect: catastrophic forgetting of instruction-following [Luo et al. 2025], interactions between the fine-tuning chat template and our prompt formats, and LoRA configuration choices [Biderman et al. 2024] are equally consistent with the observed pattern.

6. Conclusion

This work evaluated six prompt formats across three linguistically diverse 9B SLMs and four PT-BR sentiment datasets, anchored by a fine-tuned encoder (0.1B) and a 685B MoE model under a unified protocol. Three findings emerged: (i) 9B SLMs recover 71–101% of the weak-to-strong gap under optimal prompting, with Qwen3.5 matching the 685B baseline in robustness; (ii) instructions, not demonstrations, drive ICL performance on balanced and moderately skewed datasets, with zero-shot *Inst-only* capturing most of the achievable accuracy and instructions narrowing the ordering-induced *BAcc* range by 3.8× (*Demos-only* → *Inst+Demos*), while *Demos-only* degrades both accuracy and output compliance; and (iii) elaborate formats (*Roleplay*, *JSON-Schema*, *Meta-prompt*) provide no consistent gain over *Inst+Demos*. The PT-BR fine-tuned Boto showed the highest sensitivity, suggesting that narrow-distribution LoRA adaptation may reduce robustness. Future work should extend this analysis to other NLP tasks (e.g., QA, NER, NLI), broader scales (1–3B, 20–70B), and additional languages, and examine the interaction between demonstration selection and prompt format in highly skewed settings such as ReLI.

Limitations

This study has several limitations that inform result interpretation and future work. The evaluation is restricted to SA, the 9B parameter range, and four PT-BR datasets, limiting generalization across tasks, scales, and languages [Peres 2023, Alves et al. 2023, Geng et al. 2024]; exploring smaller (1–3B) and larger (20–70B) models could clarify whether prompt sensitivity decreases with scale. The cross-scale comparison in RQ1 is exploratory, as models differ in architecture (MoE vs. dense), execution (API vs. local BFloat16), and reproducibility, so reported differences in amplitude and CV are indicative rather than causal. The high accuracy on IMDB-PT may reflect cross-lingual contamination, since it derives from a widely used English benchmark [Maas et al. 2011].

Methodological trade-offs include discarding Neutral probabilities in MSA-DistilBERT, a conservative bias in majority-voting ties toward FNC, and a fixed `max_new_tokens` that can penalize verbose formats. Few-shot results also rest on a single demonstration-selection strategy (k-means over *jina-embeddings-v3*), which maximizes semantic diversity but presumes access to a labeled training split and to an embedding model, a non-trivial requirement in production settings. Since selection is known to interact with ordering sensitivity [Liu et al. 2021a, Lu et al. 2022], and RQ2 indicates that demonstrations contribute marginally once instructions are present, we expect this choice to have limited impact here, although a random-selection control remains necessary to confirm it.

Ethics Statement

This work uses only publicly available datasets (IMDB-PT, OPCovid-BR, ReLI, TweetSentBr) and pre-trained models accessed under their respective licenses. No human-subject data was collected. Automated SA systems may amplify linguistic biases in training data when deployed in high-stakes contexts; our findings are presented strictly as a contribution to the scientific understanding of prompt engineering for PT-BR NLP.

Generative AI Usage

Generative AI was used in two capacities: Claude 3.5 Sonnet (Anthropic) produced the instruction component of the Meta-prompt format (one of six evaluated configurations); LLM tools also assisted research, manuscript drafting and revision. All scientific claims and conclusions are the sole responsibility of the authors.

Acknowledgments

This study was partially funded by the São Paulo Research Foundation (FAPESP), Brazil, under Grant Nos. #2025/13172-1, #2024/22853-0, and #2023/14427-8. This work was also supported by Petrobras through research grants No. #2025/00607-0 and No. #2023/00466-1.

References

Ajith, A., Pan, C., Xia, M., et al. (2023). Instructeval: Systematic evaluation of instruction selection methods. https://arxiv.org/abs/2307.00259.

Alves, D. M., Guerreiro, N. M., Alves, J., et al. (2023). Steering large language models for machine translation with finetuning and in-context learning. https://arxiv.org/abs/2310.13448.

Assis, G., Amorim, A., Carvalho, J., et al. (2024). Exploring Portuguese hate speech detection in low-resource settings: Lightly tuning encoder models or in-context learning of large models? In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 301–311, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Berg-Kirkpatrick, T., Burkett, D., and Klein, D. (2012). An empirical investigation of statistical significance in NLP. In Tsujii, J., Henderson, J., and Paşca, M., editors, *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Biderman, D., Portes, J., Ortiz, J. J. G., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., and Cunningham, J. P. (2024). Lora learns less and forgets less.

Birjali, M., Kasri, M., and Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Brígida, L. S. (2024). Boto 9b it. https://huggingface.co/lucianosb/boto-9B-it.

Brown, C. E. (1998). Coefficient of variation. In *Applied multivariate statistics in geohydrology and related sciences*, pages 155–157. Springer.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

DeepSeek-AI (2025). Deepseek-v3.2: Pushing the frontier of open large language models.

Devlin, J., Chang, M.-W., Lee, K., et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dong, Q., Li, L., Dai, D., et al. (2024). A survey on in-context learning. https://arxiv.org/abs/2301.00234.

Dror, R., Baumer, G., Shlomov, S., et al. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Fu, J., Ng, S.-K., and Liu, P. (2022). Polyglot prompt: Multilingual multitask promptraining. https://arxiv.org/abs/2204.14264.

Gao, S., Wen, X.-C., Gao, C., et al. (2023). What makes good in-context demonstrations for code intelligence tasks with llms? In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE.

Gemma Team, Riviere, M., Pathak, S., et al. (2024). Gemma 2: Improving open language models at a practical size. https://doi.org/10.48550/arXiv.2408.00118.

Geng, M., Wang, S., Dong, D., et al. (2024). Large language models are few-shot summarizers: Multi-intent comment generation via in-context learning. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, ICSE '24, New York, NY, USA. Association for Computing Machinery.

Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. https://arxiv.org/abs/2008.05756.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Lima, J. V. R. J., Pinheiro, V., and Caminha, C. (2026). Portuguese sentiment analysis with open-source LLMs: Models, prompts, and efficient deployment. In Souza, M., de Dios-Flores, I., Santos, D., Freitas, L., Souza, J. W. d. C., and Ribeiro, E., editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 212–221, Salvador, Brazil. Association for Computational Linguistics.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Springer International Publishing.

Liu, J., Shen, D., Zhang, Y., et al. (2021a). What makes good in-context examples for gpt-3? https://arxiv.org/abs/2101.06804.

Liu, P., Yuan, W., Fu, J., et al. (2021b). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. https://doi.org/10.48550/arXiv.2107.13586.

Lu, Y., Bartolo, M., Moore, A., et al. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. https://arxiv.org/abs/2104.08786.

Lu, Z., Li, X., Cai, D., et al. (2025). Small language models: Survey, measurements, and insights. https://arxiv.org/abs/2409.15790.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. (2025). An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

Maas, A. L., Daly, R. E., Pham, P. T., et al. (2011). Learning word vectors for sentiment analysis. https://aclanthology.org/P11-1015.

Marreira, E., De Melo, T., De Oliveira, M., et al. (2025). Rating prediction in brazilian portuguese: A benchmark of large language models. *Journal of the Brazilian Computer Society*, 31:828–839.

Min, S., Lyu, X., Holtzman, A., et al. (2022). Rethinking the role of demonstrations: What makes in-context learning work? https://arxiv.org/abs/2202.12837.

Nguyen, C. V., Shen, X., Aponte, R., et al. (2025). A survey on small language models. In Angelova, G., Kunilovskaya, M., Escribe, M., and Mitkov, R., editors, *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 807–821, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Oliveira, A., Nascimento, E., Pinheiro, J., et al. (2025). Small, medium, and large language models for text-to-sql. In Maass, W., Han, H., Yasar, H., and Multari, N., editors, *Conceptual Modeling*, pages 276–294, Cham. Springer Nature Switzerland.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135.

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artif. Intell. Rev.*, 54(2):1087–1115.

Peres, R. S. (2023). Grandes modelos de linguagem na resolução de questões de vestibular: o caso dos institutos militares brasileiros. Master's thesis, Universidade Federal do Estado do Rio de Janeiro.

Piorino, G., Moreira, V., Lima, L. H. Q., et al. (2025). Sentiment analysis of shared content in brazilian reddit communities. *Journal on Interactive Systems*, 16:666–686.

Qwen Team (2026). Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5.

Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. https://arxiv.org/abs/2102.07350.

Sanh, V., Debut, L., Chaumond, J., et al. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. https://arxiv.org/abs/1910.01108.

Schoch, S. and Ji, Y. (2025). The good, the bad, and the debatable: A survey on the impacts of data for in-context learning. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29798–29812, Suzhou, China. Association for Computational Linguistics.

Schuck, A. d. F., Garcia, G. L., Manesco, J. R. R., et al. (2025). Evaluating large language models for brazilian portuguese sentiment analysis: A comparative study

of multilingual state-of-the-art vs. brazilian portuguese fine-tuned llms. *Journal of the Brazilian Computer Society*, 31(1):885–917.

Schulhoff, S., Ilie, M., Balepur, N., et al. (2024). The prompt report: A systematic survey of prompting techniques. https://arxiv.org/abs/2406.06608.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2024). Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.

Shin, T., Razeghi, Y., au2, R. L. L. I., et al. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. https://arxiv.org/abs/2010.15980.

Simmering, P. F. and Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. https://arxiv.org/abs/2310.18025.

Sturua, S., Mohr, I., Akram, M. K., et al. (2024). jina-embeddings-v3: Multilingual embeddings with task lora. https://arxiv.org/abs/2409.10173.

TabularisAI, Gyamfi, S., Borisov, V., et al. (2025). multilingual-sentiment-analysis (revision 69afb83). https://huggingface.co/tabularisai/multilingual-sentiment-analysis.

Wang, F., Lin, M., Ma, Y., et al. (2025). A survey on small language models in the era of large language models: Architecture, capabilities, and trustworthiness. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pages 6173–6183, New York, NY, USA. Association for Computing Machinery.

Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-consistency improves chain of thought reasoning in language models. https://arxiv.org/abs/2203.11171.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. (2022). Finetuned language models are zero-shot learners.

Wei, J., Wei, J., Tay, Y., et al. (2023). Larger language models do in-context learning differently. https://arxiv.org/abs/2303.03846.

White, J., Fu, Q., Hays, S., et al. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. https://arxiv.org/abs/2302.11382.

Xu, H., Wang, Q., Zhang, Y., Yang, M., Zeng, X., Qin, B., and Xu, R. (2024). Improving in-context learning with prediction feedback for sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3879–3890.

Zhao, W. X., Zhou, K., Li, J., et al. (2024). A survey of large language models. https://arxiv.org/abs/2303.18223.

Zheng, M., Pei, J., Logeswaran, L., et al. (2024). When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. https://arxiv.org/abs/2311.10054.

Zhou, Y., Muresanu, A. I., Han, Z., et al. (2023). Large language models are human-level prompt engineers. https://arxiv.org/abs/2211.01910.