**Results.** The evaluation yielded the following averaged metrics across five seeds for the default level of parsing (Level 1, the immediate children of the root node):

| Metric | Mean ± Std |
|---|---|
| Precision | 0.956 ± 0.001 |
| Recall | 0.956 ± 0.001 |
| F1-Score | 0.956 ± 0.001 |
| Accuracy | 0.956 ± 0.001 |

Table 6: Aggregated evaluation metrics for Level 1 constituents using the Benepar parser, averaged across five seeds.

**Interpretation.** The results demonstrate consistently high parsing accuracy across all evaluation metrics, with minimal variability (as indicated by the low standard deviation). These findings validate the Benepar parser's reliability for parsing Level 1 constituents, which form the backbone of sentence structure. Consequently, the parser's impact on CAP results is minimal, ensuring robustness and validity of our conclusions.

# E Detailed Performance Evaluation and Results

## E.1 Baseline Performance

Table 4 summarises the baseline performance of the models used in this paper on the three tasks. The results include the accuracy of each model's original and FT versions on the test set described in Table 2. Fine-tuning generally improves performance, particularly in the larger models such as Gemma-2B and Llama3-8B, which show notable increases in accuracy in most tasks, except the IDM task.

## E.2 Qualitative Analysis of CAP-Induced Prediction Shifts.

Tables 7 and 8 present representative examples of predictions from multiple models across all the tasks, before and after CAP is applied. These examples are drawn from inputs that the model originally predicted correctly, allowing us to isolate the effects of CAP perturbations without confounding them with unrelated model failures. Each example specifies the CAP layer, CAP type (token-to-word or token-to-phrase), and the model involved. Table 7 focuses on predictions made by original (non-fine-tuned) models, while Table 8 includes outputs from fine-tuned variants. Observed shifts include truncation of multi-token terms (e.g., "diary" $\rightarrow$ "di"), polarity inversion (e.g., "plain" $\rightarrow$ "ornament"), loss of abstraction ("polygon" $\rightarrow$ "plane"), and domain misalignment (e.g., "tree" $\rightarrow$ "street").

These qualitative differences provide interpretability insights that complement the aggregate metrics reported earlier. They reveal how CAP affects not only performance but the nature of model outputs, especially in terms of semantic generalisation, abstraction shifts, and lexical precision. While we do not observe a uniform trend across layers or model families, TP-CAP consistently induces more severe semantic degradation. This suggests that as model capacity increases, internal representations may become more sensitive to disruptions from externally imposed syntactic structures, potentially due, as argued in the main paper, to a stronger reliance on learned token-level dependencies that diverge from higher-level compositional groupings. This analysis highlights the nature of semantic and lexical shifts induced by CAP, reinforcing the need for future task-specific fine-tuning strategies that improve robustness to structured representation pooling.

## E.3 Comprehensive CAP Results for All Models and Tasks

Figure 3 and Tables 9- 13, and 14 present the reduction in accuracy when applying word-level and phrasal CAP, respectively, across models and the three tasks: IDM, SP, and HP. The results of phrasal-level CAP for Gemma-2B and Llama3-8B are not reported due to the severe degradation in model performance under these conditions, rendering the outputs effectively unusable.

Let $A_o$ represent the original accuracy and $A_c$ represent the accuracy after applying CAP. The reported drop in accuracy, $\Delta A$, is calculated as:

$$\Delta A = A_o - A_c \tag{12}$$

This $\Delta A$ value is expressed in percentage points. For example, $\Delta A = 40$ indicates that the model's accuracy has decreased by 40 percentage points from its original performance, which could represent a change from $A_o = 100\%$ to $A_c = 60\%$, or any other pair of accuracies with a 40 percentage point difference. The tables report $\Delta A$ for different layer positions (1%, 25%, 75%, and 100%) in both Original and Fine-tuned settings, using three CAP protocols: Max, Mean, and Sum. This representation allows for a direct comparison of CAP's impact across different models and tasks, independent of their baseline performance levels.

| Task / Input Prompt | Model | CAP Layer (Type) | Prediction (No CAP) | Prediction (W/ CAP) | Observation / Interpretation |
|---|---|---|---|---|---|
| IDM: *lacking embellishment or ornamentation is called a: "* | Qwen2.5-1.5B | Layer 8 (TW) | *plain* | *ornament* | Prediction shifts from correct to antonymic, likely due to token merging altering polarity. |
| IDM: *remaining after all deductions is called a: "* | LLaMA3.1-8B | Layer 4 (TW) | *net* | *gain* | Subtle financial distinction lost; CAP causes confusion between output and intermediate step. |
| IDM: *make an effort or attempt is called a:"* | Gemma-2B | Layer 1 (TP) | *try* | *<h1>* | Invalid token generation suggests breakdown in early compositional encoding. |
| IDM: *a formal series of statements showing that if one thing is true something else necessarily follows from it is called a:"* | GPT2-L | Layer 24 (TP) | *proof* | *form* | Loss of logical structure leads to a more abstract or vague concept. |
| SP: *"journal" is a synonym of* | Qwen2.5-1.5B | Layer 18 (TW) | *diary* | *di* | Output truncated, likely due to disruption in longer multi-token word embedding. |
| SP: *"get" is a synonym of* | Qwen2.5-0.5B | Layer 16 (TW) | *catch* | *break* | Semantic drift under CAP; verb meaning shifts from acquisition to interruption. |
| HP: *"voice" is a type of* | Gemma1-2B | Layer 16 (TW) | *sound* | *noise* | Precision reduced; CAP causes substitution with a noisier, less neutral concept. |
| HP: *"guama" is a type of* | LLaMA3.2-3B | Layer 12 (TW) | *tree* | *street* | The output reflects a contextual domain shift, likely due to token-level confusion post-CAP. |

Table 7: Representative examples of model predictions with and without CAP applied at various layers. Examples highlight semantic degradation and conceptual drift caused by TW-CAP or TP-CAP applied to original models.

| Task / Input Prompt | Model | CAP Layer (Type) | Prediction (No CAP) | Prediction (W/ CAP) | Observation / Interpretation |
|---|---|---|---|---|---|
| IDM: *prepare for eating by applying heat is called a: "* | GPT2-S | Layer 4 (TW) | *cook* | *heat* | CAP leads to a shift from action to cause, indicating surface-level generalisation. |
| IDM: *fail to attend an event or activity is called a: "* | LLaMA3.2-3B | Layer 1 (TW) | *miss* | *catch* | CAP appears to invert the meaning, suggesting confusion in early compositional buildup. |
| IDM: *general term for any insect or similar creeping or crawling invertebrate is called a:"* | Gemma-2B | Layer 11 (TP) | *bug* | *un* | Invalid token generation suggests breakdown in compositional encoding |
| IDM: *an institution of higher education created to educate and grant degrees often a part of a university is called a:"* | GPT2-S | Layer 1 (TP) | *college* | *regular* | CAP reduces specificity, misclassifying to a generic adjective. |
| SP: *"one fourth" is a synonym of* | Gemma1-2B | Layer 10 (TW) | *fourth* | *half* | CAP merges related quantities but loses precision, leading to broader but incorrect substitution. |
| HP: *"hotel" is a type of* | Qwen2.5-3B | Layer 16 (TW) | *building* | *room* | Shift from category to subcomponent suggests CAP disrupted higher-level abstraction. |
| HP: *"hexagon" is a type of* | Qwen2.5-3B | Layer 16 (TW) | *polygon* | *plane* | Hierarchical class (shape) replaced by domain (geometry); abstraction misaligned. |

Table 8: Representative examples of model predictions with and without CAP applied at various layers. Each example shows the prompt, model, CAP configuration (layer and type), predictions, and qualitative interpretation. All examples applied to fine-tuned (FT) models.

| Model | Layer Position | Original | | | Fine-tuned | | |
|---|---|---|---|---|---|---|---|
| | | Max | Mean | Sum | Max | Mean | Sum |
| **IDM (Inverse Dictionary Modelling)** | | | | | | | |
| **GPT2-small** | 1% | 4.76% | 4.44% | 4.69% | 8.04% | 7.72% | 7.22% |
| | 25% | 3.09% | 2.74% | 3.26% | 5.87% | 5.85% | 6.24% |
| | 75% | 2.64% | 2.36% | 2.74% | 2.72% | 2.47% | 2.35% |
| | 100% | 1.43% | 1.24% | 1.24% | 0.46% | 0.39% | 0.39% |
| **GPT2-medium** | 1% | 16.75% | 16.36% | 13.77% | 24.51% | 12.70% | 7.44% |
| | 25% | 6.73% | 5.692% | 6.22% | 5.04% | 4.84% | 5.36% |
| | 75% | 18.61% | 2.13% | 2.89% | 11.79% | 2.09% | 1.72% |
| | 100% | 1.58% | 0.41% | 0.41% | 2.27% | 1.29% | 1.29% |
| **GPT2-large** | 1% | 8.06% | 9.15% | 6.70% | 10.61% | 10.01% | 7.83% |
| | 25% | 5.19% | 4.94% | 5.63% | 6.25% | 5.77% | 6.32% |
| | 75% | 5.28% | 2.62% | 2.39% | 3.66% | 1.62% | 0.88% |
| | 100% | 0.84% | 0.12% | 0.19% | 0.22% | 0.16% | 0.16% |
| **Gemma-2B** | 1% | 97.91% | 23.51% | 23.75% | 57.58% | 22.70% | 21.99% |
| | 25% | 86.32% | 16.20% | 19.27% | 50.45% | 14.08% | 15.57% |
| | 75% | 52.38% | 31.03% | 24.74% | 21.77% | 14.99% | 12.80% |
| | 100% | 6.87% | 10.61% | 10.61% | 2.21% | 2.05% | 2.05% |
| **Llama3-8B** | 1% | 25.49% | 24.99% | 24.94% | 24.44% | 23.42% | 23.48% |
| | 25% | 20.02% | 5.87% | 5.74% | 8.81% | 6.03% | 5.92% |
| | 75% | 7.31% | 3.40% | 3.54% | 5.16% | 3.47% | 3.29% |
| | 100% | 2.80% | 1.77% | 1.77% | 1.55% | 1.33% | 1.33% |
| **Llama3-3B** | 1% | 28.79% | 26.36% | 25.96% | 25.54% | 22.71% | 22.74% |
| | 25% | 31.73% | 8.08% | 6.99% | 13.44% | 5.84% | 5.8% |
| | 75% | 12.27% | 5.84% | 5.22% | 8.54% | 5.03% | 5.15% |
| | 100% | 3.62% | 1.99% | 1.99% | 2.37% | 1.82% | 1.85% |
| **Qwen2.5-0.5B** | 1% | 10.12% | 8.2% | 8.23% | 7.85% | 6.39% | 6.00% |
| | 25% | 5.19% | 4.21% | 4.45% | 4.35% | 3.29% | 3.49% |
| | 75% | 3.56% | 2.82% | 3.15% | 2.39% | 2.24% | 2.15% |
| | 100% | 0.98% | 0.98% | 0.98% | 0.23% | 0.28% | 0.33% |
| **Qwen2.5-1.5B** | 1% | 14.56% | 11.04% | 10.22% | 9.47% | 7.36% | 7.48% |
| | 25% | 13.29% | 4.45% | 5.34% | 6.83% | 3.86% | 4.00% |
| | 75% | 7.03% | 2.68% | 2.84% | 4.21% | 2.74% | 2.79% |
| | 100% | 0.7% | 0.4% | 0.4% | 0.65% | 0.23% | 0.23% |
| **Qwen2.5-3B** | 1% | 12.63% | 12.27% | 11.44% | 7.85% | 6.71% | 6.48% |
| | 25% | 18.61% | 8.59% | 9.11% | 10.66% | 4.75% | 5.82% |
| | 75% | 7.23% | 4.00% | 3.79% | 3.65% | 2.83% | 2.8% |
| | 100% | 0.39% | 0.4% | 0.4% | 0.31% | 0.17% | 0.2% |

Table 9: 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 Inverse Dictionary Modelling (IDM) 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.