

**A Averaged Performance of the Baseline Dynamic Prompt Model**

Precision			Recall	F <sub>1</sub> -score	Precision			Recall	F <sub>1</sub> -score
GPT 4	5-shot	16.40	50.00	24.70	Llama3	5-shot	12.56	56.32	20.55
		17.67	50.62	26.20			13.68	55.81	21.97
		17.48	53.09	26.30			12.14	59.52	20.16
		23.91	54.32	33.20			14.25	59.79	23.02
	AVG	18.865	52.0075	27.60		AVG	13.1575	57.86	21.425
	10-shot	25.41	58.75	35.47		10-shot	24.51	59.52	34.72
		21.67	54.32	30.99			19.63	62.35	29.86
		21.43	59.26	31.48			23.96	57.44	33.81
		20.47	54.32	29.73			21.38	60.43	31.59
	AVG	22.245	56.6625	31.9175		AVG	22.37	59.935	32.495
	20-shot	28.74	58.63	38.57		20-shot	27.22	57.65	36.98
		27.03	57.54	36.78			23.24	52.44	32.21
		26.52	59.79	36.74			22.12	52.08	31.06
		28.65	59.02	38.57			25.49	53.06	34.44
	AVG	27.735	58.745	37.665		AVG	24.5175	53.8075	33.6725

**Table 1.** Averaged performance of the baseline dynamic prompt model on the REDDIT-IMPACTS dataset across different shot settings.

		Precision	Recall	F <sub>1</sub> -score			Precision	Recall	F <sub>1</sub> -score
GPT 4	5-shot	69.97	91.43	79.27	Llama3	5-shot	68.09	84.35	75.35
		68.32	88.3	77.29			69.70	76.67	73.02
		64.58	90.15	75.25			71.00	79.15	74.86
		71.59	91.39	80.29			67.10	73.25	70.04
	AVG	68.615	90.3175	78.025		AVG	68.9725	78.355	73.3175
	10-shot	75.11	90.84	82.23		10-shot	72.38	79.18	75.63
		74.41	90.32	81.60			73.31	74.24	73.77
		74.13	85.71	79.50			71.15	77.05	73.98
		77.65	86.35	81.73			73.40	81.17	77.2
	AVG	75.325	88.305	81.265		AVG	72.56	77.91	75.145
	20-shot	75.36	88.33	81.33		20-shot	75.12	74.84	74.98
		73.13	91.74	81.38			72.03	71.88	71.96
		71.84	91.1	80.33			76.88	77.20	77.04
		77.94	85.54	81.57			77.64	78.40	78.02
	AVG	74.5675	89.1775	81.1525		AVG	75.4175	75.58	75.5

**Table 2.** Averaged performance of the baseline dynamic prompt model on the BC5CDR dataset across different shot settings.

		Precision	Recall	F <sub>1</sub> -score			Precision	Recall	F <sub>1</sub> -score
GPT 4	5-shot	62.38	63.78	63.07	Llama3	5-shot	57.69	68.39	62.58
		61.88	62.19	62.03			61.83	61.33	61.58
		65.26	68.14	66.67			58.24	68.06	62.77
		62.71	62.36	62.54			59.44	71.29	64.82
	AVG	63.0575	64.1175	63.5775		AVG	59.3	67.2675	62.9375
	10-shot	66.38	74.24	70.09		10-shot	59.13	71.63	64.78
		68.37	74.86	71.47			62.68	63.8	63.23
		66.72	75.33	70.77			61.71	67.58	64.51
		66.46	73.34	69.73			62.24	62.84	62.54
	AVG	66.9825	74.4425	70.515		AVG	61.44	66.4625	63.765
	20-shot	70.41	69.84	70.12		20-shot	63.89	62.04	62.95
		70.11	71.24	70.67			63.96	62.83	63.39
		71.45	73.39	72.41			59.22	61.71	60.44
		70.64	70.80	70.72			60.96	61.88	61.42
	AVG	70.6525	71.3175	70.98		AVG	62.0075	62.115	62.05

**Table 3.** Averaged performance of the baseline dynamic prompt model on the MIMIC III dataset across different shot settings.

		Precision	Recall	F <sub>1</sub> -score			Precision	Recall	F <sub>1</sub> -score
GPT 4	5-shot	46.17	50.52	48.25	Llama3	5-shot	33.88	40.67	36.97
		45.69	49.07	47.32			32.26	39.14	35.37
		40.24	43.65	41.88			36.38	28.49	31.95
		47.96	52.83	50.28			40.73	30.53	34.9
	AVG	45.015	49.0175	46.9325		AVG	35.8125	34.7075	34.7975
	10-shot	52.98	52.67	52.82		10-shot	42.79	31.08	35.66
		53.71	53.36	53.54			40.18	28.73	36.27
		52.99	50.35	51.64			35.75	32.05	33.76
		53.85	52.3	53.06			39.95	34.11	36.71
	AVG	53.3825	52.17	52.765		AVG	39.6675	31.4925	35.6
	20-shot	54.57	54.73	54.65		20-shot	40.59	42.07	41.32
		44.86	45.43	45.14			40.87	42.8	41.81
		51.6	51.75	51.68			41.48	43.02	42.24
		55.7	57.24	56.46			39.88	42.41	41.1
	AVG	51.6825	52.2875	51.9825		AVG	40.705	42.575	41.6175

**Table 4.** Averaged performance of the baseline dynamic prompt model on the NCBI dataset across different shot settings.

		Precision	Recall	F <sub>1</sub> -score			Precision	Recall	F <sub>1</sub> -score
GPT 4	5-shot	27.24	62.73	37.99	Llama3	5-shot	24.94	73.03	37.18
		26.71	58.58	36.69			25.95	61.96	36.58
		29.23	60.47	39.41			26.11	73.57	38.54
		25.86	58.41	35.85			26.55	59.62	36.74
	AVG	27.26	60.0475	37.485		AVG	25.8875	67.045	37.26
	10-shot	26.92	58.26	36.82		10-shot	25.67	70.61	37.65
		26.1	59.15	36.22			25.1	59.19	35.15
		24.41	58.56	34.46			25.76	70.06	37.67
		29.19	60.82	39.45			25.73	57.46	35.54
	AVG	26.655	59.1975	36.7375		AVG	25.565	64.33	36.5025
	20-shot	27.86	59.12	37.87		20-shot	26.18	63.59	37.09
		25.02	60.64	35.42			26.63	61.14	37.1
		29.33	61.71	39.76			25.93	63.57	36.84
		30.18	61.66	40.52			27.54	70.85	39.66
	AVG	28.0975	60.7825	38.3925		AVG	26.57	64.7875	37.6725

**Table 5.** Averaged performance of the baseline dynamic prompt model on the Med-Mentions dataset across different shot settings.

## B Results of 95% CIs for Each Metric

	Reddit_Impacts	BC5CDR	MIMIC III	NCBI	Med-Mentions
<b>GPT-3.5</b>					
<b>Basic Prompt (BP)</b>	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]
<b>BP + Description of datasets</b>	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]
<b>BP + High-frequency instances</b>	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]
<b>BP + UMLS knowledge</b>	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]
<b>BP + Error analysis</b>	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]
<b>BP + 5-shot learning with sentences</b>	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]
<b>BP + 5-shot learning with tokens</b>	<b>21.69 [15.92, 28.89]</b>	<b>70.79 [68.87, 73.15]</b>	<b>61.21 [56.81, 66.05]</b>	<b>43.01 [41.43, 48.21]</b>	<b>24.57 [22.88, 26.64]</b>
<b>BP + All above</b>	<b>23.91 [15.87, 30.97]</b>	<b>72.73 [70.32, 74.86]</b>	<b>61.99 [57.24, 66.38]</b>	<b>45.24 [42.64, 50.58]</b>	<b>31.63 [29.36, 34.74]</b>
<b>GPT-4</b>					
<b>Basic Prompt (BP)</b>	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]
<b>BP + Description of datasets</b>	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]
<b>BP + High-frequency instances</b>	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]
<b>BP + UMLS knowledge</b>	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]
<b>BP + Error analysis</b>	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]
<b>BP + 5-shot learning with sentences</b>	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]
<b>BP + 5-shot learning with tokens</b>	<b>25.95 [18.50, 32.07]</b>	<b>76.65 [74.15, 77.92]</b>	<b>62.94 [57.56, 66.87]</b>	<b>44.24 [42.93, 48.28]</b>	<b>33.20 [31.64, 35.70]</b>
<b>BP + All above</b>	<b>27.60 [19.43, 33.80]</b>	<b>78.03 [75.51, 80.02]</b>	<b>63.58 [58.73, 67.18]</b>	<b>46.93 [44.85, 51.58]</b>	<b>37.95 [35.88, 39.90]</b>
<b>Llama3</b>					
<b>Basic Prompt (BP)</b>	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]
<b>BP + Description of datasets</b>	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]
<b>BP + High-frequency instances</b>	<b>20.44 [13.79, 27.51]</b>	68.39 [66.48, 70.35]	56.06 [52.62, 61.42]	26.62 [22.16, 28.31]	27.12 [26.37, 29.35]
<b>BP + UMLS knowledge</b>	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]
<b>BP + Error analysis</b>	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]
<b>BP + 5-shot learning with sentences</b>	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]
<b>BP + 5-shot learning with tokens</b>	20.04 [14.81, 27.29]	<b>71.76 [69.58, 73.51]</b>	<b>61.98 [56.59, 65.18]</b>	<b>33.42 [28.72, 35.12]</b>	<b>35.23 [33.17, 37.08]</b>
<b>BP + All above</b>	<b>21.43 [14.24, 28.80]</b>	<b>73.32 [72.27, 74.26]</b>	<b>62.94 [57.07, 65.79]</b>	<b>34.80 [28.57, 35.44]</b>	<b>37.26 [35.45, 39.08]</b>

**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>
<b>GPT-4</b>						
5-shot	<b>Base</b>	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]
	<b>TF-IDF</b>	28.47 [21.78, 35.47]	<b>85.88 [84.53, 86.42]</b>	<b>76.24 [72.98, 79.63]</b>	<b>60.08 [56.70, 63.32]</b>	37.96 [35.90, 39.84]
	<b>SBERT</b>	<b>33.72 [26.28, 42.20]</b>	83.37 [82.51, 84.22]	73.44 [69.91, 76.81]	57.56 [54.05, 60.73]	39.12 [36.84, 41.34]
	<b>ColBERT</b>	32.39 [25.10, 39.85]	79.82 [78.24, 80.98]	75.56 [72.06, 78.94]	52.38 [49.06, 55.55]	<b>39.93 [37.93, 41.73]</b>
	<b>DPR</b>	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	<b>Base</b>	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]
	<b>TF-IDF</b>	31.14 [24.33, 38.13]	<b>86.64 [85.15, 88.09]</b>	75.53 [72.18, 79.10]	<b>62.05 [58.79, 65.11]</b>	40.37 [38.23, 42.43]
	<b>SBERT</b>	<b>35.47 [27.17, 43.21]</b>	85.92 [83.09, 87.27]	73.89 [70.22, 77.80]	60.83 [57.47, 64.03]	40.37 [38.23, 42.39]
	<b>ColBERT</b>	33.81 [26.24, 41.55]	85.71 [84.42, 86.07]	<b>76.34 [73.01, 79.68]</b>	57.25 [53.75, 60.72]	<b>40.48 [38.13, 42.54]</b>
	<b>DPR</b>	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	<b>Base</b>	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]
	<b>TF-IDF</b>	38.35 [30.77, 46.28]	87.16 [85.77, 88.62]	<b>77.66 [71.91, 78.88]</b>	<b>64.36 [61.18, 67.87]</b>	<b>41.32 [39.21, 43.26]</b>
	<b>SBERT</b>	38.22 [28.57, 44.90]	<b>87.42 [85.26, 89.12]</b>	75.14 [71.77, 78.75]	62.21 [59.01, 65.18]	39.37 [35.05, 40.39]
	<b>ColBERT</b>	<b>42.49 [32.52, 48.33]</b>	83.00 [81.39, 84.40]	76.70 [73.11, 79.89]	57.69 [54.22, 61.18]	40.53 [37.61, 43.26]
	<b>DPR</b>	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]
<b>Llama3</b>						
5-shot	<b>Base</b>	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]
	<b>TF-IDF</b>	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]
	<b>SBERT</b>	<b>34.42 [26.28, 41.52]</b>	<b>80.39 [79.50, 81.33]</b>	67.88 [64.09, 71.69]	<b>50.12 [46.89, 53.66]</b>	37.91 [36.02, 39.81]
	<b>ColBERT</b>	32.94 [25.00, 39.84]	71.76 [70.75, 72.69]	<b>71.68 [68.08, 75.21]</b>	45.50 [41.95, 49.49]	<b>38.99 [36.15, 41.34]</b>
	<b>DPR</b>	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	<b>Base</b>	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]
	<b>TF-IDF</b>	34.21 [27.24, 42.03]	<b>80.57 [79.65, 81.47]</b>	55.56 [53.11, 60.44]	49.50 [46.05, 52.92]	35.51 [34.75, 37.45]
	<b>SBERT</b>	32.45 [25.33, 39.63]	81.17 [80.26, 82.03]	71.63 [67.75, 75.15]	<b>51.35 [47.49, 55.16]</b>	<b>39.08 [36.39, 41.38]</b>
	<b>ColBERT</b>	32.89 [20.35, 35.05]	80.34 [79.53, 81.24]	<b>72.85 [69.46, 76.49]</b>	38.77 [34.91, 42.29]	38.06 [35.52, 40.71]
	<b>DPR</b>	<b>34.29 [26.11, 41.98]</b>	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	<b>Base</b>	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]
	<b>TF-IDF</b>	39.11 [31.34, 47.70]	<b>78.36 [77.42, 79.30]</b>	57.66 [51.19, 59.80]	47.50 [43.87, 50.84]	<b>38.83 [37.54, 39.11]</b>
	<b>SBERT</b>	<b>41.43 [31.58, 48.98]</b>	76.85 [74.86, 78.96]	65.35 [60.44, 70.40]	44.14 [40.57, 47.86]	36.01 [34.16, 37.75]
	<b>ColBERT</b>	34.66 [24.07, 36.31]	72.19 [71.17, 73.20]	57.63 [53.19, 61.93]	<b>48.44 [45.07, 51.78]</b>	36.85 [34.10, 39.29]
	<b>DPR</b>	37.30 [27.13, 44.76]	74.80 [72.49, 76.36]	<b>65.80 [61.82, 69.69]</b>	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 F<sub>1</sub>-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.

## C Detailed Task-specific Static Prompts

Prompt Strategies	Reddit-Impacts
Basic Prompt	<p><b>[Task Description]:</b> You are a medical AI trained to identify and classify tokens into <a href="#">three categories: Clinical Impacts, Social Impacts, and Outside ('O')</a>. Your task is to extract and classify the clinical and social impacts from this dataset, considering your knowledge of the lifestyle of this population and the potential clinical and social impacts they might experience.</p>
	<p><b>[Entity Types with Definitions]:</b> '<a href="#">Clinical Impacts</a>' <a href="#">refer to</a> tokens describing the effects, consequences, or impacts of substance use on individual health or well-being, as defined in UMLS. '<a href="#">Social Impacts</a>' <a href="#">describe</a> the societal, interpersonal, or community-level effects, also based on UMLS definitions. Any token not falling into these categories should be labeled as 'O'.</p>
	<p><b>[Format Specification]:</b> For example, <a href="#">the sentence</a> 'I was a codeine addict.' <a href="#">is tokenized and labeled</a> as follows: ['I', 'was', 'a', 'codeine', 'addict', '.'] with labels ['O', 'O', 'O', 'Clinical Impacts', 'Clinical Impacts', 'O']. Your task is to predict and return the label for each provided token, ensuring the number of output labels matches the number of input tokens exactly. <a href="#">The output format should be tokens with their labels:</a> ['I-O', 'was-O', 'a-O', 'codeine-Clinical Impacts', 'addict-Clinical Impacts', '.-O'].</p>
Description of datasets	The data you are working with has been collected from 14 forums on Reddit (subreddits) that focused on prescription and illicit opioids, and medications for opioid use disorder. This dataset represents a social media context, coming from individuals who may use prescription and illicit opioids and stimulants.
High-frequency instances	In this dataset, <a href="#">high-frequency clinical impacts</a> include 'withdrawal', 'rehab', 'addicted', 'detox', 'overdosed', and 'rehabs'. <a href="#">High-frequency social impacts</a> include 'lost', 'homeless', 'charged', 'streets', 'jail', and 'disorderly'.
UMLS knowledge	<a href="#">The Unified Medical Language System (UMLS)</a> is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. <a href="#">You understand medical terminology and concepts from UMLS.</a>
Error analysis	<a href="#">Possible analysis of prediction errors:</a> If a sentence describes the background information of an event, facility, or project, then even if it mentions keywords related to social impact like 'at jail', it still cannot be determined as describing a patient being in jail. It is essential to clearly determine whether the sentence is describing the patient's condition. Second, if the sentence is about the usage, operation, or introduction of a drug or medicine, it does not belong to the patient's clinical impacts, even if it mentions some symptoms. Pay attention to whether the sentence contains words like 'if' that indicate conditions.

**Table 8.** Specific static prompts for each component we used for the REDDIT-IMPACTS dataset.

Prompt Strategies	BC5CDR
	<p><b>[Task Description]:</b> You are a medical AI trained to identify and classify tokens into <a href="#">three categories</a>: 'Disease', 'Chemical' and Outside ('O'). Your task is to extract and classify the Disease and Chemical related concepts from this dataset.</p> <p><b>[Entity Types with Definitions]:</b> 'Disease' is a particular abnormal condition that adversely affects the structure or function of all or part of an organism and is not immediately due to any external injury. Diseases are often known to be medical conditions that are associated with specific signs and symptoms. A disease may be caused by external factors such as pathogens or by internal dysfunctions. For example, internal dysfunctions of the immune system can produce a variety of different diseases, including various forms of immunodeficiency, hypersensitivity, allergies, and autoimmune disorders. 'Chemical' in this context refers to substances or compounds with specific chemical properties and structures. These can include drugs, neurotransmitters, elements or ions, vitamins, and other medically relevant chemicals. Any token not falling into Disease categories should be labeled as 'O'.</p> <p><b>[Format Specification]:</b> For example, the sentence 'The hypotensive effect of 100 mg / kg alpha-methyldopa was also partially reversed by naloxone.' is <a href="#">tokenized and labeled</a> as follows: ['The', 'hypotensive', 'effect', 'of', '100', 'mg', '/', 'kg', 'alpha-methyldopa', 'was', 'also', 'partially', 'reversed', 'by', 'naloxone', '.']. with labels ['O', 'Disease', 'O', 'O', 'O', 'O', 'O', 'O', 'Chemical', 'O', 'O', 'O', 'O', 'O', 'Chemical', 'O']. Your task is to predict and return the label for each provided token, ensuring the number of output labels matches the number of input tokens exactly. The output format should include tokens with their labels: ['The-O', 'hypotensive-Disease', 'effect-O', 'of-O', '100-O', 'mg-O', '/-O', 'kg-O', 'alpha-methyldopa-Chemical', 'was-O', 'also-O', 'partially-O', 'reversed-O', 'by-O', 'naloxone-Chemical', '.-O'].  </p>
Basic Prompt	
Description of datasets	<p>The data you are working with is <a href="#">BC5CDR dataset</a>, a benchmark dataset for biomedical natural language processing, created from PubMed abstracts. It includes annotations for two entity types—chemicals and diseases—and their relationships, specifically chemical-induced disease interactions. The dataset is widely used for tasks such as named entity recognition and relation extraction, supporting research in biomedical text mining and information extraction.</p>
High-frequency instances	<p>In this dataset, <a href="#">high frequency of 'Disease'</a> include 'pain', 'toxicity', 'renal', 'failure', 'disease', 'hypotension'; <a href="#">high frequency of 'Chemical'</a> include 'cocaine', 'acid', 'dopamine', 'nicotine', 'morphine', 'lithium'.</p>
UMLS knowledge	<p><a href="#">The Unified Medical Language System (UMLS)</a> is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. <a href="#">You understand medical terminology and concepts from UMLS.</a></p>
Error analysis	<p><b>Possible analysis of prediction errors:</b> The prediction errors mainly stem from challenges in distinguishing between entity boundaries and contextual usage. For instance, multi-token entities were partially labeled, causing boundary mismatches. Additionally, certain terms such as "receptor" or "antagonist" were incorrectly labeled as 'O', despite being part of chemical-related entities. Misclassification also occurred in sentences with conditional phrases or background information, where the relation between entities was not accurately captured. Furthermore, entities mentioned in descriptive or abstract contexts, were sometimes overlooked. These errors highlight difficulties in handling complex sentence structures, context-specific classification, and multi-token entity recognition.</p>

**Table 9.** Specific static prompts for each component we used for the BC5CDR dataset.

Prompt Strategies	MIMIC III
Basic Prompt	<p><b>[Task Description]:</b> You are a medical AI trained to identify and classify tokens into 13 categories: 'CONDITION/SYMPТОМ', 'DRUG', 'AMOUNT', 'TIME', 'MEASUREMENT', 'LOCATION', 'EVENT', 'FREQUENCY', 'ORGANIZATION', 'DATE', 'AGE', 'GENDER' and Outside ('O'). Your task is to extract and classify the concepts from this dataset.</p> <p><b>[Entity Types with Definitions]:</b> 'ORGANIZATION' refers to entities or groups associated with healthcare or emergency medical services. These could be specific departments, teams, or services within a medical or emergency response organization. 'DATE' in this context refers to specific calendar dates. These dates are typically used to mark particular events, appointments, or deadlines. 'AGE' in this context refers to the length of time that a person has lived or the number of years since their birth. It can be expressed in various formats, including numerical values, abbreviated forms, or written out in words. 'GENDER' in this context refers to the socially constructed roles, behaviors, activities, and attributes that a given society considers appropriate for men and women. It encompasses the identities of 'male' and 'female,' which are often associated with biological sex but are also shaped by cultural and social factors. 'FREQUENCY' in this context refers to the rate or regularity at which an event or phenomenon occurs. It can describe how often something happens, ranging from sporadic or irregular occurrences to more regular or constant patterns. 'EVENT' in this context refers to specific occurrences or actions that take place, particularly in a medical or clinical setting. These can include procedures, assessments, or other significant incidents. 'LOCATION' in this context refers to specific places or areas, particularly within a healthcare or medical setting. These can include types of facilities, specific locations within a facility, or other relevant places. 'MEASUREMENT' in this context refers to quantitative assessments or values used to evaluate specific physiological or medical parameters. These can include vital signs, laboratory test results, numerical values, or other metrics related to patient health. 'TIME' in this context refers to specific points or periods in the temporal continuum, particularly as they relate to healthcare or medical events. These can include general time references, specific durations, or events tied to time. 'AMOUNT' in this context refers to specific quantities or dosages, particularly in a medical or pharmaceutical setting. These can include measurements of medication, frequency or number of administrations, and methods of delivery. 'DRUG' in this context refers to specific medications or pharmaceutical substances used in the treatment, prevention, or diagnosis of diseases. These can include brand names, generic names, or forms of administration. 'CONDITION/SYMPТОМ' in this context refers to physical or subjective signs that indicate a medical condition or disease. These can include sensations of discomfort, specific types of pain or discomfort, respiratory issues, or gastrointestinal symptoms. Any token not falling into categories above should be labeled as 'O'.</p> <p><b>[Format Specification]:</b> For example, the sentence "The patient was readmitted to the hospital on 2195-6-6 due to fevers to 103 at the rehabilitation facility despite being on intravenous antibiotics HISTORY OF PRESENT ILLNESS 55 year-old female presents with 2/5 week history of non-bloody diarrhea" is tokenized and labeled as follows: ["The", "patient", "was", "readmitted", "to", "the", "hospital", "on", "2195-6-6", "due", "to", "fevers", "to", "103", "at", "the", "rehabilitation", "facility", "despite", "being", "on", "intravenous", "antibiotics", "HISTORY", "OF", "PRESENT", "ILLNESS", "55", "year-old", "female", "presents", "with", "2/5", "week", "history", "of", "non-bloody", "diarrhea"] with labels ['O', 'O', 'O', 'EVENT', 'O', 'LOCATION', 'LOCATION', 'O', 'O', 'O', 'O', 'MEASUREMENT', 'MEASUREMENT', 'MEASUREMENT', 'O', 'LOCATION', 'LOCATION', 'LOCATION', 'O', 'O', 'O', 'DRUG', 'DRUG', 'O', 'O', 'O', 'O', 'AGE', 'O', 'GENDER', 'O', 'O', 'AMOUNT', 'AMOUNT', 'O', 'O', 'CONDITION/SYMPТОМ', 'CONDITION/SYMPТОМ']. Your task is to predict and return the label for each provided token, ensuring the number of output labels matches the number of input tokens exactly. The output format should include tokens with their labels: ["The-O", "patient-O", "was-O", "readmitted-EVENT", "to-O", "the-LOCATION", "hospital-LOCATION", "on-O", "2195-6-6-O", "due-O", "to-O", "fevers-MEASUREMENT", "to-MEASUREMENT", "103-MEASUREMENT", "at-O", "the-LOCATION", "rehabilitation-LOCATION", "facility-LOCATION", "despite-O", "being-O", "on-O", "intravenous-DRUG", "antibiotics-DRUG", "HISTORY-O", "OF-O", "PRESENT-O", "ILLNESS-O", "55-AGE", "year-old-O", "female-GENDER", "presents-O", "with-O", "2/5-AMOUNT", "week-AMOUNT", "history-O", "of-O", "non-bloody-CONDITION/SYMPТОМ", "diarrhea-CONDITION/SYMPТОМ"].</p>
	<p>The data you are working with is MIMIC-III (Medical Information Mart for Intensive Care) dataset, a large, publicly available database containing de-identified health data from critical care patients at the Beth Israel Deaconess Medical Center. It includes structured data, such as demographics, lab results, and vital signs, as well as unstructured data, such as clinical notes and discharge summaries. The dataset is widely used for research in machine learning, natural language processing, and clinical decision support to improve healthcare outcomes.</p>
	<p>In this dataset, high frequency of 'CONDITION/SYMPТОМ' include 'pain', 'chest', 'cough', 'breath', 'nausea', 'abdominal'; high frequency of 'DRUG' include 'iv', 'lasix', 'ceftriaxone', 'oxygen', 'ns', 'coumadin'; high frequency of 'AMOUNT' include 'iv', '2', '1', 'mg', 'days', 'one'; high frequency of 'TIME' include 'day', 'admission', 'prior', 'last', 'ago', 'morning'; high frequency of 'MEASUREMENT' include 'bp', 'hr', 'pressure', 'blood', 'rr', 'rate', 'heart'; high frequency of 'LOCATION' include 'hospital', 'right', 'home', 'floor', 'emergency', 'micu'; high frequency of 'EVENT' include 'ct', 'placed', 'cxr', 'intubated', 'exam', 'review'; high frequency of 'FREQUENCY' include 'chronic', 'intermittent', 'daily', 'occasionally', 'frequent', 'intermittently'; high frequency of 'ORGANIZATION' include 'ems', 'service', 'surgery', 'pcp', 'emergency', 'neuro', 'medicine'; high frequency of 'DATE' include '2171114', '21491117'; high frequency of 'AGE' include '60', '80yo', '78', '61', 'seventyeightyearold', '69'; high frequency of 'GENDER' include 'man', 'woman', 'f', 'male', 'female', 'm'.</p>
	<p>The Unified Medical Language System (UMLS) is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. You understand medical terminology and concepts from UMLS.</p>
	<p>The prediction errors stem from several factors. Entity boundary recognition issues were common, particularly with multi-token entities like "shortness of breath" or "paroxysmal nocturnal dyspnea," where some tokens were missed or incorrectly labeled as 'O.' Additionally, the model struggled with entity type confusion, such as distinguishing "pain" as a symptom versus its contextual use related to location. Context-dependent misinterpretations also contributed to errors, especially in handling negations like "denies chest pain" or temporal references such as "last few months." Overlapping entities posed further challenges, where closely related terms (e.g., "MI" and "CABG") interfered with accurate classification. Finally, rare or unseen entities in the training data led to occasional misclassifications, highlighting gaps in the model's ability to generalize.</p>

**Table 10.** Specific static prompts for each component we used for the MIMIC III dataset.

Prompt Strategies	NCBI
	<p><b>[Task Description]:</b> You are a medical AI trained to identify and classify tokens into <b>five categories</b>: DiseaseClass, SpecificDisease, Modifier, CompositeMention and Outside ('O'). Your task is to extract and classify the DiseaseClass, SpecificDisease, Modifier and CompositeMention from this dataset.</p> <p><b>[Entity Types with Definitions]:</b> 'DiseaseClass' refers to a classification system or category used to group various medical conditions or diseases based on certain characteristics, such as their nature, affected biological systems, or underlying causes. 'SpecificDisease' appears to describe particular diseases that are identified and classified based on their specific clinical features, genetic origins, or biochemical abnormalities. 'Modifier' refers to specific attributes or variations or conditions that can modify or influence the presentation, progression, or characteristics of a disease, alter the manifestation or course of a disease, potentially affecting its diagnosis, treatment, and prognosis. 'CompositeMention' describes medical conditions or characteristics that are composed of several elements or features, often involving multiple tissues, organs, or systems. Any token not falling into these categories should be labeled as 'O'.</p> <p><b>[Format Specification]:</b> For example, the sentence 'Histidinemia. Classical and atypical form in siblings.' is tokenized and labeled as follows: ['Histidinemia.', 'Classical', 'and', 'atypical', 'form', 'in', 'siblings.']. with labels ['SpecificDisease', 'O', 'O', 'O', 'O', 'O', 'O']. Your task is to predict and return the label for each provided token, ensuring the number of output labels matches the number of input tokens exactly. The output format should include tokens with their labels: ['Histidinemia.-SpecificDisease', 'Classical-O', 'and-O', 'atypical-O', 'form-O', 'in-O', 'siblings.-O'].</p>
Basic Prompt	
Description of datasets	<p>The data you are working with is NCBI disease corpus, a collection of 793 PubMed abstracts fully annotated at the mention and concept level to serve as a research resource for the biomedical natural language processing community. Each PubMed abstract was manually annotated by two annotators with disease mentions and their corresponding concepts in Medical Subject Headings (MeSH) or Online Mendelian Inheritance in Man (OMIM). The public release of the NCBI disease corpus contains 6892 disease mentions, which are mapped to 790 unique disease concepts. Of these, 88 percent link to a MeSH identifier, while the rest contain an OMIM identifier. We were able to link 91 percent of the mentions to a single disease concept, while the rest are described as a combination of concepts.</p>
High-frequency instances	<p>In this dataset, high-frequency 'DiseaseClass' include 'disorder', 'abnormalities', 'tumors', 'mental', 'disorders', 'retardation'. High-frequency 'SpecificDisease' include 'deficiency', 'syndrome', 'dystrophy', 'familial', 'myotonic', 'colorectal'. High-frequency 'Modifier' include 'tumor', 'tumour', 'APC', 'choroideremia', 'DM', 'DMD'. High-frequency 'CompositeMention' include 'breast', 'ovarian', 'cancer', 'muscular', 'andor', 'becker'.</p>
UMLS knowledge	<p>The Unified Medical Language System (UMLS) is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. You understand medical terminology and concepts from UMLS.</p>
Error analysis	<p>The prediction errors in the NCBI dataset primarily stem from challenges in distinguishing composite mentions and modifiers within complex biomedical contexts. For instance, entities like "BRCA1 gene" were incorrectly segmented, with "BRCA1" labeled as a modifier instead of being part of the composite mention. Additionally, multi-token composite mentions such as "breast and ovarian cancer" were not consistently labeled, with individual tokens occasionally missed or misclassified. Contextual ambiguity, such as distinguishing between mentions of general biological terms (e.g., "tumor") and their specific functional roles (e.g., "tumor suppressor"), also contributed to errors.</p>

**Table 11.** Specific static prompts for each component we used for the NCBI dataset.

Prompt Strategies	Med-Mentions
	<p><b>[Task Description]:</b> You are a medical AI trained to identify and classify tokens into <b>two categories</b>: Disease and Outside ('O'). Your task is to extract and classify the Disease related concepts from this dataset.</p> <p><b>[Entity Types with Definitions]:</b> 'Disease' is a particular abnormal condition that adversely affects the structure or function of all or part of an organism and is not immediately due to any external injury. Diseases are often known to be medical conditions that are associated with specific signs and symptoms. A disease may be caused by external factors such as pathogens or by internal dysfunctions. For example, internal dysfunctions of the immune system can produce a variety of different diseases, including various forms of immunodeficiency, hypersensitivity, allergies, and autoimmune disorders. Any token not falling into Disease categories should be labeled as 'O'.</p> <p><b>[Format Specification]:</b> For example, the sentence 'A total of 200 children and adolescents with type 1 diabetes, ages 9-18 years, completed the DEPS-R Turkish version.' is tokenized and labeled as follows: ['A', 'total', 'of', '200', 'children', 'and', 'adolescents', 'with', 'type', '1', 'diabetes', 'ages', '9-18', 'years', 'completed', 'the', 'DEPS-R', 'Turkish', 'version.']. with labels ['O', 'O', 'O', 'O', 'Disease', 'O', 'Disease', 'O', 'Disease', 'Disease', 'Disease', 'Disease', 'O', 'Disease', 'O', 'O', 'Disease', 'Disease', 'Disease']. The output format should include tokens with their labels: ['A-O', 'total-O', 'of-O', '200-O', 'children-Disease', 'and-O', 'adolescents-Disease', 'with-O', 'type-Disease', '1-Disease', 'diabetes.-Disease', 'ages-Disease', '9-18-O', 'years.-Disease', 'completed-O', 'the-O', 'DEPS-R-Disease', 'Turkish-Disease', 'version.-Disease'].</p>
Basic Prompt	
Description of datasets	<p>The data you are working with is Med-Mentions, a new manually annotated resource for the recognition of biomedical concepts. What distinguishes Med-Mentions from other annotated biomedical corpora is its size (over 4,000 abstracts and over 350,000 linked mentions), as well as the size of the concept ontology (over 3 million concepts from UMLS 2017) and its broad coverage of biomedical disciplines.</p>
High-frequency instances	<p>In this dataset, high frequency 'Disease' related entities include 'patients', 'cells', 'treatment', 'cancer', 'analysis', 'disease', 'clinical'.</p>
UMLS knowledge	<p>The Unified Medical Language System (UMLS) is developed by the U.S. National Library of Medicine (NLM) to integrate and standardize diverse medical terminologies and coding systems. It consists of three main components: the Metathesaurus, Semantic Network, and SPECIALIST Lexicon, supporting medical information retrieval and semantic analysis. You understand medical terminology and concepts from UMLS.</p>
Error analysis	<p>The prediction errors in the Med-Mentions dataset are primarily due to challenges in identifying complex and overlapping disease mentions, as well as distinguishing between general biomedical terms and specific disease entities. Multi-token entities such as "renal pedicle occlusion" or "intention-to-treat analyses" were often partially labeled, with some tokens being misclassified or excluded. Additionally, the presence of nested or overlapping mentions, such as "prostate cancer" and its relationship to broader contexts like "treatment disparities," led to inconsistent labeling. The model also struggled with domain-specific terminology, misclassifying general terms like "maternal genotype" or "outcome" as disease mentions. These errors highlight limitations in handling nuanced biomedical language, especially when entities span multiple tokens or overlap with related terms.</p>

**Table 12.** Specific static prompts for each component we used for the Med-Mentions dataset.