Model	Layer Position	Original			Fine-tuned		
		Max	Mean	Sum	Max	Mean	Sum
SP (Synonym Prediction)							
GPT2-small	1%	99.04%	99.04%	99.04%	59.68%	49.40%	34.68%
	25%	98.56%	98.56%	97.60%	61.09%	30.85%	29.64%
	75%	96.15%	94.23%	93.75%	40.12%	9.68%	10.48%
	100%	6.73%	7.21%	7.21%	3.23%	2.42%	2.42%
GPT2-medium	1%	96.43%	96.43%	96.43%	83.35%	82.50%	84.06%
	25%	96.13%	96.43%	96.43%	79.22%	80.22%	80.79%
	75%	63.93%	48.30%	56.63%	48.36%	23.23%	24.53%
	100%	6.68%	3.41%	3.41%	6.55%	5.12%	5.12%
GPT2-large	1%	98.49%	98.49%	98.06%	78.61%	78.33%	80.17%
	25%	97.63%	97.63%	97.63%	80.93%	81.78%	79.89%
	75%	34.27%	27.59%	28.52%	11.91%	10.02%	10.49%
	100%	1.29%	1.51%	1.51%	1.22%	39.12%	0.61%
Gemma-2B	1%	99.99%	99.80%	83.47%	99.93%	99.15%	96.38%
	25%	99.99%	97.46%	63.68%	90.20%	90.24%	65.82%
	75%	84.63%	60.66%	61.15%	89.87%	75.68%	68.65%
	100%	4.30%	8.69%	8.69%	2.98%	4.57%	4.57%
Llama3-8B	1%	99.99%	99.90%	99.90%	99.99%	99.88%	99.88%
	25%	85.55%	83.50%	82.81%	87.63%	85.75%	85.63%
	75%	53.35%	50.55%	49.77%	31.29%	30.29%	29.91%
	100%	9.28%	9.96%	9.96%	5.20%	5.82%	5.82%
Llama3-3B	1%	100%	100%	100%	100%	100%	100%
	25%	85.81%	86.2%	85.16%	88.47%	84.54%	85.48%
	75%	40.18%	39.3%	38.91%	14.77%	16.48%	15.64%
	100%	5.77%	6.16%	6.16%	5.8%	6.12%	6.12%
Qwen2.5-0.5B	1%	81.77%	88.89%	79.17%	64.24%	58.36%	53.3%
	25%	90.8%	91.15%	86.11%	54.51%	54.38%	37.22%
	75%	63.72%	66.32%	39.06%	48.87%	48.57%	24.29%
	100%	8.51%	10.07%	8.51%	3.67%	3.8%	3.8%
Qwen2.5-1.5B	1%	89.35%	84.52%	84.23%	64.55%	56.79%	56.03%
	25%	90.58%	83.48%	83.19%	60.45%	55.5%	54.79%
	75%	22.06%	22.21%	18.8%	10.88%	10.34%	10.02%
	100%	6.82%	3.55%	3.55%	8.19%	7.87%	7.87%
Qwen2.5-3B	1%	81.39%	81.53%	73.58%	55.93%	49.35%	49.57%
	25%	93.04%	89.91%	82.81%	72.41%	42.78%	38.47%
	75%	77.84%	69.6%	49.43%	43.24%	22.13%	15.25%
	100%	3.98%	3.13%	3.13%	1.4%	1.29%	1.29%

Table 10: Performance drop (in percentage points) for GPT2 (small, medium, large), Gemma-2B, Llama3 (3B, 8B), and Qwen2.5 (0.5B, 1.5B, 3B) models after applying word-level CAP for the Synonym Prediction (SP) task. Results are reported for different layer positions (1%, 25%, 75%, and 100%) in both Original and Fine-tuned settings, using three CAP protocols: Max, Mean, and Sum.

Model	Layer Position	Original			Fine-tuned		
		Max	Mean	Sum	Max	Mean	Sum
HP (Hypernym Prediction)							
GPT2-small	1%	99.75%	99.75%	99.75%	91.19%	91.08%	88.20%
	25%	99.47%	99.29%	98.94%	81.35%	76.76%	72.63%
	75%	95.40%	91.16%	91.32%	48.75%	38.54%	38.40%
	100%	8.12%	6.39%	6.39%	1.35%	1.38%	1.28%
GPT2-medium	1%	99.42%	99.40%	99.44%	93.42%	92.17%	91.69%
	25%	99.11%	98.55%	97.85%	91.64%	86.11%	85.76%
	75%	74.83%	33.22%	41.52%	3.86%	2.23%	2.33%
	100%	4.42%	1.79%	1.79%	3.86%	2.23%	2.32%
GPT2-large	1%	99.27%	99.32%	99.20%	91.49%	90.90%	89.80%
	25%	98.81%	98.75%	98.10%	87.30%	87.54%	84.16%
	75%	45.17%	29.85%	35.66%	7.61%	6.89%	6.22%
	100%	2.14%	0.45%	0.90%	0.69%	0.50%	0.56%
Gemma-2B	1%	99.99%	98.97%	70.22%	99.88%	95.39%	74.03%
	25%	99.98%	90.58%	86.35%	90.98%	73.78%	86.01%
	75%	68.14%	80.06%	80.20%	58.56%	72.57%	66.56%
	100%	5.89%	10.99%	10.99%	1.58%	2.12%	2.12%
Llama3-8B	1%	99.99%	99.99%	99.14%	99.99%	99.10%	99.14%
	25%	80.85%	76.97%	76.81%	72.67%	71.86%	71.40%
	75%	24.43%	24.39%	23.11%	19.65%	19.71%	18.77%
	100%	3.83%	4.49%	4.49%	4.63%	4.04%	4.20%
Llama3-3B	1%	100%	99.95%	99.95%	99.93%	99.86%	99.82%
	25%	88.04%	83.87%	84.34%	65.53%	63.92%	64.17%
	75%	26.06%	24.47%	23.4%	11.06%	10.52%	10.79%
	100%	4.34%	4.31%	4.31%	3.85%	4.08%	3.86%
Qwen2.5-0.5B	1%	93.76%	90.95%	85.27%	86.33%	80.55%	77.91%
	25%	97.12%	97.51%	89.18%	74.83%	75.41%	75.77%
	75%	76.74%	77.96%	55.39%	50.69%	49.71%	48.81%
	100%	6.15%	5.56%	5.56%	2.48%	2.34%	2.34%
Qwen2.5-1.5B	1%	97.14%	90.5%	88.96%	88.52%	83.19%	77.21%
	25%	98.12%	95.66%	94.04%	72.29%	68.18%	68.33%
	75%	18.27%	18.72%	17.94%	8.94%	9.64%	9.51%
	100%	7.13%	6.81%	6.81%	3.95%	3.8%	3.8%
Qwen2.5-3B	1%	83.26%	82.41%	68.8%	75.13%	72.56%	70.69%
	25%	97.36%	96.32%	88.81%	92.69%	79.67%	79.63%
	75%	86.56%	71.45%	45.47%	40.87%	30.95%	33.04%
	100%	2.07%	1.89%	1.89%	0.45%	0.35%	0.41%

Table 11: Performance drop (in percentage points) for GPT2 (small, medium, large), Gemma-2B, Llama3 (3B, 8B), and Qwen2.5 (0.5B, 1.5B, 3B) models after applying word-level CAP for the Hypernym Prediction (HP) task. Results are reported for different layer positions (1%, 25%, 75%, and 100%) in both Original and Fine-tuned settings, using three CAP protocols: Max, Mean, and Sum.

Model	Layer Position	Original			Fine-tuned		
		Max	Mean	Sum	Max	Mean	Sum
IDM (Inverse Dictionary Modelling)							
GPT2-small	1%	93.00%	93.94%	96.56%	77.912%	77.73%	80.28%
	25%	90.20%	87.85%	91.41%	65.73%	62.95%	72.31%
	75%	87.81%	78.66%	84.90%	55.74%	46.81%	55.73%
	100%	48.10%	45.10%	38.04%	11.11%	8.45%	8.11%
GPT2-medium	1%	87.96%	89.87%	92.52%	81.12%	82.37%	81.83%
	25%	77.06%	82.71%	86.54%	69.53%	75.19%	77.55%
	75%	76.35%	48.76%	57.68%	60.60%	29.52%	33.12%
	100%	29.23%	23.12%	23.21%	13.03%	9.75%	9.94%
GPT2-large	1%	87.06%	89.91%	88.44%	81.14%	85.35%	79.46%
	25%	73.54%	78.18%	82.48%	69.39%	73.85%	71.90%
	75%	49.02%	42.06%	40.38%	20.59%	19.78%	21.45%
	100%	28.14%	24.22%	24.78%	6.46%	6.67%	8.44%
Qwen2.5-0.5B	1%	93.97%	91.19%	87.15%	90.94%	84.44%	78.85%
	25%	84.64%	76.78%	78.00%	76.36%	66.24%	67.16%
	75%	61.75%	57.95%	63.86%	48.86%	41.8%	46.25%
	100%	32.29%	26.8%	19.5%	13.55%	10.17%	15.08%
Qwen2.5-1.5B	1%	98.24%	95.8%	95.82%	93.31%	87.33%	80.81%
	25%	96.4%	84.72%	89.41%	79.52%	63.00%	65.53%
	75%	69.68%	64.6%	60.33%	19.11%	14.72%	24.01%
	100%	68.03%	60.04%	56.6%	12.01%	7.46%	12.72%
Qwen2.5-3B	1%	96.51%	94.37%	94.64%	90.11%	86.02%	80.57%
	25%	96.82%	89.89%	92.39%	90.24%	76.55%	76.28%
	75%	82.27%	74.71%	77.07%	47.45%	36.06%	39.95%
	100%	62.26%	62.21%	58.12%	7.41%	5.52%	8.18%

Table 12: Performance drop (in percentage points) for GPT2-small, GPT2-medium, and GPT2-large models after applying phrasal-level CAP across three tasks: Inverse Dictionary Modelling (IDM), Synonym Prediction (SP), and Hypernym Prediction (HP). Results are reported for different layer positions (1%, 25%, 75%, and 100%) in both Original and Fine-tuned settings, using three CAP protocols: Max, Mean, and Sum. Results for Gemma-2B and Llama3-8B are omitted due to severe performance degradation under phrasal-level CAP.

Model	Layer Position	Original			Fine-tuned		
		Max	Mean	Sum	Max	Mean	Sum
SP (Synonym Prediction)							
GPT2-small	1%	99.99%	99.99%	99.99%	64.90%	58.47%	53.22%
	25%	92.97%	93.36%	93.36%	61.27%	37.19%	74.69%
	75%	92.58%	90.63%	92.19%	43.35%	20.57%	52.22%
	100%	58.46%	47.92%	51.43%	13.27%	7.57%	12.45%
GPT2-medium	1%	97.55%	95.11%	99.99%	88.92%	84.23%	84.80%
	25%	97.55%	99.73%	97.55%	75.00%	76.85%	85.65%
	75%	71.20%	68.21%	77.45%	47.72%	22.16%	45.88%
	100%	66.30%	39.40%	52.17%	12.93%	6.68%	9.52%
GPT2-large	1%	96.67%	98.33%	96.67%	92.55%	80.76%	79.58%
	25%	96.67%	96.44%	97.90%	79.44%	80.48%	82.86%
	75%	78.83%	66.72%	66.32%	18.63%	15.80%	21.00%
	100%	67.10%	45.83%	56.68%	9.69%	7.15%	8.33%
Qwen2.5-0.5B	1%	99.32%	95.88%	92.87%	81.67%	61.89%	57.95%
	25%	98.65%	95.91%	96.45%	60.19%	58.75%	58.43%
	75%	93.21%	84.66%	77.4%	56.29%	49.3%	44.94%
	100%	68.78%	45.74%	43.92%	13.56%	7.47%	16.79%
Qwen2.5-1.5B	1%	98.1%	96.33%	94.43%	72.33%	58.5%	59.55%
	25%	97.55%	96.2%	95.38%	63.79%	55.84%	68.93%
	75%	75.72%	55.17%	48.41%	19.33%	14.48%	26.87%
	100%	70.39%	38.68%	36.29%	18.73%	10.41%	20.97%
Qwen2.5-0.5B	1%	96.47%	95.52%	90.31%	74.05%	67.1%	56.57%
	25%	99.32%	98.1%	94.29%	94.89%	56.93%	57.38%
	75%	94.02%	89.46%	83.4%	86.43%	64.01%	43.39%
	100%	47.00%	35.56%	31.32%	20.07%	15.19%	21.15%

Table 13: Performance drop (in percentage points) for GPT2-small, GPT2-medium, and GPT2-large models after applying phrasal-level CAP across three tasks: Inverse Dictionary Modelling (IDM), Synonym Prediction (SP), and Hypernym Prediction (HP). Results are reported for different layer positions (1%, 25%, 75%, and 100%) in both Original and Fine-tuned settings, using three CAP protocols: Max, Mean, and Sum. Results for Gemma-2B and Llama3-8B are omitted due to severe performance degradation under phrasal-level CAP.