Entity Tracking in Small Language Models: An Attention-Based Study of Parameter-Efficient Fine-Tuning

Sungho Jeon
Samsung Electronics
s4.jeon@samsung.com*

Michael Strube
Heidelberg Institute of Theoretical Studies
michael.strube@h-its.org

Abstract

The ability to track entities is fundamental for language understanding, yet the internal mechanisms governing this capability in Small Language Models (SLMs) are poorly understood. Previous studies often rely on indirect probing or complex interpretability methods, leaving a gap for lightweight diagnostics that connect model behavior to performance. To bridge this gap, we introduce a framework to analyze entity tracking by measuring the attention flow between entity and non-entity tokens within SLMs. We apply this to analyze models both before and after Parameter-Efficient Fine-Tuning (PEFT). Our analysis reveals two key findings. First, SLMs' attentional strategies vary significantly with text type, but entities consistently receive a high degree of focus. Second, we show that PEFT – specifically QLoRA – dramatically improves classification performance on entity-centric tasks by increasing the model's attentional focus on entity-related tokens. Our work provides direct evidence for how PEFT can refine a model's internal mechanisms and establishes attention analysis as a valuable, lightweight diagnostic tool for interpreting and improving SLMs[1].

1 Introduction

A fundamental aspect of natural language understanding is the ability to track entities as a discourse unfolds (Grosz et al., 1995). This ability is a prerequisite for maintaining coherence, performing complex reasoning, and succeeding in a wide array of downstream natural language processing (NLP) tasks. For language models to generate coherent text or answer questions accurately, they must implicitly recognize entities and update their states based on new information (Grosz and Sidner, 1986).

Despite the remarkable capabilities demonstrated by modern Large and Small Language Models (LLMs and SLMs) (Brown et al., 2020), the internal mechanisms by which these models manage and track entities remain largely unexplained (Li et al., 2024), especially in SLMs, which are often deployed for efficiency and on-device AI. These models are often treated as "black boxes". While SLMs may replicate human-like output behavior, it is not clear whether they rely on linguistically grounded cues—such as noun phrases—or whether their performance stems from spurious correlations learned during pretraining.

Efforts to interpret model behavior typically fall into one of three categories: (i) evaluating input-output behaviors on benchmark tasks (Schuster and Linzen, 2022; Kim and Schuster, 2023), (ii) probing hidden state representations to see if they encode entity information (Loáiciga et al., 2022), or (iii) modifying architectures to better handle discourse entities (Fagnou et al., 2024). While these approaches provide valuable insights, they often leave a gap. They either do not directly inspect the internal mechanisms of standard architectures or they require complex, computationally intensive analysis. A direct, lightweight method for analyzing how the native attention mechanism facilitates entity tracking, particularly in the widely used Transformer architecture, is less explored.

This paper addresses this gap by proposing a novel framework to investigate entity tracking through the lens of attention scores. We treat attention as a direct, interpretable signal of the model's focus during processing (Section 2.2). Our central hypothesis is that the allocation of attention to entity tokens is a direct correlate of a model's entity tracking capability and that performance improvements from fine-tuning can be explained by specific, measurable shifts in these attention patterns. By systematically analyzing the attention scores between entity tokens and their surrounding

*This work was conducted while Sungho Jeon was at Heidelberg Institute of Theoretical Studies.

[1]Our code is available at https://github.com/sdeva14/codi25_entity_attn_tracking_slm

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 42–53

context, we aim to build a mechanistic bridge between an observable performance change and an internal model behavior.

This research makes the following contributions:

- A systematic analysis of entity-centric attention patterns in several modern SLMs, revealing how attentional strategies adapt to different text types and qualities.
- A key finding that Parameter-Efficient Fine-Tuning (PEFT) with QLoRA (Dettmers et al., 2023) substantially improves performance on an entity-centric classification task by mechanistically intensifying the model's attention on entity tokens.
- A demonstration of attention analysis as a valuable and accessible diagnostic tool for understanding and explaining the effects of fine-tuning on a model's internal mechanisms.

2 Related Work

Our work is situated at the intersection of three active research areas: entity tracking in language models, the use of attention for interpretability, and the mechanistic understanding of fine-tuning.

2.1 Probing Entity Representations in Language Models

The study of how language models manage entities has evolved from linguistic tests to sophisticated analyses of internal model states. Earlier work identified significant challenges, showing that even large models struggle with fundamental aspects of discourse, such as recognizing when a new entity is introduced (Schuster and Linzen, 2022). Subsequent research shifted from model outputs to internal representations, finding a disconnect between a model's latent knowledge of entities and its ability to apply it effectively (Loáiciga et al., 2022). More recent work has created benchmarks to test dynamic entity tracking, discovering that this ability can be taught via fine-tuning (Kim and Schuster, 2023). Other studies propose architectural changes to better handle dynamic entity tracking (Fagnou et al., 2024). Unlike these prior approaches, our framework interprets entity tracking behavior through the model's native attention weights, which directly reflect token-level interactions in Transformer models.

2.2 Attention as an Interpretability Tool

The attention mechanism, introduced as the core component of the Transformer architecture, was initially proposed as a window into the model's reasoning process. Early work suggested that visualizing attention weights could serve as a proxy for interpreting model decisions. However, this view was contested by a line of research arguing that "attention is not explanation" (Jain and Wallace, 2019; Serrano and Smith, 2019). These studies demonstrated that attention weights could be manipulated without significantly affecting model output, suggesting they might be a symptom of the model's reasoning rather than its cause.

Nevertheless, more recent work has revealed that specific attention heads often specialize in meaningful linguistic functions, including syntactic relations and coreference resolution (Clark et al., 2019). This suggests that attention, when interpreted systematically, offers insight into the model's internal processing. Rather than treating attention as a complete explanation, we adopt a pragmatic perspective: we use it as a measurable correlate of focus, with a particular emphasis on how attention is distributed over discourse entities. In doing so, we aim to reconcile the interpretability of attention with its utility as a diagnostic signal.

2.3 Mechanistic Insights into Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2022) enable the adaptation of large pretrained models to specific tasks by modifying only a small subset of parameters. While PEFT methods are valued for their efficiency and scalability, their effect on the internal computations of language models has only recently begun to receive systematic attention.

Recent work attempts to reverse-engineer fine-tuned models using circuit analysis and other mechanistic tools (Wang et al., 2023; Prakash et al., 2024). These studies identify sub-network pathways responsible for specific behaviors, but their analyses are computationally intensive and often require considerable expertise. In contrast, our framework uses attention interactions to trace the effects of PEFT on entity focus directly. We show that LoRA fine-tuning leads to measurable shifts in attention toward entity tokens, which correspond to improved task performance. Our approach is both

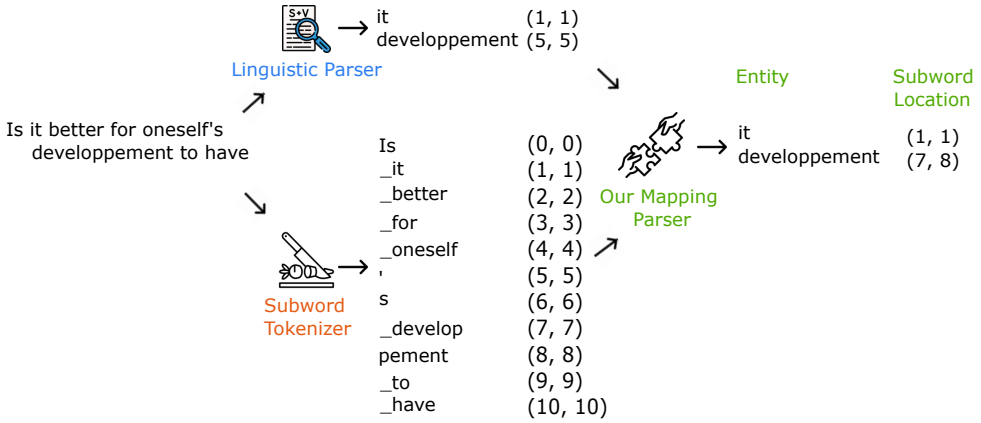

Figure 1: This example illustrates tokenization and mapping to noun phrases using the subword tokenizer of "google/gemma-2-2b-it". The misspelling of "developpement" results in two subword tokens, "develop" and "pement", a phenomenon commonly observed in real-world data.

computationally efficient and grounded in linguistic theory, making it suitable for broader adoption in small-model development and evaluation.

3 A Framework for Analyzing Entity-Centric Attention

In this section, we introduce a lightweight, linguistically motivated framework to analyze how Small Language Models (SLMs) allocate attention to entities during text processing. Our method is grounded in the assumption that coherent language understanding involves selectively focusing on salient discourse elements—primarily noun phrases—while integrating relevant context. We capture this behavior by systematically quantifying attention flows between entity and non-entity tokens.

3.1 Identifying and Mapping Entity Tokens

A primary challenge in analyzing the internal processing of linguistic phenomena is the discrepancy between human-readable words or phrases and the subword tokens that models actually operate on (Table 5). To bridge this gap, our framework employs a two-stage mapping process: first identifying linguistic units, then mapping them to model tokens.

3.1.1 Noun Phrase Extraction

For the purposes of this study, we define an "entity" as a noun phrase. This simplification provides a consistent and scalable method for identifying key subjects and objects across a large corpus. We use the Stanza constituency parser[2], which segments input texts into syntactic constituents and extracts noun phrases based on their syntactic labels. We impose a constraint that limits the length of extracted noun phrases to a maximum of four words to reduce structural complexity and exclude deeply nested constructions. In cases of nested noun phrases, we retain only the outermost phrase to maintain consistent granularity across analyses.

3.1.2 Tokenization and Mapping

Once noun phrases are identified, we must align them with the subword tokens generated by the target model's tokenizer. This alignment is a non-trivial task, as tokenization schemes like Byte-Pair Encoding (BPE) can fragment single words and handle whitespace in model-specific, often proprietary ways. To address this, we developed a mapping algorithm which uses the character-level spans of each noun phrase to identify all subword tokens within its boundaries. As illustrated in Figure 1, this process creates a definitive mapping from each linguistic entity to a set of token indices, a critical step that enables our subsequent analysis of attention flow.

3.2 Quantifying Attention Flow Across Linguistic Boundaries

With entities mapped to tokens, we can now quantify how the model allocates attention with respect to these linguistic categories (Figure 2).

3.2.1 Attention Score Extraction

We extract attention values from the final layer of the model's Transformer architecture. This layer is chosen because it represents the culmination of the model's processing, where representations are expected to be the most semantically rich and task-

[2] https://stanfordnlp.github.io/stanza/

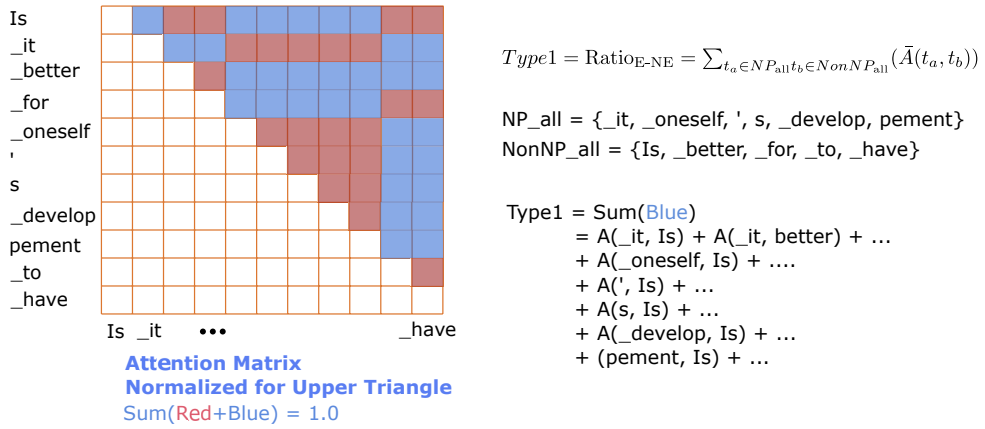

Figure 2: Example of calculating Attention Type 1: between entities and non-entities. The word "developpement" is a typo found in a real TOEFL dataset, and it causes the subword tokenizer to split it into multiple subword tokens.

relevant. While individual attention heads may specialize in different functions (Clark et al., 2019), we average the attention scores across all heads in the final layer to obtain a holistic measure of the model's aggregate focus. This provides a robust, high-level signal of information flow. The raw attention scores are normalized via softmax for each query token. The final averaged attention score between a query token t_a and a key token t_b is calculated as:

$$\bar{A}_{L_{last}}(t_a, t_b) = \frac{1}{|H|} \sum_{h \in H} A_{L_{last},h}(t_a, t_b) \quad (1)$$

where L_{last} denotes the last layer, H is the set of all attention heads, and $A_{L_{last,h}}(t_a, t_b)$ is the attention score from token t_a to token t_b in the last layer and head h. Additionally, we investigate the different attention interaction patterns across various layers in our evaluation to provide a more comprehensive understanding.

3.3 Analysis of Attention Score Interactions

Using the extracted attention values and the LLM tokens that match noun phrases, we measure three distinct types of interactions. This helps us understand how the LLM processes context. For each interaction type, we define which tokens are involved and how we combine their attention scores.

In terms of formulation, for any input text, let N be the total number of LLM tokens: $T = \{t_1, t_2, ..., t_N\}$. Let NP_k be the LLM tokens for the k-th noun phrase: $NP_k = \{t_{i_1}, t_{i_2}, ..., t_{i_m}\}$. Let NP_{all} be the set of all tokens that are part of any noun phrase (entity): $NP_{\text{all}} = \bigcup_k NP_k$. Let $NonNP_{all}$ be the set of all tokens that are not part of any noun phrase (non-entities): $NonNP_{\text{all}} = T \setminus NP_{\text{all}}$. The attention score from token t_a to token t_b is written as $\bar{A}_{L_{last}}(t_a, t_b)$. As detailed in Section 3.1, we focus on the last layer.

When investigating the interactions between different tokens, we focus on unique pairs of elements, effectively excluding self-attention (diagonal elements) and avoiding duplicate pairs (e.g., considering (t_a, t_b) and (t_b, t_a) as a single interaction). This is conceptually equivalent to considering only the upper triangle of the attention matrix and summing the attention for each unique pair. Each interaction type captures the ratio of attention taken by these specific pairs of tokens, normalized by the total amount of attention values between all distinct pairs of tokens in the sequence. Let $Attn\_Total$ be the sum of all attention values between distinct token pairs in the sequence, considering both directions for each unique unordered pair:

$$\text{Attn\_Total} = \sum_{t_x \in T} \sum_{t_y \in T, t_x < t_y} (\bar{A}(t_x, t_y)) \quad (2)$$

Our analysis focuses on three specific types of interactions: 1) between entities and non-entities, 2) between tokens of entities, and 3) between tokens of non-entities. This structured approach allows us to isolate and quantify specific linguistic phenomena, providing insights into how LLMs encode and leverage different types of relationships.

3.3.1 Type 1: Between entities and non-entities

This quantifies the attention flow between any subword token identified as part of an entity and any subword token identified as a non-entity. This captures how entities interact with their broader non-entity context within the sentence.

We calculate the average attention where tokens are from NP_{all} and tokens are from $NonNP_{all}$, then normalize by $TotalAttention$.

$$\text{Ratio}_{\text{E-NE}} = \frac{1}{\text{TotalAttention}} \times \sum_{\substack{t_a \in NP_{\text{all}} \\ t_b \in NonNP_{\text{all}}}} (\bar{A}(t_a, t_b)) \quad (3)$$

3.3.2 Type 2: Between tokens of entities

This measures the ratio of attention among subword tokens within the collective set of all entities, relative to the total attention in the sequence. It reflects the internal coherence and interconnectedness of all identified entities in the text.

We calculate the sum of attention between distinct tokens within NP_{all}, considering both directions for each unique unordered pair, then normalize by $TotalAttention$.

$$\text{Ratio}_{\text{E-E}} = \frac{1}{\text{TotalAttention}} \times \sum_{\substack{t_a \in NP_{\text{all}} \\ t_b \in NP_{\text{all}}, t_a < t_b}} (\bar{A}(t_a, t_b)) \quad (4)$$

3.3.3 Type 3: Between tokens of non-entities

This quantifies the ratio of attention among subword tokens within the collective set of all non-entities, relative to the total attention in the sequence. It reflects the internal coherence and interconnectedness of the non-entity context.

We calculate the sum of attention between distinct tokens within $NonNP_{all}$, considering both directions for each unique unordered pair, then normalize by TotalAttention.

$$\text{Ratio}_{\text{NE-NE}} = \frac{1}{\text{TotalAttention}} \times \sum_{\substack{t_a \in NonNP_{\text{all}} \\ t_b \in NonNP_{\text{all}}, t_a < t_b}} (\bar{A}(t_a, t_b)) \quad (5)$$

4 Experimental Setup

We evaluate our attention-based analysis framework in the context of two representative classification tasks using Small Language Models (SLMs). Our goal is to examine how SLMs allocate attention over entity and non-entity tokens across different discourse settings and how this distribution changes under Parameter-Efficient Fine-Tuning (PEFT). This section describes the datasets, models, and evaluation metrics used in our experiments.

4.1 Datasets

To ensure generalizability across different textual domains and discourse structures, we select two datasets that differ markedly in length, coherence structure, and task type.

- **SST-5 (Stanford Sentiment Treebank):** A benchmark for fine-grained sentiment analysis, consisting of 11,855 individual movie review sentences[3] (Socher et al., 2013). The task involves assigning one of five sentiment labels: "very negative" to "very positive". These short texts typically contain a small number of entities, often representing film titles or actors. Thus, SST-5 enables us to analyze attention patterns when entity information is concentrated in compact, sentiment-focused utterances.

- **TOEFL11:** A dataset for proficiency-level classification, composed of essays written by English language learners (Blanchard et al., 2013). Each essay is labeled with a language proficiency score (low, medium, or high). With an average length of over 400 words, the dataset provides a setting for analyzing long-form discourse. The essays include multiple entities and exhibit varied discourse organization, making it suitable for studying attention flow over extended contexts.

4.2 LLM Models for Evaluation

We perform our experiments on a representative set of modern, instruction-tuned SLMs available via the HuggingFace Hub. These models vary in parameter size, tokenizer behavior, and pretraining objectives, offering a diverse testbed for our attention analysis:

- google/gemma-2-2b-it
- meta-llama/Llama-3.2-1B-Instruct
- meta-llama/Llama-3.2-3B-Instruct
- microsoft/Phi-3.5-mini-instruct
- Qwen/Qwen2.5-1.5B-Instruct

To evaluate the effects of fine-tuning, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022), a parameter-efficient tuning technique. LORA introduces trainable low-rank matrices into the model's

[3] https://huggingface.co/datasets/SetFit/sst5

attention projections while keeping the original model weights frozen. We fine-tune each model on the SST-5 dataset using LoRA and analyze the resulting changes in both performance and attention allocation.

Hyperparameters used during fine-tuning (e.g., rank, learning rate, and epochs) are listed in Appendix A.2. All fine-tuning experiments are conducted using a consistent setup across models to ensure comparability.

4.3 Evaluation Metrics

We assess our framework using both interpretability metrics derived from attention interactions and standard performance metrics for classification.

- Attention Analysis: The core of our interpretability analysis relies on the three attention interaction ratios ($\text{Ratio}_{\text{E-NE}}$, $\text{Ratio}_{\text{E-E}}$, and $\text{Ratio}_{\text{NE-NE}}$) defined in Section 3.3. These values quantify the internal focus of the model and allow us to track systematic shifts in attention behavior across datasets and tuning conditions.

- Classification Performance: We measured model performance on the SST-5 test set using standard metrics for multi-class classification: Accuracy, Linear Weighted Kappa (κ_L), and Quadratic Weighted Kappa (κ_Q). Kappa scores are particularly important as they correct for agreement that could occur by chance and are sensitive to the ordinal nature of the sentiment labels (e.g., misclassifying "positive" as "very positive" is less of an error than misclassifying it as "negative").

5 Evaluations

Our evaluation proceeds in three stages. First, we analyze baseline attention patterns in SLMs across different textual domains. Second, we examine how attention patterns vary with text granularity and writing quality. Finally, we investigate the impact of Parameter-Efficient Fine-Tuning (PEFT) using LoRA on both performance and attention allocation. Our findings demonstrate that entity-centric attention is a consistent and informative signal for tracking discourse focus and that LoRA fine-tuning meaningfully enhances this behavior.

For SST-5, we treat each review as a single unit, as reviews are typically single sentences. For TOEFL11, we analyze each sentence independently rather than encoding entire essays, allowing us to capture fine-grained variations in local attention and entity focus across discourse units.

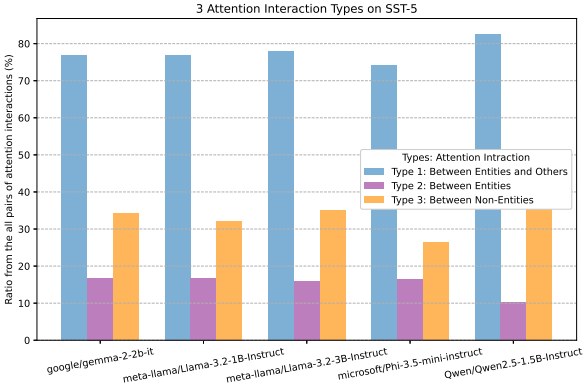

Figure 3: Attention allocation in pre-trained SLMs on the short-text SST-5 dataset. Entity-related interactions (Type 1) dominate.

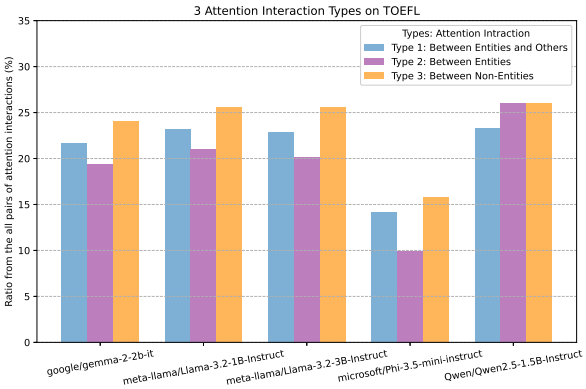

Figure 4: Attention allocation in pre-trained SLMs on the long-text TOEFL dataset. Attention is more distributed compared to SST-5.

5.1 Entity Focus Depends on Text Length and Quality

We first examine how pretrained SLMs allocate attention scores across entity and non-entity tokens in two different textual settings: short-form reviews in SST-5 and long-form essays in TOEFL11. Our goal is to determine whether the model's internal focus shifts based on text length and discourse complexity.

Dependence on Discourse Granularity: On the short, sentiment-focused sentences of the SST-5 dataset, all models dedicated the vast majority of their attention to interactions involving entities. As shown in Figure 3, the sum of Entity-NonEntity ($\text{Ratio}_{\text{E-NE}}$) and Entity-Entity ($\text{Ratio}_{\text{E-E}}$) interactions consistently accounts for over 70% of the total attention. This indicates that for concise, opinionated text, entities serve as the primary attentional

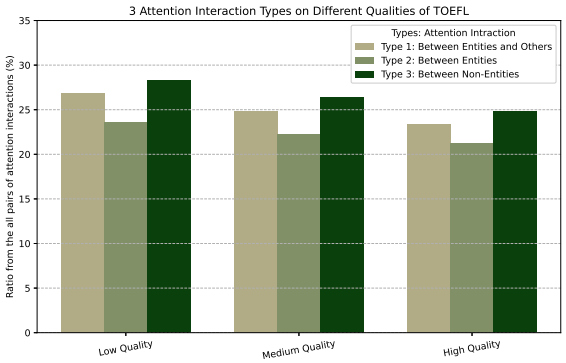

Figure 5: Attention patterns across different qualities of TOEFL essays. As text quality improves, the relative attention on Entity-NonEntity interactions (Type 1) slightly decreases.

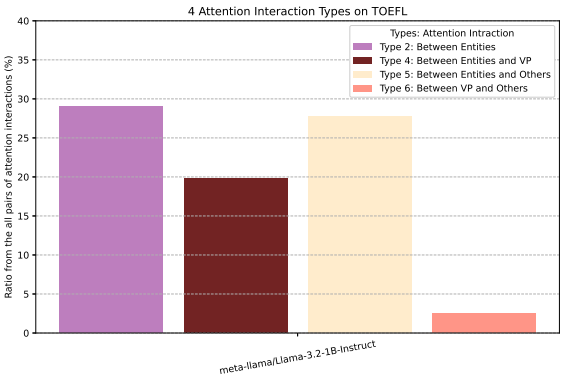

Figure 6: Attention patterns across different qualities of TOEFL essays. As text quality improves, the relative attention on Entity-NonEntity interactions (Type 1) slightly decreases.

anchors for the model. In contrast, on the long-form essays of the TOEFL dataset (Figure 4), attention is more distributed. Entity-related interactions still command a significant share but constitute a smaller portion of the total, ranging from 20% to 30%. This suggests that in complex, descriptive prose, models balance their focus between key entities and broader contextual and structural cues.

Effect of Writing Quality: To analyze whether attention patterns are sensitive to writing quality, we examine the TOEFL11 subset with labeled proficiency levels ("low", "medium", "high"). After controlling for essay length and sentence count, we observe a subtle inverse correlation: as writing quality improves, the proportion of Entity-NonEntity attention (Type 1) slightly decreases. Figure 5 illustrates this trend, with low-quality essays exhibiting approximately 26% Type 1 interaction, compared to 23% for high-quality essays.

This observation aligns with previous findings that well-written texts exhibit richer lexical diversity and syntactic variety (Louis and Nenkova, 2013), allowing models to rely on a broader set of discourse cues. Hence, entity tracking remains essential but is less dominant when more reliable and structured context is available.

5.2 Entities Receive Most Attention in Complex Texts

To better understand how SLMs process long-form texts, we conduct a fine-grained analysis of the TOEFL11 dataset by expanding our attention scope beyond noun phrases. Specifically, we compare attention interactions between entities and verb phrases (VPs), as well as between entities and other non-labeled tokens.

Our results reveal a clear attentional hierarchy. As shown in Figure 6, interactions involving entity tokens (e.g., Entity-Entity, Entity-VP, Entity-Other) consistently account for more than 70% of total attention: the sum of Type 2, 4, and 5. By contrast, interactions between verb phrases and non-entity tokens are minimal (approximately 2.5%). This finding confirms that even in linguistically complex environments, SLMs focus on entities as central nodes in the discourse structure.

This behavior is in line with Centering Theory (Grosz et al., 1995), which posits that entities serve as coherence anchors during discourse progression. Our results suggest that pretrained SLMs implicitly adopt a similar processing strategy, prioritizing entities as focal elements in attention allocation.

5.3 PEFT Increases Entity Attention and Improves Accuracy

We next investigate whether QLoRA-based PEFT affects entity-focused attention behavior and model performance. The attention layers of all models are fine-tuned on the SST-5 training set (Appendix B). We then compare their attention distributions and classification accuracy on a balanced evaluation set, which was constructed by sampling 200 instances from each label of the test set to address class imbalance.

Performance Gains: Prior to fine-tuning, the models perform poorly on the 5-class sentiment classification task, with accuracy scores around 40%. After applying LoRA, we observe substantial improvements in classification accuracy and kappa scores (Table 1). In particular, PEFT let SLMs predict extreme emotions well, which was not possible

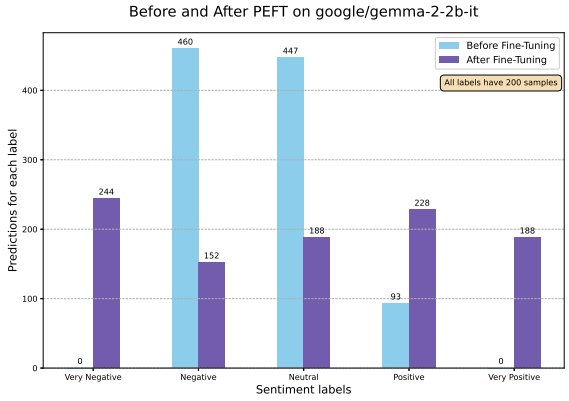

Figure 7: Error analysis: Predictions per label, before and after PEFT on SST5

before fine-tuning (Figure 7). The models become better at distinguishing closely related sentiment categories, such as "positive" vs. "very positive", confirming that LoRA tuning effectively enhances task-specific capabilities.

Classification Performance before LORA			
	Acc	**Kappa-L**	**Kappa-Q**
gemma-2-2b-it	0.38	0.50	0.71
Llama-3.2-1B-it	0.26	0.16	0.27
Llama-3.2-3B-it	0.45	0.55	0.71
Qwen2.5-1.5B-it	0.41	0.54	0.71
Phi-3.5-mini-it	0.42	0.54	0.74
Classification Performance After LORA			
	Acc	**Kappa-L**	**Kappa-Q**
gemma-2-2b-it	0.60	0.73	0.88
Llama-3.2-1B-it	0.52	0.66	0.83
Llama-3.2-3B-it	0.52	0.66	0.83
Qwen2.5-1.5B-it	0.52	0.67	0.83
Phi-3.5-mini-it	0.61	0.74	0.88

Table 1: Classification performance on the SST-5 test set before and after LoRA fine-tuning. PEFT leads to substantial improvements in accuracy (Acc) and both Linear (κ_L) and Quadratic (κ_Q) Weighted Kappa scores.

Shifts in Attention Patterns: Crucially, these performance gains are accompanied by consistent and measurable shifts in attention allocation. Table 2 shows that after LoRA fine-tuning, the proportion of Type 2 interactions (Entity-Entity) increases across all models. This suggests that LoRA encourages the model to more explicitly model semantic relationships between entities. Simultaneously, the proportion of non-entity interactions (Type 3) decreases, reflecting a redistribution of attention toward discourse-salient elements.

This result supports our central hypothesis: LoRA fine-tuning refines the model's internal attention mechanisms by enhancing focus on linguistically meaningful units—specifically, entities. It also validates the use of our attention analysis framework as a lightweight, model-agnostic diagnostic tool for tracking internal behavioral changes induced by fine-tuning.

Model	**Δ E-NE**	**Δ E-E**	**Δ NE-NE**
gemma-2-2b-it	+0.95	-0.91	+0.63
Llama-3.2-1B-it	+0.52	-0.94	+3.22
Llama-3.2-3B-it	-0.27	-0.62	+0.26
Qwen2.5-1.5B-it	+0.40	-0.49	+0.32
Phi-3.5-mini-it	-0.02	+0.04	-0.16

Table 2: Change in attention allocation ratios (in percentage points, pp) on SST-5 after LoRA fine-tuning. The columns show the change in Entity-NonEntity (Δ E-NE), Entity-Entity (Δ E-E), and NonEntity-NonEntity (Δ NE-NE) attention.

6 Conclusion

Our findings suggest that attention weights – often dismissed as unreliable – can, when anchored in syntactic structure, serve not only as effective diagnostic tools but also as a valuable clue for model development. By tracing attention flow through entity representations, we provide an interpretable and lightweight method that not only probes the internal behavior of SLMs but also points toward directions for improving or tailoring such models to better capture entity-based coherence.

We emphasize that our findings should not be taken as a comprehensive explanation of how Small Language Models operate. The scope of our experiments is necessarily limited, and broader generalizations would require further study. Nonetheless, our work highlights an intriguing avenue: entity-focused attention analysis provides a promising perspective on model interpretability that may inspire future research. Extensions could include multi-sentence coherence modeling, cross-lingual entity behavior, or alignment of model outputs with formal discourse theories.

Limitations

This study, while providing clear findings, has several limitations that offer avenues for future research.

First, our definition of an "entity" as a noun phrase is a pragmatic simplification. This approach does not capture more abstract entities, such as events or concepts, and a more sophisticated entity identification method could yield further insights.

Second, our analysis treats attention as a diagnostic correlate, not a definitive causal mechanism. The final output of a Transformer layer is also influenced by the value vector transformations and the computations within the feed-forward networks. A complete mechanistic explanation would require analyzing the interplay between all these components, which was beyond the scope of this work.

Third, the scope of our study is confined to a specific set of SLMs and two classification tasks. While the consistency of our findings across multiple models is encouraging, they may not generalize to all model architectures (e.g., non-Transformers), significantly larger models (LLMs), or different task modalities, such as text generation.

Finally, our method of averaging attention scores across all heads in the final layer provides a high-level, aggregate view of the model's focus. This approach necessarily obscures the diverse and specialized functions that individual attention heads are known to perform (Clark et al., 2019). A more granular, head-level analysis could reveal which specific heads are most affected by fine-tuning and what linguistic roles they play, representing a promising direction for future work.

Acknowledgments

The authors would like to thank the anonymous reviewers for their comments. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author had been supported by a Heidelberg Institute for Theoretical Studies Ph.D. scholarship.

References

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. BERT: A study of attention in BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5717–5726, Hong Kong, China. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Erwan Fagnou, Paul Caillon, Blaise Delattre, and Alexandre Allauzen. 2024. Chain and causal attention for efficient entity tracking. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13174–13188, Miami, Florida, USA. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.

Zichao Li, Yanshuai Cao, and Jackie Chi Kit Cheung. 2024. Do LLMs build world representations? probing through the lens of state abstraction. In *The Twelfth International Conference on Learning Representations*.

Sharid Loáiciga, Anne Beyer, and David Schlangen. 2022. New or old? exploring how pre-trained language models represent discourse entities. In

Proceedings of the 29th International Conference on Computational Linguistics, pages 875–886, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Bhavya Prakash, Tamar Rott Shaham, Tal Linzen, and Yonatan Belinkov. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *The Twelfth International Conference on Learning Representations*.

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kevin Wang, Vikrant Varma, Neel Nanda, Jacob Steinhardt, and David McAllester. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

A Appendix: Dataset Details

The TOEFL11 dataset consists of 12,378 essays written in response to eight distinct open-ended prompts, which are detailed in Table 4. On average, each essay is approximately 411 words long, with further statistics provided in Table 3. The dataset is labeled by proficiency, with a distribution of 1,308 low-quality, 6,568 medium-quality, and 4,502 high-quality essays.

For our sentence-level analysis, we began with a total of 128,549 sentences. We applied several filtering criteria to ensure data quality, excluding: 2,046 sentences that lacked any identifiable entities; 1,970 sentences that our entity-subword mapping parser could not process correctly; and 3,545 sentences shorter than five words. This filtering process resulted in a final set of 120,999 sentences used in our analysis.

Dataset	#Texts	Avg len (Std)	Max len	Scores
T-P1	1,656	401 (97)	902	1-3
T-P2	1,562	423 (97)	902	1-3
T-P3	1,396	407 (102)	837	1-3
T-P4	1,509	405 (99)	852	1-3
T-P5	1,648	424 (101)	993	1-3
T-P6	960	425 (101)	925	1-3
T-P7	1,686	396 (87)	755	1-3
T-P8	1,683	407 (92)	795	1-3

Table 3: Dataset statistics on tokenization: each TOEFL prompt (T-P).

The Stanford Sentiment Treebank (SST-5) dataset contains 5,992 movie reviews for 5-class sentiment classification (from "very negative" to "very positive"). The average sentence length is 23.44 subwords. From this initial set, we excluded 24 sentences shorter than five words and one sentence that failed parsing, resulting in a final analysis set of 5,967 sentences (99.6% of the original dataset).

B Appendix: LORA Hyperparameter Details

The LoRA fine-tuning was conducted using the HuggingFace PEFT library. We employed 4-bit quantization (QLoRA) with the nf4 data type and loaded the base models with fp16 precision. The target modules for LoRA were the attention layers of the SLMs: "q_proj", "k_proj", "v_proj", "o_proj". This results in 0.12% trainable parameters for `google/gemma-2-2b-it`, and 0.14% for `meta-llama/Llama-3.2-1B`. The primary hyperparameters were set as follows: rank=16, alpha=32, lora_dropout=0.05, and a learning rate of $1e$-4 with AdamW optimizer. Models were trained for 2 epochs with a batch size of 4.

C Appendix: Subwords Tokenization as SLM

Our entity-subword mapping parser was designed to handle model-specific tokenization schemes. We observed that the SLMs in our study primarily use one of two conventions to mark word boundaries: a prefix _ (e.g., _word) or a special character "Ġ" (e.g., "Ġ"word). Our parser correctly interprets these conventions for each model to ensure accurate alignment between linguistic noun phrases and their corresponding subword tokens, as illustrated

T-Prompt 1	Agree or Disagree: It is better to have broad knowledge of many academic subjects than to specialize in one specific subject.
T-Prompt 2	Agree or Disagree: Young people enjoy life more than older people do.
T-Prompt 3	Agree or Disagree: Young people nowadays do not give enough time to helping their communities.
T-Prompt 4	Agree or Disagree: Most advertisements make products seem much better than they really are.
T-Prompt 5	Agree or Disagree: In twenty years, there will be fewer cars in use than there are today.
T-Prompt 6	Agree or Disagree: The best way to travel is in a group led by a tour guide.
T-Prompt 7	Agree or Disagree: It is more important for students to understand ideas and concepts than it is for them to learn facts.
T-Prompt 8	Agree or Disagree: Successful people try new things and take risks rather than only doing what they already know how to do well.

Table 4: Topic description: TOEFL (T).

in Table 5.

Origin Text	Is it better for oneself's developpement to have broad knowledge of many academic subjects than to specialize in one specific subject?
google/gemma-2-2b-it	
Tokenized-Ids	tensor([[2, 2437, 665, 2525, 604, 63320, 235303, 235256, 2115, 227070, 577, 791, 7209, 5567, 576, 1767, 15459, 12749, 1178, 577, 78292, 575, 974, 3724, 5091, 235336]])
Tokenized-Subwords	[‘<bos>’, ‘Is’, ‘_it’, ‘_better’, ‘_for’, ‘_oneself’, ‘́’, ‘s’, ‘_develop’, ‘pement’, ‘_to’, ‘_have’, ‘_broad’, ‘_knowledge’, ‘_of’, ‘_many’, ‘_academic’, ‘_subjects’, ‘_than’, ‘_to’, ‘_specialize’, ‘_in’, ‘_one’, ‘_specific’, ‘_subject’, ‘?’]
meta-llama/Llama-3.2-1B	
Tokenized-Ids	tensor([[128000, 3957, 433, 2731, 369, 57669, 596, 2274, 79, 1133, 311, 617, 7353, 6677, 315, 1690, 14584, 15223, 1109, 311, 48444, 304, 832, 3230, 3917, 30]])
Tokenized-Subwords	[‘<|begin_of_text|>’, ‘Is’, ‘Ġit’, ‘Ġbetter’, ‘Ġfor’, ‘Ġoneself’, ‘’s’, ‘Ġdevelop’, ‘p’, ‘ement’, ‘Ġto’, ‘Ġhave’, ‘Ġbroad’, ‘Ġknowledge’, ‘Ġof’, ‘Ġmany’, ‘Ġacademic’, ‘Ġsubjects’, ‘Ġthan’, ‘Ġto’, ‘Ġspecialize’, ‘Ġin’, ‘Ġone’, ‘Ġspecific’, ‘Ġsubject’, ‘?’]

Table 5: Examples of different subword tokenization schemes deployed on SLMs.