| Model | Layer Position | Original | | | Fine-tuned | | |
|---|---|---|---|---|---|---|---|
| | | Max | Mean | Sum | Max | Mean | Sum |
| **HP (Hypernym Prediction)** | | | | | | | |
| **GPT2-small** | 1% | 99.40% | 99.26% | 47.24% | 89.31% | 89.86% | 88.76% |
| | 25% | 99.31% | 98.12% | 46.38% | 77.72% | 73.12% | 76.08% |
| | 75% | 95.63% | 91.78% | 45.57% | 47.73% | 336.59% | 48.32% |
| | 100% | 65.62% | 45.84% | 34.80% | 4.80% | 3.64% | 4.00% |
| **GPT2-medium** | 1% | 99.77% | 99.56% | 99.950% | 92.67% | 90.40% | 92.54% |
| | 25% | 99.92% | 99.35% | 99.47% | 90.38% | 84.29% | 86.84% |
| | 75% | 77.77% | 58.17% | 80.58% | 63.00% | 21.55% | 23.32% |
| | 100% | 59.28% | 27.47% | 30.54% | 8.46% | 5.10% | 5.10% |
| **GPT2-large** | 1% | 99.77% | 99.71% | 99.76% | 91.63% | 92.56% | 88.92% |
| | 25% | 99.82% | 98.72% | 98.82% | 85.31% | 85.35% | 84.58% |
| | 75% | 66.58% | 49.79% | 63.56% | 9.87% | 8.79% | 9.73% |
| | 100% | 35.57% | 24.79% | 26.69% | 6.99% | 5.05% | 4.82% |
| **Qwen2.5-0.5B** | 1% | 99.06% | 97.77% | 92.97% | 94.46% | 81.39% | 79.64% |
| | 25% | 99.85% | 98.54% | 96.95% | 75.14% | 76.07% | 86.94% |
| | 75% | 94.87% | 87.81% | 88.37% | 56.27% | 53.09% | 63.33% |
| | 100% | 68.71% | 27.91% | 27.92% | 10.6% | 7.68% | 15.16% |
| **Qwen2.5-1.5B** | 1% | 99.81% | 97.07% | 92.75% | 90.34% | 84.61% | 78.76% |
| | 25% | 99.64% | 97.97% | 96.98% | 72.81% | 68.48% | 77.13% |
| | 75% | 84.28% | 47.63% | 43.15% | 17.12% | 14.76% | 28.18% |
| | 100% | 82.22% | 26.00% | 27.7% | 13.49% | 9.08% | 17.98% |
| **Qwen2.5-3B** | 1% | 93.95% | 91.81% | 82.05% | 77.6% | 73.86% | 71.41% |
| | 25% | 99.24% | 98.54% | 95.97% | 93.6% | 80.32% | 80.77% |
| | 75% | 94.48% | 88.91% | 78.88% | 54.32% | 38.19% | 57.87% |
| | 100% | 55.28% | 27.4% | 25.1% | 15.1% | 8.77% | 13.77% |

Table 14: 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.