

B Results of 95% CIs for Each Metric

	Reddit_Impacts	BC5CDR	MIMIC III	NCBI	Med-Mentions
GPT-3.5					
Basic Prompt (BP)	16.73 [11.53, 22.83]	64.56 [61.55, 67.73]	54.70 [49.60, 58.73]	26.96 [24.43, 30.98]	9.27 [7.81, 12.22]
BP + Description of datasets	21.15 [14.88, 26.64]	68.61 [66.74, 70.72]	56.73 [52.58, 61.22]	34.48 [31.08, 39.25]	12.71 [10.93, 15.65]
BP + High-frequency instances	21.15 [15.75, 27.40]	69.01 [66.24, 70.98]	57.72 [52.75, 62.26]	35.95 [33.36, 38.44]	17.22 [14.47, 19.80]
BP + UMLS knowledge	16.44 [8.43, 23.07]	64.83 [61.83, 66.41]	50.57 [46.17, 55.04]	30.75 [27.73, 33.26]	10.88 [8.81, 12.29]
BP + Error analysis	19.24 [12.91, 26.17]	67.67 [65.53, 70.32]	59.52 [54.96, 64.47]	33.15 [31.24, 38.87]	15.52 [13.14, 17.20]
BP + 5-shot learning with sentences	19.30 [12.26, 25.78]	68.84 [67.25, 70.49]	57.03 [53.06, 62.85]	40.16 [38.78, 46.45]	20.61 [17.58, 22.29]
BP + 5-shot learning with tokens	21.69 [15.92, 28.89]	70.79 [68.87, 73.15]	61.21 [56.81, 66.05]	43.01 [41.43, 48.21]	24.57 [22.88, 26.64]
BP + All above	23.91 [15.87, 30.97]	72.73 [70.32, 74.86]	61.99 [57.24, 66.38]	45.24 [42.64, 50.58]	31.63 [29.36, 34.74]
GPT-4					
Basic Prompt (BP)	20.16 [13.29, 26.54]	69.43 [66.28, 72.44]	56.63 [51.27, 60.83]	33.56 [31.59, 37.25]	13.83 [11.85, 15.09]
BP + Description of datasets	23.52 [16.46, 30.84]	70.65 [67.47, 72.72]	59.68 [55.18, 64.09]	35.75 [33.54, 40.58]	15.30 [13.61, 17.15]
BP + High-frequency instances	24.64 [17.72, 31.11]	72.60 [71.17, 74.28]	60.08 [56.33, 65.37]	37.96 [36.95, 41.73]	19.50 [17.11, 22.97]
BP + UMLS knowledge	20.46 [13.84, 27.07]	69.86 [66.05, 72.62]	55.13 [50.20, 60.29]	30.90 [28.68, 34.30]	14.50 [12.57, 16.46]
BP + Error analysis	23.13 [16.65, 30.69]	74.61 [71.44, 77.29]	60.11 [55.44, 64.72]	37.84 [34.13, 42.71]	18.25 [15.06, 20.43]
BP + 5-shot learning with sentences	22.88 [16.23, 30.59]	73.00 [71.26, 76.22]	58.25 [53.28, 63.95]	40.86 [39.37, 45.36]	28.80 [26.71, 30.20]
BP + 5-shot learning with tokens	25.95 [18.50, 32.07]	76.65 [74.15, 77.92]	62.94 [57.56, 66.87]	44.24 [42.93, 48.28]	33.20 [31.64, 35.70]
BP + All above	27.60 [19.43, 33.80]	78.03 [75.51, 80.02]	63.58 [58.73, 67.18]	46.93 [44.85, 51.58]	37.95 [35.88, 39.90]
Llama3					
Basic Prompt (BP)	15.61 [8.20, 22.12]	62.13 [59.24, 63.58]	50.70 [45.93, 54.19]	19.15 [15.21, 21.38]	21.23 [19.24, 23.42]
BP + Description of datasets	19.28 [11.71, 25.96]	67.68 [64.86, 69.10]	56.22 [52.77, 60.25]	21.44 [20.80, 24.65]	21.57 [19.30, 24.76]
BP + High-frequency instances	20.44 [13.79, 27.51]	68.39 [66.48, 70.35]	56.06 [52.62, 61.42]	26.62 [22.16, 28.31]	27.12 [26.37, 29.35]
BP + UMLS knowledge	12.91 [7.40, 18.71]	64.71 [61.44, 67.01]	48.92 [44.75, 53.37]	20.91 [17.07, 22.61]	23.68 [20.59, 25.17]
BP + Error analysis	18.87 [13.34, 25.13]	68.07 [65.41, 70.58]	58.92 [53.90, 63.84]	24.46 [20.97, 25.20]	25.78 [23.48, 27.56]
BP + 5-shot learning with sentences	17.65 [13.62, 24.69]	70.70 [69.36, 72.83]	56.85 [52.32, 61.33]	30.52 [26.50, 33.96]	34.87 [32.18, 37.25]
BP + 5-shot learning with tokens	20.04 [14.81, 27.29]	71.76 [69.58, 73.51]	61.98 [56.59, 65.18]	33.42 [28.72, 35.12]	35.23 [33.17, 37.08]
BP + All above	21.43 [14.24, 28.80]	73.32 [72.27, 74.26]	62.94 [57.07, 65.79]	34.80 [28.57, 35.44]	37.26 [35.45, 39.08]

Table 6. Evaluation of static prompting strategies using GPT-3.5, GPT-4 and Llama 3 across five biomedical datasets. The table presents F_1 -score with 95% confidence intervals reported for each metric to indicate the statistical reliability of the results.

		<i>Reddit_Impacts</i>	<i>BC5CDR</i>	<i>MIMIC III</i>	<i>NCBI</i>	<i>Med-Mentions</i>
GPT-4						
5-shot	Base	27.60 [19.43, 33.80]	78.03 [75.51, 80.02]	63.58 [58.73, 67.18]	46.93 [44.85, 51.58]	37.95 [35.88, 39.90]
	TF-IDF	28.47 [21.78, 35.47]	85.88 [84.53, 86.42]	76.24 [72.98, 79.63]	60.08 [56.70, 63.32]	37.96 [35.90, 39.84]
	SBERT	33.72 [26.28, 42.20]	83.37 [82.51, 84.22]	73.44 [69.91, 76.81]	57.56 [54.05, 60.73]	39.12 [36.84, 41.34]
	ColBERT	32.39 [25.10, 39.85]	79.82 [78.24, 80.98]	75.56 [72.06, 78.94]	52.38 [49.06, 55.55]	39.93 [37.93, 41.73]
	DPR	32.64 [25.42, 40.17]	83.58 [82.30, 84.88]	69.89 [65.75, 73.63]	49.37 [45.37, 52.94]	39.13 [34.44, 41.35]
10-shot	Base	31.92 [23.77, 38.44]	81.27 [80.81, 82.37]	70.52 [66.10, 73.81]	52.67 [49.36, 56.76]	36.74 [32.29, 38.83]
	TF-IDF	31.14 [24.33, 38.13]	86.64 [85.15, 88.09]	75.53 [72.18, 79.10]	62.05 [58.79, 65.11]	40.37 [38.23, 42.43]
	SBERT	35.47 [27.17, 43.21]	85.92 [83.09, 87.27]	73.89 [70.22, 77.80]	60.83 [57.47, 64.03]	40.37 [38.23, 42.39]
	ColBERT	33.81 [26.24, 41.55]	85.71 [84.42, 86.07]	76.34 [73.01, 79.68]	57.25 [53.75, 60.72]	40.48 [38.13, 42.54]
	DPR	32.61 [24.50, 40.33]	84.79 [82.96, 86.78]	72.13 [68.06, 75.85]	58.70 [54.99, 61.91]	40.25 [30.83, 50.75]
20-shot	Base	37.67 [30.04, 43.44]	81.15 [80.40, 82.24]	70.98 [65.77, 73.82]	51.98 [50.33, 58.84]	38.39 [35.26, 40.29]
	TF-IDF	38.35 [30.77, 46.28]	87.16 [85.77, 88.62]	77.66 [71.91, 78.88]	64.36 [61.18, 67.87]	41.32 [39.21, 43.26]
	SBERT	38.22 [28.57, 44.90]	87.42 [85.26, 89.12]	75.14 [71.77, 78.75]	62.21 [59.01, 65.18]	39.37 [35.05, 40.39]
	ColBERT	42.49 [32.52, 48.33]	83.00 [81.39, 84.40]	76.70 [73.11, 79.89]	57.69 [54.22, 61.18]	40.53 [37.61, 43.26]
	DPR	38.84 [29.01, 44.44]	85.60 [84.28, 86.93]	72.28 [68.56, 75.95]	60.34 [56.54, 63.72]	39.23 [34.22, 41.56]
Llama3						
5-shot	Base	21.43 [14.24, 28.80]	73.32 [72.27, 74.26]	62.94 [57.07, 65.79]	34.80 [28.57, 35.44]	37.26 [35.45, 39.08]
	TF-IDF	28.57 [21.74, 36.06]	80.11 [79.25, 81.00]	70.41 [66.87, 73.76]	49.80 [46.38, 53.03]	38.68 [35.67, 40.81]
	SBERT	34.42 [26.28, 41.52]	80.39 [79.50, 81.33]	67.88 [64.09, 71.69]	50.12 [46.89, 53.66]	37.91 [36.02, 39.81]
	ColBERT	32.94 [25.00, 39.84]	71.76 [70.75, 72.69]	71.68 [68.08, 75.21]	45.50 [41.95, 49.49]	38.99 [36.15, 41.34]
	DPR	29.00 [22.86, 36.36]	75.67 [74.67, 76.70]	68.97 [65.05, 72.70]	44.54 [41.24, 48.25]	38.66 [36.78, 40.50]
10-shot	Base	32.50 [26.94, 42.26]	75.15 [74.65, 76.67]	63.77 [58.59, 67.75]	35.60 [32.17, 39.12]	36.50 [35.73, 39.57]
	TF-IDF	34.21 [27.24, 42.03]	80.57 [79.65, 81.47]	55.56 [53.11, 60.44]	49.50 [46.05, 52.92]	35.51 [34.75, 37.45]
	SBERT	32.45 [25.33, 39.63]	81.17 [80.26, 82.03]	71.63 [67.75, 75.15]	51.35 [47.49, 55.16]	39.08 [36.39, 41.38]
	ColBERT	32.89 [20.35, 35.05]	80.34 [79.53, 81.24]	72.85 [69.46, 76.49]	38.77 [34.91, 42.29]	38.06 [35.52, 40.71]
	DPR	34.29 [26.11, 41.98]	74.72 [73.77, 75.73]	69.54 [65.61, 73.17]	46.28 [42.77, 49.65]	37.85 [36.06, 39.69]
20-shot	Base	33.67 [24.09, 40.88]	75.50 [73.57, 76.36]	62.05 [58.23, 67.15]	41.62 [38.83, 45.71]	37.67 [35.22, 40.57]
	TF-IDF	39.11 [31.34, 47.70]	78.36 [77.42, 79.30]	57.66 [51.19, 59.80]	47.50 [43.87, 50.84]	38.83 [37.54, 39.11]
	SBERT	41.43 [31.58, 48.98]	76.85 [74.86, 78.96]	65.35 [60.44, 70.40]	44.14 [40.57, 47.86]	36.01 [34.16, 37.75]
	ColBERT	34.66 [24.07, 36.31]	72.19 [71.17, 73.20]	57.63 [53.19, 61.93]	48.44 [45.07, 51.78]	36.85 [34.10, 39.29]
	DPR	37.30 [27.13, 44.76]	74.80 [72.49, 76.36]	65.80 [61.82, 69.69]	40.36 [36.96, 43.96]	36.89 [34.46, 38.84]

Table 7. Evaluation of dynamic prompting strategies (5-shot, 10-shot, and 20-shot) using GPT-4 and Llama 3 across five biomedical datasets. The table presents F₁-score for each retrieval method: Base Prompt, TF-IDF, SBERT, ColBERT, and DPR, with 95% confidence intervals reported for each metric to indicate the statistical reliability of the results.