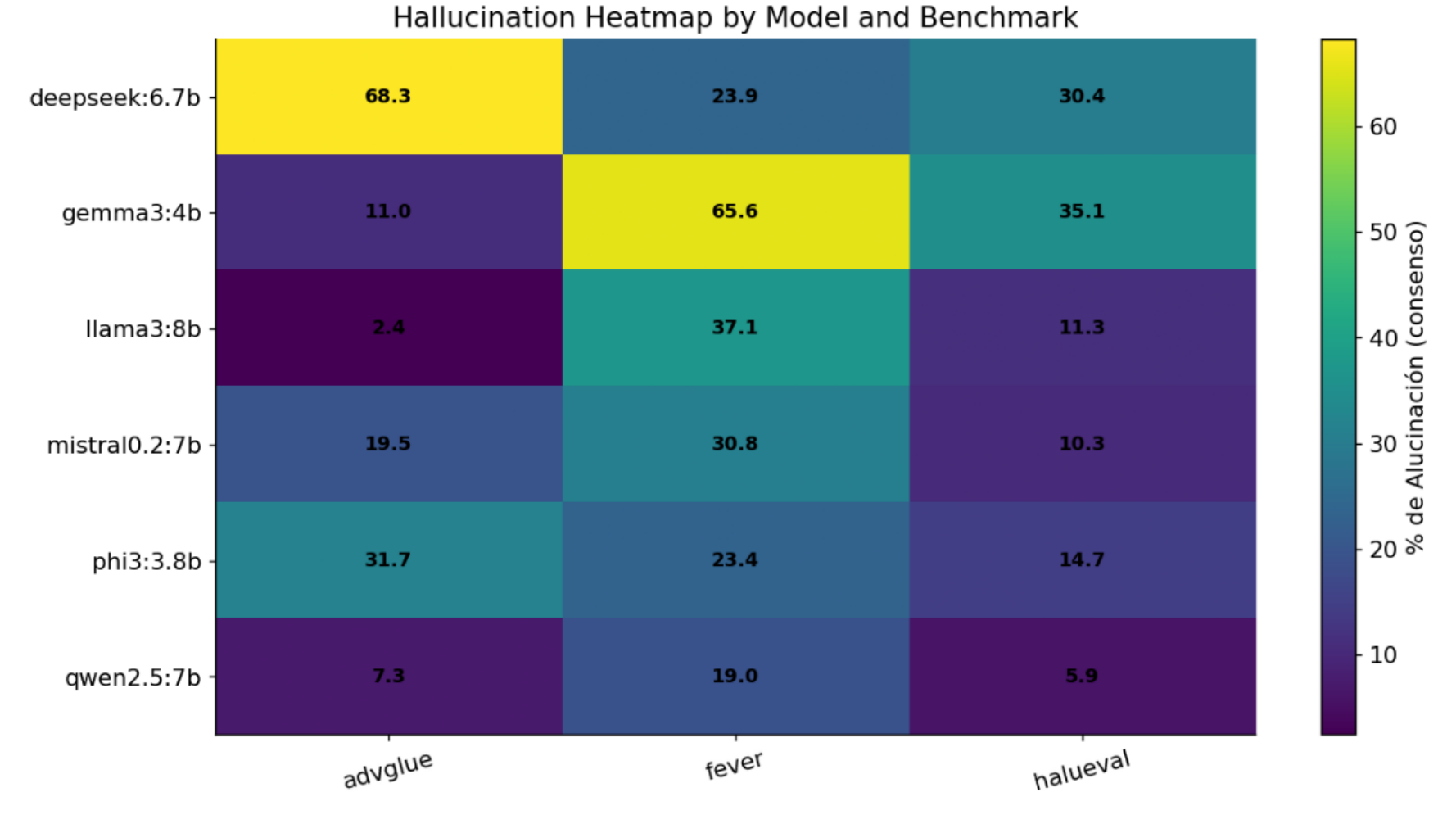
Reporte de Visualizaciones – 06_

Generado: 2025-08-27 10:55:14

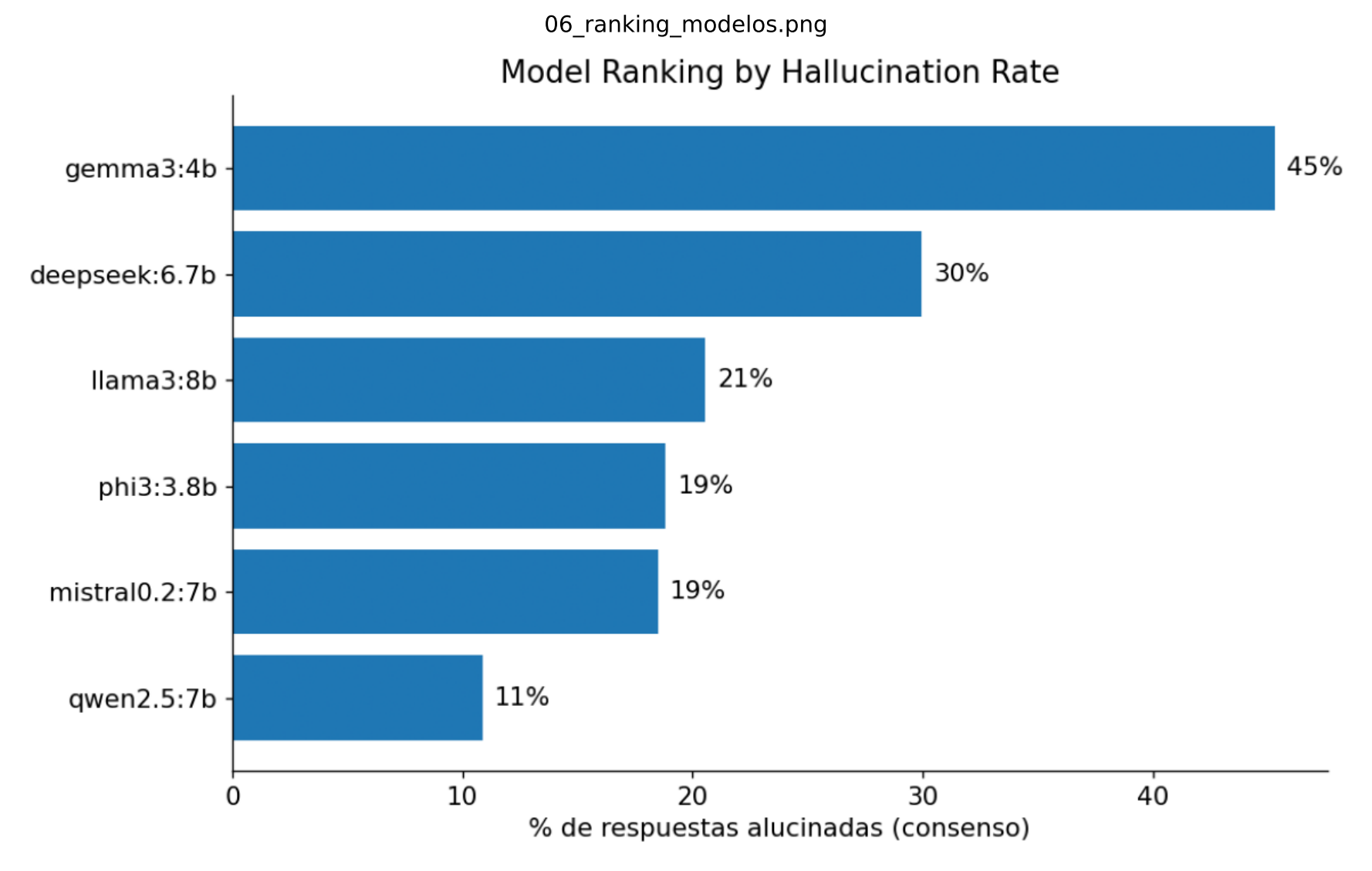
Carpeta de salida: /home/dslab/teilor/newtron/06_visualizaciones

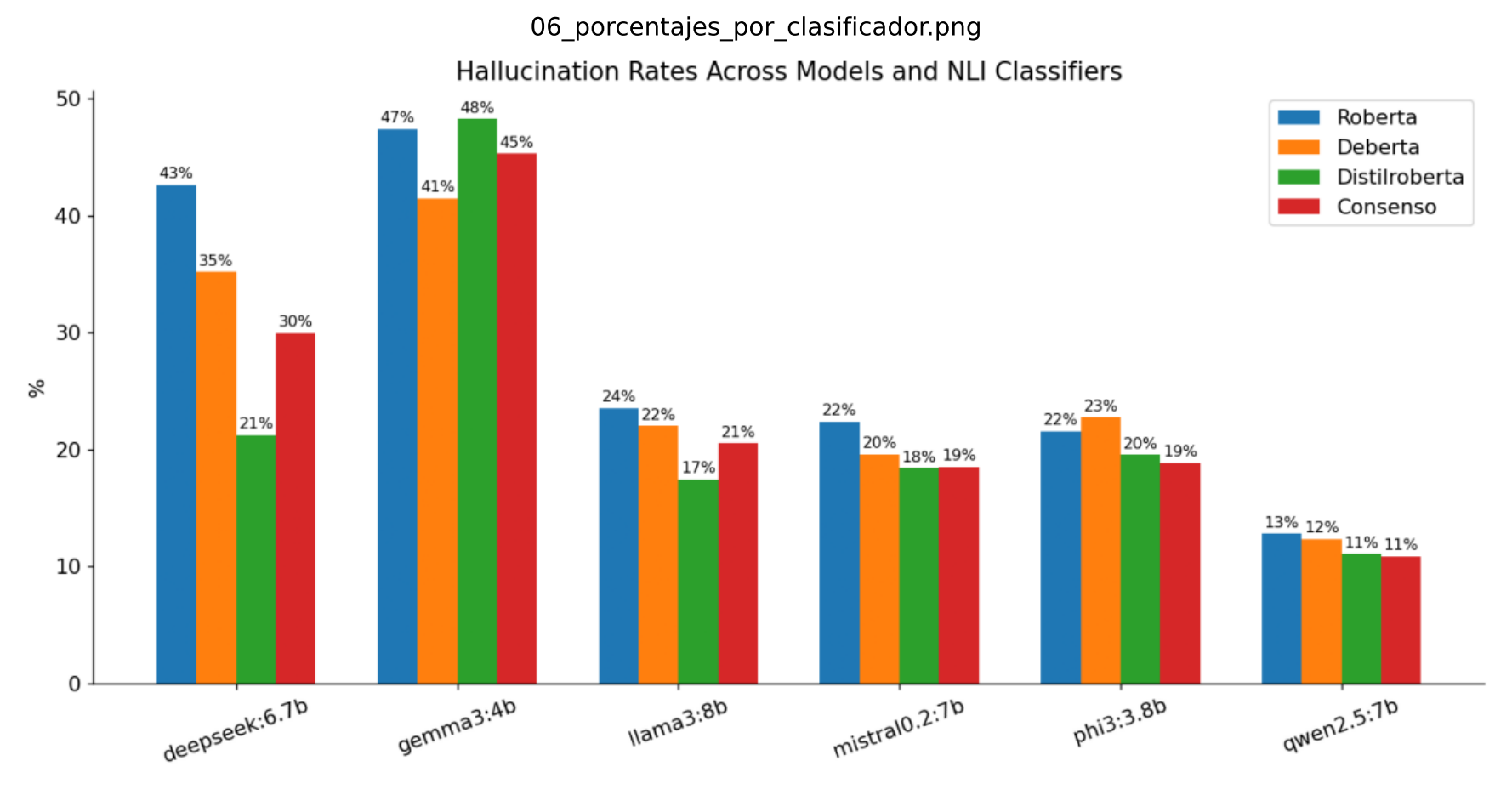
Nota: porcentajes ⇒ % de respuestas alucinadas (consenso). Δ en Mann–Whitney: μ(aluc.) − μ(no aluc.)

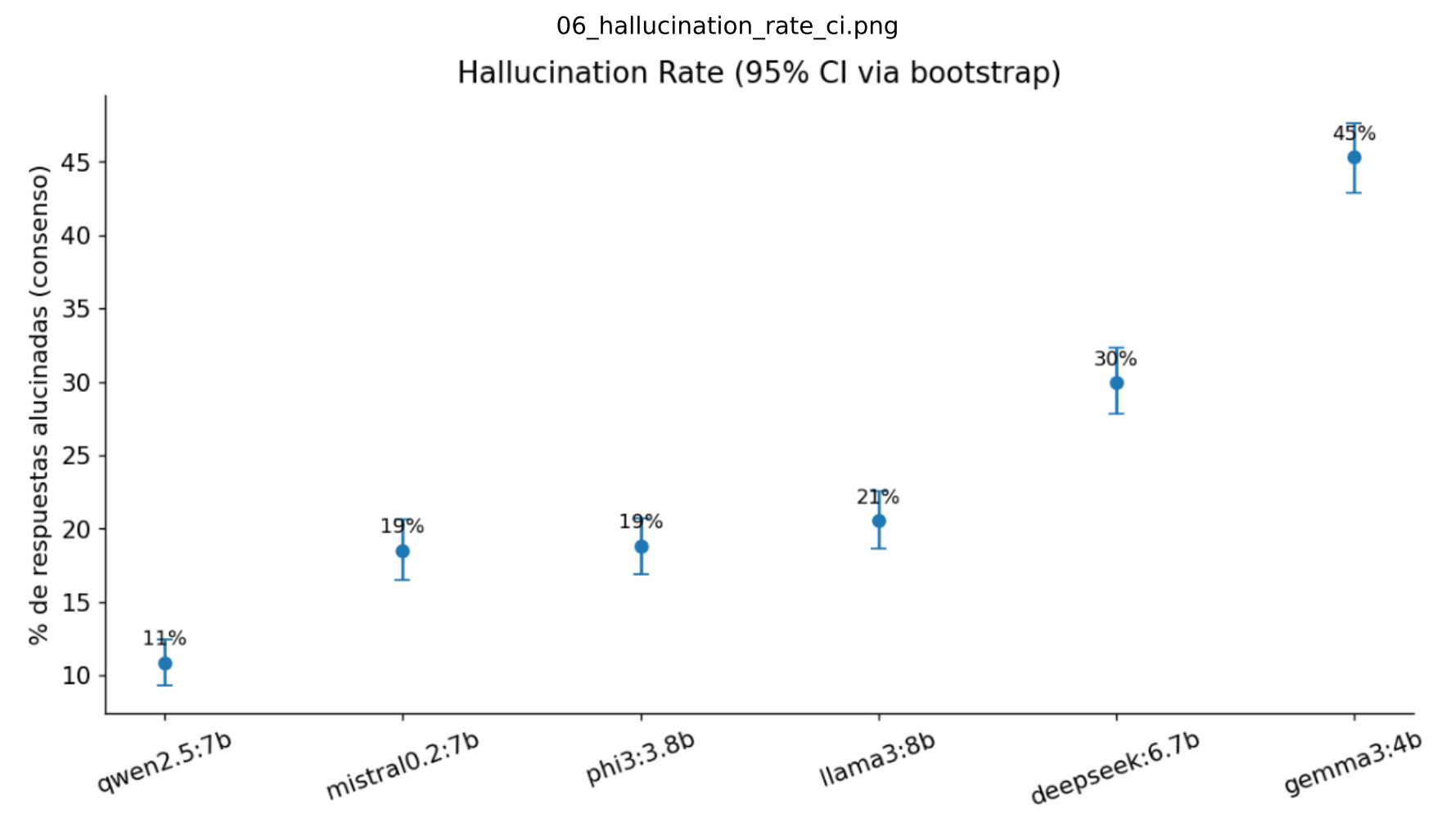
06_heatmap_modelo_benchmark.png

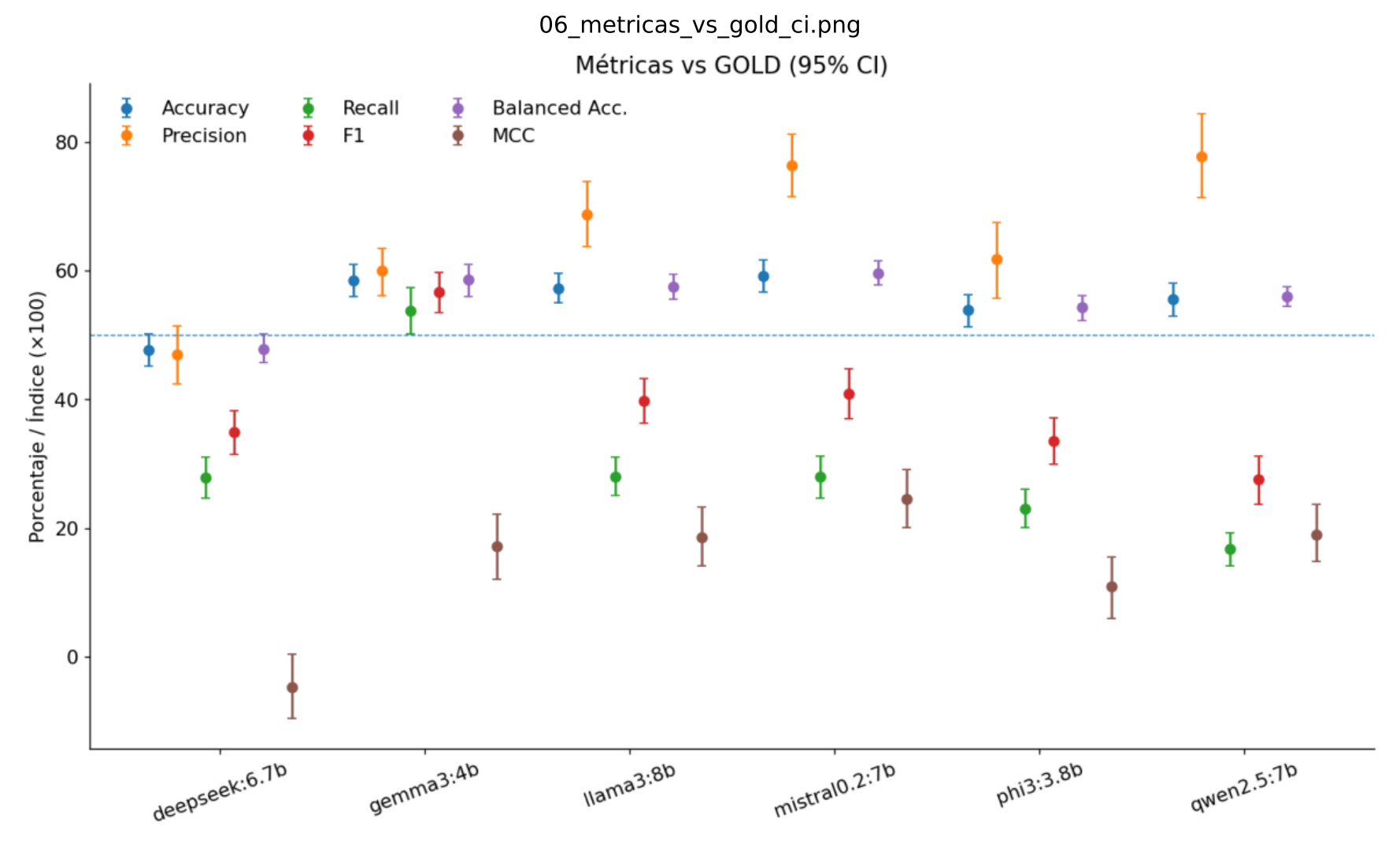


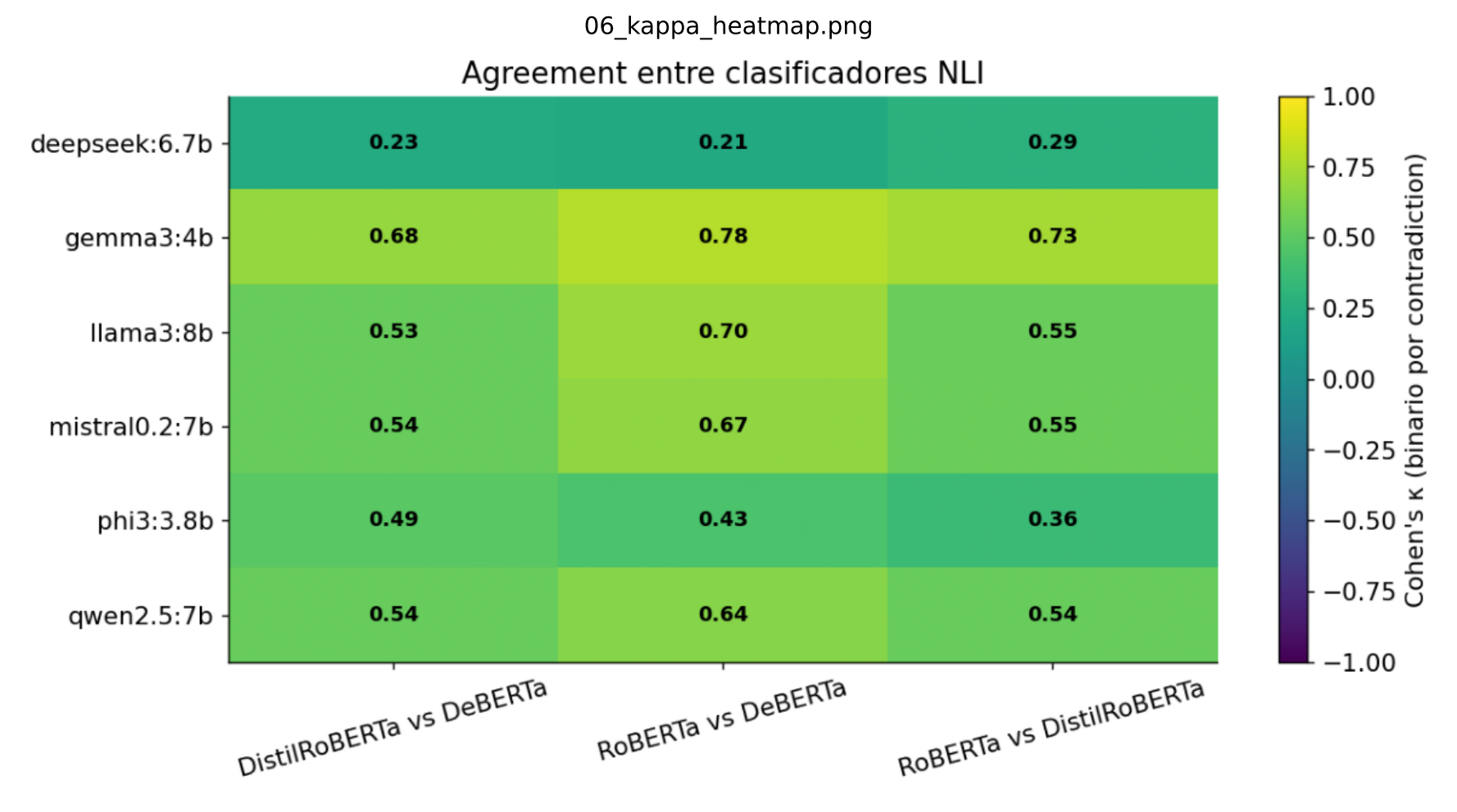
06_ranking_modelos.png



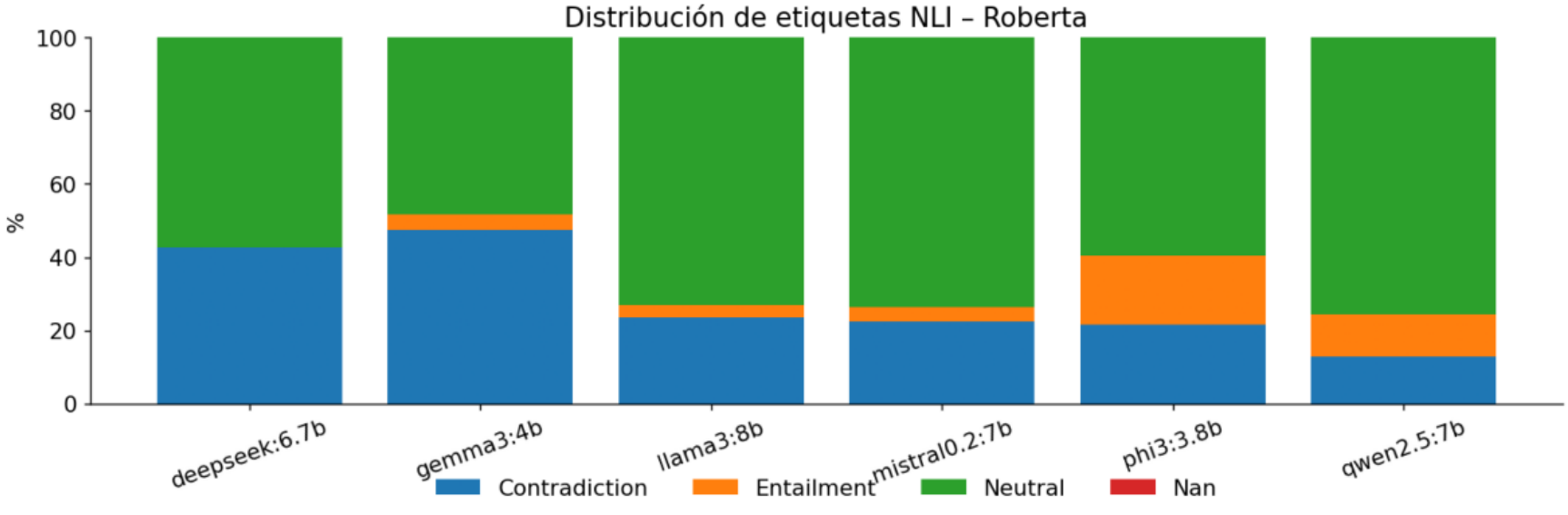

06_porcentajes_por_clasificador.png
Hallucination Rates Across Models and NLI Classifiers
%
0
10
20
30
40
50
Roberta
Deberta
Distilroberta
Consenso
43%
35%
21%
30%
47%
41%
48%
45%
24%
22%
17%
21%
22%
20%
18%
19%
22%
23%
20%
19%
13%
12%
11%
11%
deepseek:6.7b
gemma3:4b
llama3:8b
mistral0.2:7b
phi3:3.8b
qwen2.5:7b


06_hallucination_rate_ci.png
Hallucination Rate (95% CI via bootstrap)
% de respuestas alucinadas (consenso)
10
15
20
25
30
35
40
45
11%
19%
19%
21%
30%
45%
qwen2.5:7b
mistral0.2:7b
phi3:3.8b
llama3:8b
deepseek:6.7b
gemma3:4b

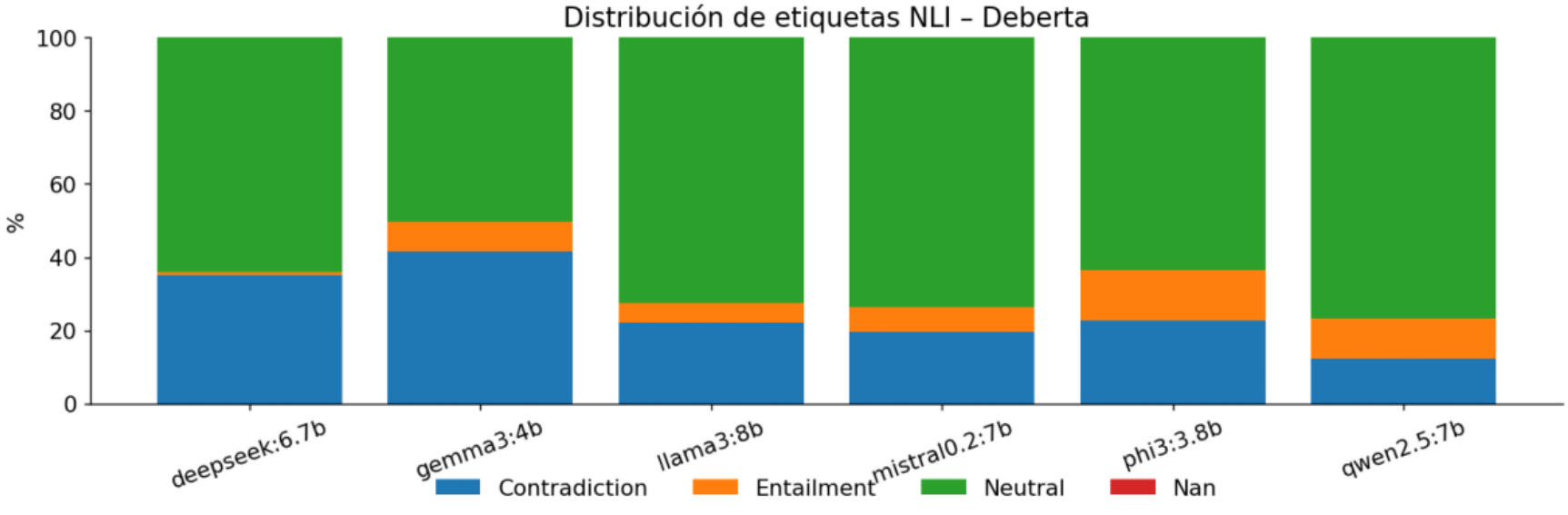

06_metricas_vs_gold_ci.png
Métricas vs GOLD (95% CI)
Accuracy
Precision
Recall
F1
Balanced Acc.
MCC
Porcentaje / Índice (×100)
80
60
40
20
0
deepseek:6.7b
gemma3:4b
llama3:8b
mistral0.2:7b
phi3:3.8b
qwen2.5:7b

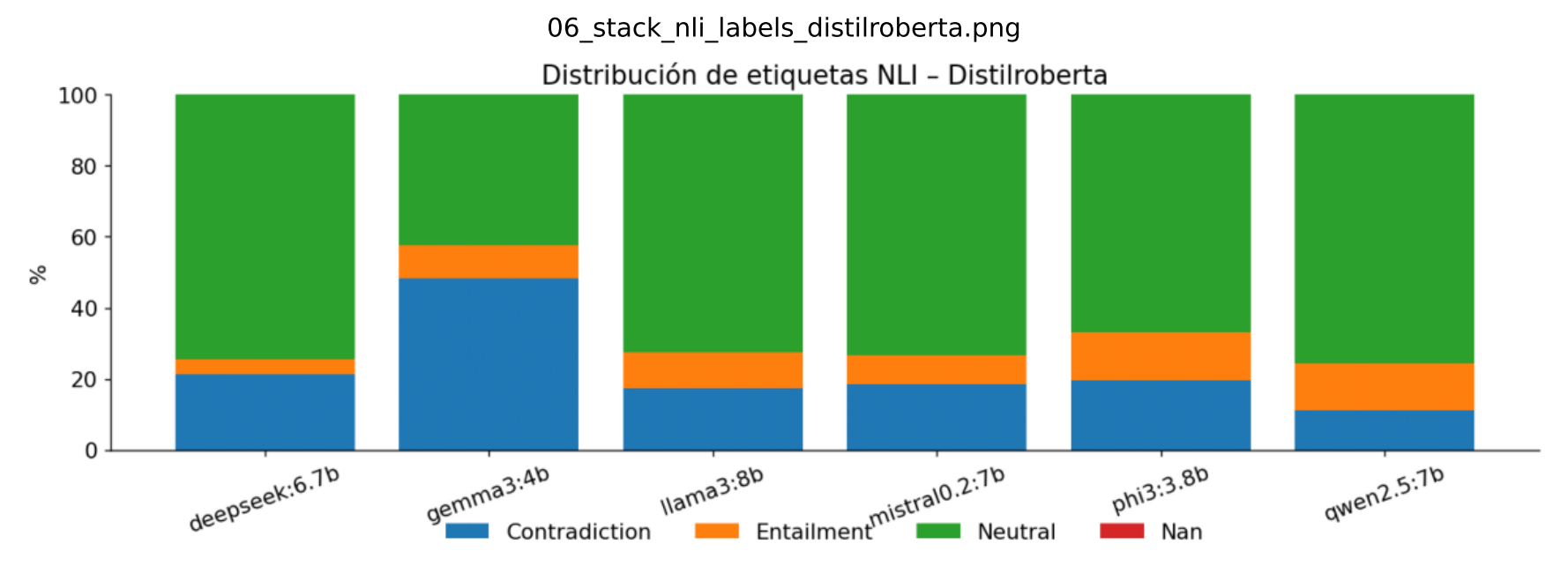

06_kappa_heatmap.png
Agreement entre clasificadores NLI
deepseek:6.7b
gemma3:4b
llama3:8b
mistral0.2:7b
phi3:3.8b
qwen2.5:7b
0.23
0.21
0.29
0.68
0.78
0.73
0.53
0.70
0.55
0.54
0.67
0.55
0.49
0.43
0.36
0.54
0.64
0.54
DistilRoBERTa vs DeBERTa
RoBERTa vs DeBERTa
RoBERTa vs DistilRoBERTa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Cohen's κ (binario por contradiction)

06_stack_nli_labels_roberta.png



06_stack_nli_labels_deberta.png




06_stack_nli_labels_distilroberta.png
Distribución de etiquetas NLI – Distilroberta
%
100
80
60
40
20
0
deepseek:6.7b
gemma3:4b
llama3:8b
mistral0.2:7b
phi3:3.8b
qwen2.5:7b
Contradiction
Entailment
Neutral
Nan

Overview por modelo

Métrica	deepseek:6.7b	gemma3:4b	llama3:8b	mistral0.2:7b	phi3:3.8b	qwen2.5:7b
N	1572	1572	1572	1572	1572	1572
Con NLI	1572	1572	1572	1572	1572	1572
% con NLI	100.0	100.0	100.0	100.0	100.0	100.0
% resp. alucinadas (CI95)	30.0 [27.9, 32.3]	45.3 [42.9, 47.6]	20.5 [18.6, 22.6]	18.5 [16.5, 20.6]	18.8 [16.9, 20.7]	10.9 [9.3, 12.5]
Consistencia mayoría	1.00	1.00	1.00	1.00	1.00	1.00

Métricas vs GOLD por modelo

Métrica	deepseek:6.7b	gemma3:4b	llama3:8b	mistral0.2:7b	phi3:3.8b	qwen2.5:7b
Accuracy % (CI95)	47.6 [45.2, 50.1]	58.5 [56.0, 61.0]	57.2 [55.0, 59.6]	59.2 [56.7, 61.7]	53.9 [51.3, 56.2]	55.5 [53.0, 58.0]
Precision % (CI95)	46.9 [42.4, 51.5]	60.0 [56.2, 63.5]	68.7 [63.8, 73.8]	76.3 [71.5, 81.2]	61.8 [55.7, 67.5]	77.8 [71.3, 84.4]
Recall % (CI95)	27.8 [24.7, 31.0]	53.8 [50.1, 57.3]	28.0 [25.1, 31.1]	28.0 [24.7, 31.2]	23.0 [20.1, 26.0]	16.8 [14.1, 19.3]
F1 % (CI95)	34.9 [31.5, 38.3]	56.7 [53.5, 59.7]	39.7 [36.3, 43.3]	40.9 [37.0, 44.7]	33.6 [29.9, 37.2]	27.6 [23.8, 31.2]
Balanced Acc. % (CI95)	47.9 [45.7, 50.2]	58.6 [56.0, 61.0]	57.5 [55.6, 59.4]	59.5 [57.7, 61.5]	54.3 [52.3, 56.1]	55.9 [54.4, 57.5]
MCC (CI95)	-0.047 [-0.095, 0.004]	0.172 [0.121, 0.221]	0.185 [0.141, 0.232]	0.246 [0.201, 0.291]	0.109 [0.059, 0.155]	0.191 [0.148, 0.237]

Matriz de confusión vs GOLD

Métrica	deepseek:6.7b	gemma3:4b	llama3:8b	mistral0.2:7b	phi3:3.8b	qwen2.5:7b
TP	221	427	222	222	183	133
FP	250	285	101	69	113	38
FN	573	367	572	572	611	661
TN	528	493	677	709	665	740

Mann–Whitney: resumen por modelo

modelo	n_tests	n_sig_fdr	min_p	min_p_adj
deepseek:6.7b	30	0	1.0	1.0
gemma3:4b	30	0	1.0	1.0
llama3:8b	30	0	1.0	1.0
mistral0.2:7b	30	0	1.0	1.0
phi3:3.8b	30	0	1.0	1.0
qwen2.5:7b	30	0	1.0	1.0

Mann–Whitney: Top-20 por p-value (global)

modelo	benchmark	variable	μ(aluc.)	μ(no aluc.)	Δ	p_value	p_adj_bh
deepseek:6.7b	advglue	longitud_tokens	58.463	58.463	0.0	1.0	1.0
deepseek:6.7b	advglue	num_verbos	7.183	7.183	0.0	1.0	1.0
deepseek:6.7b	advglue	num_sustantivos	13.293	13.293	0.0	1.0	1.0
deepseek:6.7b	advglue	num_adjetivos	3.476	3.476	0.0	1.0	1.0
deepseek:6.7b	advglue	num_adverbios	1.037	1.037	0.0	1.0	1.0
deepseek:6.7b	advglue	num_entidades	0.902	0.902	0.0	1.0	1.0
deepseek:6.7b	advglue	longitud_promedio_oracion	20.86	20.86	0.0	1.0	1.0
deepseek:6.7b	advglue	densidad_verbos	0.122	0.122	0.0	1.0	1.0
deepseek:6.7b	advglue	densidad_sustantivos	0.227	0.227	-0.0	1.0	1.0
deepseek:6.7b	advglue	densidad_entidades	0.016	0.016	0.0	1.0	1.0
deepseek:6.7b	fever	longitud_tokens	54.482	54.482	0.0	1.0	1.0
deepseek:6.7b	fever	num_verbos	6.74	6.74	0.0	1.0	1.0
deepseek:6.7b	fever	num_sustantivos	12.382	12.382	0.0	1.0	1.0
deepseek:6.7b	fever	num_adjetivos	2.99	2.99	0.0	1.0	1.0
deepseek:6.7b	fever	num_adverbios	0.615	0.615	0.0	1.0	1.0
deepseek:6.7b	fever	num_entidades	1.063	1.063	0.0	1.0	1.0
deepseek:6.7b	fever	longitud_promedio_oracion	20.015	20.015	0.0	1.0	1.0
deepseek:6.7b	fever	densidad_verbos	0.124	0.124	0.0	1.0	1.0
deepseek:6.7b	fever	densidad_sustantivos	0.221	0.221	-0.0	1.0	1.0
deepseek:6.7b	fever	densidad_entidades	0.019	0.019	0.0	1.0	1.0

Mann–Whitney: Top-10 por p-value – deepseek:6.7b

benchmark	variable	μ(aluc.)	μ(no aluc.)	Δ	p_value	p_adj_bh
advglue	longitud_tokens	58.463	58.463	0.0	1.0	1.0
advglue	num_verbos	7.183	7.183	0.0	1.0	1.0
advglue	num_sustantivos	13.293	13.293	0.0	1.0	1.0
advglue	num_adjetivos	3.476	3.476	0.0	1.0	1.0
advglue	num_adverbios	1.037	1.037	0.0	1.0	1.0
advglue	num_entidades	0.902	0.902	0.0	1.0	1.0
advglue	longitud_promedio_oracion	20.86	20.86	0.0	1.0	1.0
advglue	densidad_verbos	0.122	0.122	0.0	1.0	1.0
advglue	densidad_sustantivos	0.227	0.227	-0.0	1.0	1.0
advglue	densidad_entidades	0.016	0.016	0.0	1.0	1.0

Mann–Whitney: Top-10 por p-value – gemma3:4b

benchmark	variable	μ(aluc.)	μ(no aluc.)	Δ	p_value	p_adj_bh
advglue	longitud_tokens	73.768	73.768	0.0	1.0	1.0
advglue	num_verbos	6.988	6.988	0.0	1.0	1.0
advglue	num_sustantivos	14.585	14.585	0.0	1.0	1.0
advglue	num_adjetivos	5.39	5.39	0.0	1.0	1.0
advglue	num_adverbios	2.537	2.537	0.0	1.0	1.0
advglue	num_entidades	2.329	2.329	0.0	1.0	1.0
advglue	longitud_promedio_oracion	14.821	14.821	0.0	1.0	1.0
advglue	densidad_verbos	0.106	0.106	0.0	1.0	1.0
advglue	densidad_sustantivos	0.197	0.197	0.0	1.0	1.0
advglue	densidad_entidades	0.034	0.034	0.0	1.0	1.0

Mann–Whitney: Top-10 por p-value – llama3:8b

benchmark	variable	μ(aluc.)	μ(no aluc.)	Δ	p_value	p_adj_bh
advglue	longitud_tokens	229.915	229.915	0.0	1.0	1.0
advglue	num_verbos	24.049	24.049	0.0	1.0	1.0
advglue	num_sustantivos	42.854	42.854	0.0	1.0	1.0
advglue	num_adjetivos	17.439	17.439	0.0	1.0	1.0
advglue	num_adverbios	8.183	8.183	0.0	1.0	1.0
advglue	num_entidades	7.378	7.378	0.0	1.0	1.0
advglue	longitud_promedio_oracion	20.355	20.355	0.0	1.0	1.0
advglue	densidad_verbos	0.106	0.106	0.0	1.0	1.0
advglue	densidad_sustantivos	0.181	0.181	-0.0	1.0	1.0
advglue	densidad_entidades	0.031	0.031	0.0	1.0	1.0

Mann–Whitney: Top-10 por p-value – mistral0.2:7b

benchmark	variable	μ(aluc.)	μ(no aluc.)	Δ	p_value	p_adj_bh
advglue	longitud_tokens	200.756	200.756	0.0	1.0	1.0
advglue	num_verbos	21.598	21.598	0.0	1.0	1.0
advglue	num_sustantivos	41.037	41.037	0.0	1.0	1.0
advglue	num_adjetivos	15.854	15.854	0.0	1.0	1.0
advglue	num_adverbios	5.976	5.976	0.0	1.0	1.0
advglue	num_entidades	8.061	8.061	0.0	1.0	1.0
advglue	longitud_promedio_oracion	25.33	25.33	0.0	1.0	1.0
advglue	densidad_verbos	0.115	0.115	0.0	1.0	1.0
advglue	densidad_sustantivos	0.19	0.19	0.0	1.0	1.0
advglue	densidad_entidades	0.037	0.037	0.0	1.0	1.0

Mann–Whitney: Top-10 por p-value – phi3:3.8b

benchmark	variable	μ(aluc.)	μ(no aluc.)	Δ	p_value	p_adj_bh
advglue	longitud_tokens	170.354	170.354	0.0	1.0	1.0
advglue	num_verbos	18.037	18.037	0.0	1.0	1.0
advglue	num_sustantivos	30.793	30.793	0.0	1.0	1.0
advglue	num_adjetivos	11.512	11.512	0.0	1.0	1.0
advglue	num_adverbios	5.476	5.476	0.0	1.0	1.0
advglue	num_entidades	9.073	9.073	0.0	1.0	1.0
advglue	longitud_promedio_oracion	18.51	18.51	0.0	1.0	1.0
advglue	densidad_verbos	0.083	0.083	0.0	1.0	1.0
advglue	densidad_sustantivos	0.14	0.14	0.0	1.0	1.0
advglue	densidad_entidades	0.032	0.032	0.0	1.0	1.0

Mann–Whitney: Top-10 por p-value – qwen2.5:7b

benchmark	variable	μ(aluc.)	μ(no aluc.)	Δ	p_value	p_adj_bh
advglue	longitud_tokens	89.963	89.963	0.0	1.0	1.0
advglue	num_verbos	10.854	10.854	0.0	1.0	1.0
advglue	num_sustantivos	16.402	16.402	0.0	1.0	1.0
advglue	num_adjetivos	5.268	5.268	0.0	1.0	1.0
advglue	num_adverbios	4.012	4.012	0.0	1.0	1.0
advglue	num_entidades	3.366	3.366	0.0	1.0	1.0
advglue	longitud_promedio_oracion	20.854	20.854	0.0	1.0	1.0
advglue	densidad_verbos	0.118	0.118	0.0	1.0	1.0
advglue	densidad_sustantivos	0.18	0.18	0.0	1.0	1.0
advglue	densidad_entidades	0.039	0.039	0.0	1.0	1.0