# Model Internal Sleuthing: Finding Lexical Identity and Inflectional Features in Modern Language Models

**Michael Li**[CMU] **Nishant Subramani**[CMU]
[CMU] Carnegie Mellon University - Language Technologies Institute
{ml6, nishant2}@cs.cmu.edu

## Abstract

Large transformer-based language models dominate modern NLP, yet our understanding of how they encode linguistic information relies primarily on studies of early models like `BERT` and `GPT-2`. We systematically probe 25 models from `BERT Base` to `Qwen2.5-7B` focusing on two linguistic properties: lexical identity and inflectional features across 6 diverse languages. We find a consistent pattern: inflectional features are linearly decodable throughout the model, while lexical identity is prominent early but increasingly weakens with depth. Further analysis of the representation geometry reveals that models with aggressive mid-layer dimensionality compression show reduced steering effectiveness in those layers, despite probe accuracy remaining high. Pretraining analysis shows that inflectional structure stabilizes early while lexical identity representations continue evolving. Taken together, our findings suggest that transformers maintain inflectional features across layers, while trading off lexical identity for compact, predictive representations.

## 1 Introduction

Large transformer-based language models (LMs) are widely used for tasks such as text generation, question answering, and code completion (Workshop, 2023; Groeneveld et al., 2024; Llama, 2024; Hui et al., 2024). However, how these models internally represent linguistic information remains an active research area. Prior work suggests a hierarchical organization where different layers specialize in capturing distinct levels of linguistic structure (Jawahar et al., 2019; Tenney et al., 2019; Rogers et al., 2020). However, these studies focus only on first-generation LMs such as `BERT` and `GPT-2` (Devlin et al., 2019; Radford et al., 2019). Since then, language technology has transformed dramatically — today's models differ in architecture (encoder-only, decoder-only, encoder-decoder), pretraining objectives (masked vs. causal language modeling), training data volume (billions vs. trillions of tokens), and post-training adaptation (Brown et al., 2020; Groeneveld et al., 2024; Lambert et al., 2025). We ask: where and how do modern LMs encode lexical identity and inflectional morphology, and how do these representations vary with model scale and architecture?

To answer these questions we systematically probe 25 pretrained models ranging from `BERT Base` to `Llama-3.1 8B`, spanning multiple architectures, sizes, and training regimes. We train simple classifiers at each layer to predict word-level lexical identity and inflectional features, and evaluate where this information emerges and how linearly accessible they are. We focus on two linguistic properties: *lexical identity* and *inflectional features*, which help disentangle meaning from surface form. Consider the words *walk*, *walked*, *jump*, and *jumped*. Do language models group words by shared meaning (*walk*, *walked*) or by shared grammar (*walked*, *jumped*)? More broadly, where and how do LMs encode a word's lexeme and its inflectional features?

To test whether observed patterns generalize beyond English, we examine six typologically diverse languages: English, Chinese, German, French, Russian, and Turkish. We also investigate where lexical and inflectional information reside (attention heads vs. residual streams), evaluate the impact of editing activations via steering vectors, and track when these representations emerge during pretraining. To the best of our knowledge, this is the first systematic analysis of how lexical identity and inflectional features are encoded across 25 modern language models spanning multiple architectures, scales, and training regimes in six typologically diverse languages. We find that:

1. Lexical identity information is encoded prominently in early layers and becomes increasingly non-linear deeper in the network,



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15825–15864
July 2-7, 2026 

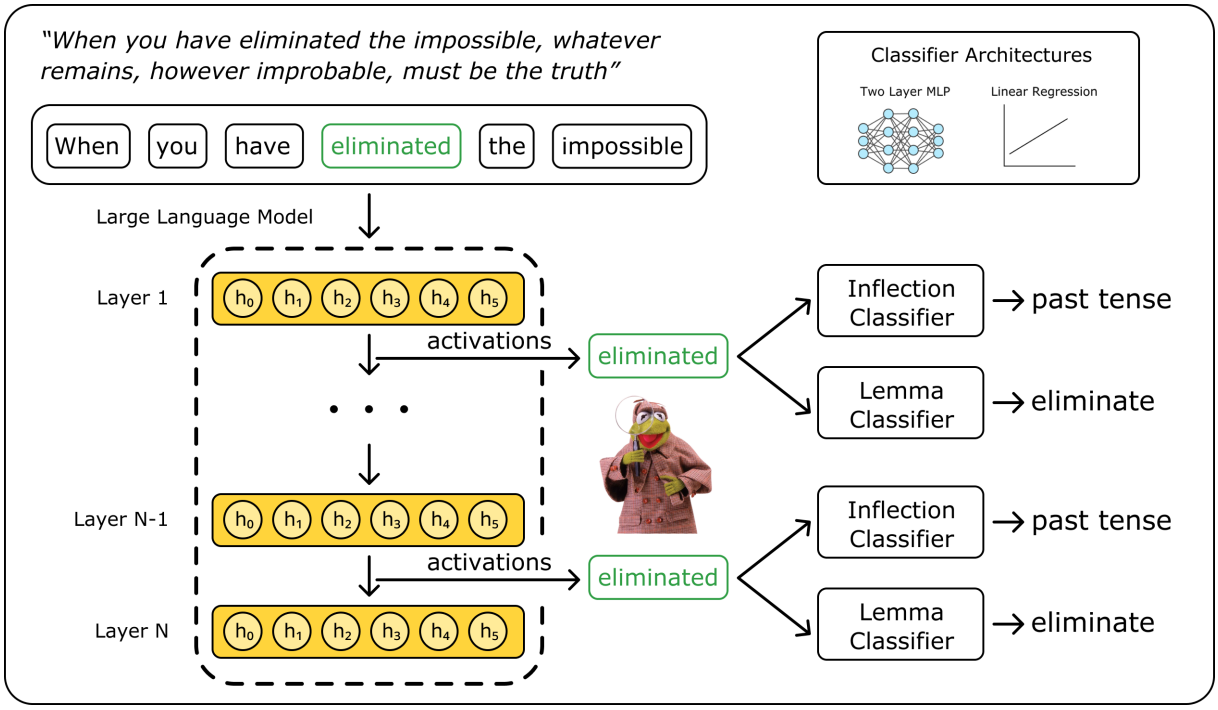



Figure 1: Overview of our probing methodology. We extract hidden state activations from each model layer for target words and train simple linear and shallow non-linear classifiers to predict word-level lexical identity and inflectional features. We compute selectivity using control labels and analyze how linear vs. non-linear accessibility varies with depth.

whereas inflectional information remains linearly accessible across all layers.

2. Across languages, we find that linguistic encoding strength varies with morphological typology, declining most sharply in Turkish.

3. Lexical and inflectional information emerge early in pretraining and reside primarily in the residual stream; inflectional features occupy compact, steerable subspaces that enable effective interventions.

## 2 Probe Design and Metrics

We investigate how language models encode linguistic information using simple classifiers (*probes*) trained on activations from individual layers. We train probes to predict each word's lexeme (*e.g.*, *walk* as the base form of *walked*) and its inflectional features (*e.g.*, plural, past tense).

### 2.1 Probe architectures

For each layer of a model we extract residual-stream representations for a target word and train two classifiers: a linear regression probe and non-linear multi-layer perceptron (MLP) probe. The linear probe measures how well information is linearly separable in the representation space, while the non-linear probe relaxes this linear assumption and searches for a non-linear decision boundary. Comparing these probes allows us to infer whether a property is encoded *linearly* or *non-linearly*. Training details are provided in Appendix B.

#### 2.1.1 Linear Regression Classifier

Consistent with best practices for probing (Hewitt and Liang, 2019; Liu et al., 2019), we use a ridge-regularized linear regression classifier. Given training representations $X_{\text{train}} \in \mathbb{R}^{m \times d}$ and one-hot encoded labels $Y_{\text{train}} \in \mathbb{R}^{m \times c}$, the optimal weight matrix $W \in \mathbb{R}^{d \times c}$ is obtained in closed form as:

$$W = \left(X_{\text{train}}^{\top} X_{\text{train}} + \lambda I\right)^{-1} X_{\text{train}}^{\top} Y_{\text{train}}, \quad (1)$$

where $\lambda$ controls the strength of $\ell_2$ regularization and $I$ is the identity matrix. Predictions on test representations $X_{\text{test}}$ are then given by $\hat{Y}_{\text{test}} = X_{\text{test}} W$.

#### 2.1.2 MLP Classifier

To test for non-linear separability, we train a simple two-layer MLP with ReLU activation:

$$\hat{Y} = \text{softmax}\Big(\text{ReLU}(XW_1)W_2\Big), \quad (2)$$

Here, $W_1 \in \mathbb{R}^{d \times h}$ and $W_2 \in \mathbb{R}^{h \times c}$ are learned weight matrices, $h$ is the hidden dimension (we use $h = 64$), and biases are omitted for brevity.

### 2.2 Metrics

We define two metrics for quantifying signal and nonlinearity across depth: selectivity and the linear separability gap.

**Selectivity.** Probes may simply memorize training data rather than extracting true linguistic information from the representations. To account for this, we construct control tasks following Hewitt and Liang (2019), assigning each unique word form a random class label and training identical probes on these labels. We call this set the control set. We define selectivity at layer $\ell$ as the difference between real and control accuracies:

$$\text{Sel}_\ell = \text{Acc}_\ell^{\text{real}} - \text{Acc}_\ell^{\text{control}} \quad (3)$$

Higher values mean the classifier is extracting true linguistic information rather than just memorizing.

**Linear separability gap.** To compare how much linguistic signal each probe type extracts, we compute the difference in selectivity:

$$\text{Gap}_\ell = \text{Sel}_\ell^{\text{nonlin}} - \text{Sel}_\ell^{\text{linear}} \quad (4)$$

Negative gap values indicate that additional (MLP) probe capacity captures spurious correlations rather than linguistic structure.

## 3 Experiments

Using the methodology introduced in Section §2, we describe the components of our experimental setup: the datasets, model suite, and procedure for extracting token-level representations.

### 3.1 Datasets

For our analysis of lexical identity and inflectional features, we use Universal Dependencies corpora across six languages - English, Chinese, German, French, Russian, Turkish (Nivre et al., 2016). We select GUM for English (Zeldes, 2017), GSD for Chinese/German/French (McDonald et al., 2013; Guillaume et al., 2019), SynTagRus for Russian (Droganova et al., 2018), and IMST for Turkish (Sulubacak et al., 2016). [1]

### 3.2 Models

We study a diverse set of pretrained transformer language models spanning different architectures, sizes, and training regimes. Table 1 lists all models used in this study (see Table 15 for the HuggingFace identifiers).

| Model | Parameters | Pretraining Data | Layers |
|---|---|---|---|
| **Encoder-only** | | | |
| `BERT Base` | 110M | 12.6B tokens[1] | 12 |
| `BERT Large` | 340M | 12.6B tokens[1] | 24 |
| `DeBERTa V3 Large` | 418M | 32B tokens[1] | 24 |
| **Decoder-only** | | | |
| `GPT 2 Small` | 124M | 8B tokens[1] | 12 |
| `GPT 2 Large` | 708M | 8B tokens[1] | 36 |
| `GPT 2 XL` | 1.5B | 8B tokens[1] | 48 |
| `Goldfish English 1000mb` | 124M | 200M tokens | 12 |
| `Goldfish Chinese 1000mb` | 124M | 200M tokens | 12 |
| `Goldfish German 1000mb` | 124M | 200M tokens | 12 |
| `Goldfish French 1000mb` | 124M | 200M tokens | 12 |
| `Goldfish Russian 1000mb` | 124M | 200M tokens | 12 |
| `Goldfish Turkish 1000mb` | 124M | 200M tokens | 12 |
| `Pythia 6.9b` | 6900M | 300B tokens | 32 |
| `Pythia 6.9b Tulu` | 6900M | 300B tokens | 32 |
| `OLMo 2 7B` | 7300M | 4T tokens | 32 |
| `OLMo 2 7B Instruct` | 7300M | 4T tokens | 32 |
| `Gemma 2 2B` | 2610M | 2T tokens | 26 |
| `Gemma 2 2B Instruct` | 2610M | 2T tokens | 26 |
| `Qwen2.5 1.5B` | 1540M | 18T tokens | 28 |
| `Qwen2.5 1.5B Instruct` | 1540M | 18T tokens | 28 |
| `Qwen2.5 7B` | 7620M | 18T tokens | 28 |
| `Qwen2.5 7B Instruct` | 7620M | 18T tokens | 28 |
| `Llama 3.1 8B` | 8000M | 15T tokens | 32 |
| `Llama 3.1 8B Instruct` | 8000M | 15T tokens | 32 |
| **Encoder-Decoder** | | | |
| `mT5-base` | 580M | 1T tokens | 12 |

Table 1: Overview of models used in this study. [1]Converted from GB to tokens using the approximation that 1GB of data is approximately 200M tokens in English (Chang et al., 2024).

For English, we evaluate 19 models: all models in Table 1 except the five non-English `Goldfish` models and `mT5-base`. For the five non-English languages (Chinese, German, French, Russian, Turkish), we use a set of models with explicit coverage of each target language: the corresponding monolingual `Goldfish <Language> 1000mb` model (Chang et al., 2024), multilingual `Qwen2.5-1.5B` (and instruct), multilingual `Qwen2.5-7B` (and instruct) (Team, 2024), and the multilingual `mT5-base` model (Xue et al., 2021). This ensures that we evaluate models on languages they were trained on while maintaining sufficient coverage.

### 3.3 Representation Extraction

Each input to the model is a complete sentence from the corpus. We tokenize inputs with model-specific tokenizers and run a forward pass to collect residual-stream activations at the target word position from every layer.[2]

[1] See Appendix §E for complete details including dataset statistics, tokenization information, and visualizations for all languages

[2] We also experiment with attention head outputs (see §4.2 and §G)

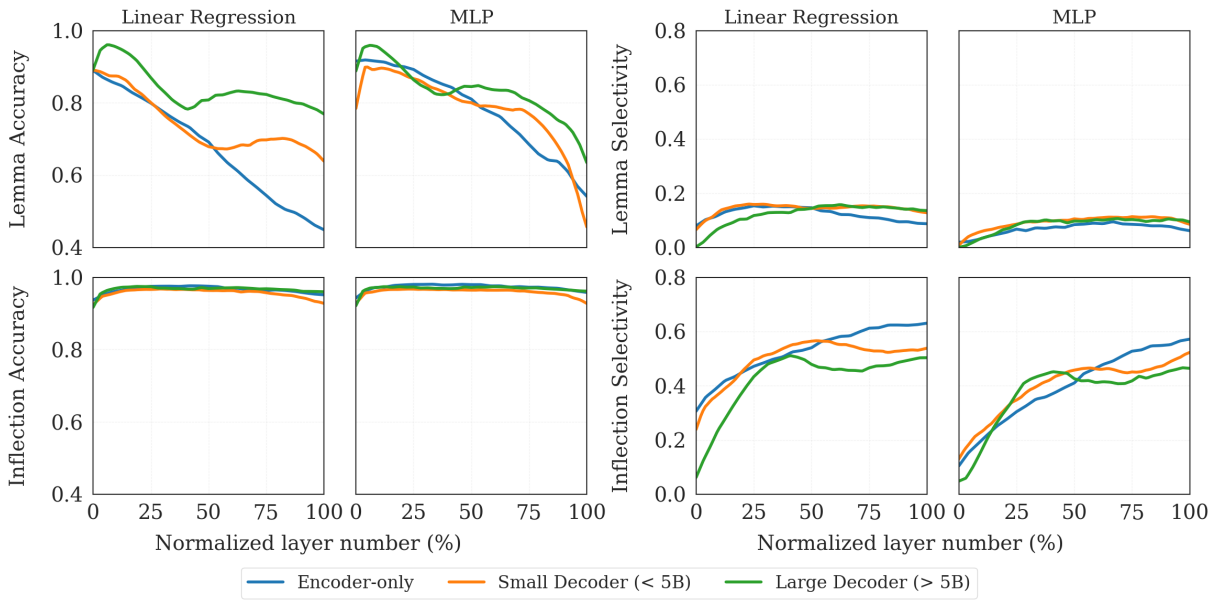



Figure 2: Lexeme and inflection probing results for English, *averaged by model category*: encoder-only (`BERT`, `DeBERTa`), small decoder <5B (`GPT-2`, `Gemma-2-2B` (and instruct), `Qwen2.5-1.5B` (and instruct)), and large decoder >5B (`Pythia-6.9B`, `OLMo-2-7B`, `Llama-3.1-8B` and instruct versions). Columns show prediction accuracy (Linear vs. MLP probes) and selectivity scores (linguistic minus control accuracy). Note that for readability, the y-axis for accuracy starts at 0.4. Full (non-averaged) results for individual models are provided in Appendix §C.

For words split into multiple subwords, we use the last subword's representation, since this is where prior work suggests word-level meaning lives (Kaplan et al., 2025; Feucht et al., 2025).

## 4 Lexical Identity and Inflectional Features

### 4.1 Results

We report layer-wise accuracies for lexeme and inflection prediction across all datasets models described in §3. Detailed layer-wise accuracy and selectivity tables are provided in Appendix §D.

**Lexeme.** Lexeme accuracy under linear regression starts high (0.8–1.0) and decreases with depth in all English model families (Figure 2, top left). Encoder-only models show the strongest decrease, while small decoders decline more gradually and large decoders maintain higher accuracy in deeper layers. Across languages (Figure 3, top left), Turkish shows the largest drop (0.95 to 0.25), while Russian and Chinese retain 0.6–0.8 accuracy in later layers. MLP accuracy is similar but slightly higher than linear at most depths (middle column). Selectivity for lexeme remains close to zero across depths and languages (right column), indicating that high lexeme accuracy early in the network is mostly driven by surface correlations rather than strongly selective lexical structure.

**Inflection.** Inflectional features remain readable across all layers and architectures. For English, linear regression accuracy stays near 0.9–1.0 throughout the layers (Figure 2, bottom left). This pattern holds cross-linguistically (Figure 3, bottom left): English, Chinese, German, French, and Russian exceed 0.9 accuracy at most depths, while Turkish is slightly lower, hovering around 0.8–0.9. MLP probes follow the same pattern (middle column). Selectivity scores for inflection remain positive (0.4–0.6) across models and languages (right column), with Russian and German at the upper end, supporting the view that inflectional features are encoded in stable, linearly accessible subspaces.

**Probe error analysis.** Frequency strongly correlates with probe accuracy for both tasks, *i.e.* frequent lexemes and inflectional categories achieve high accuracy, while rarer ones account for most of the errors. For inflection, comparative and superlative degrees and low-frequency verb forms are the most error-prone categories. Turkish shows the strongest sensitivity to frequency, likely due to its morphological complexity creating a long tail of rare forms. A detailed breakdown by part of speech and inflectional category is given in Appendix §I.

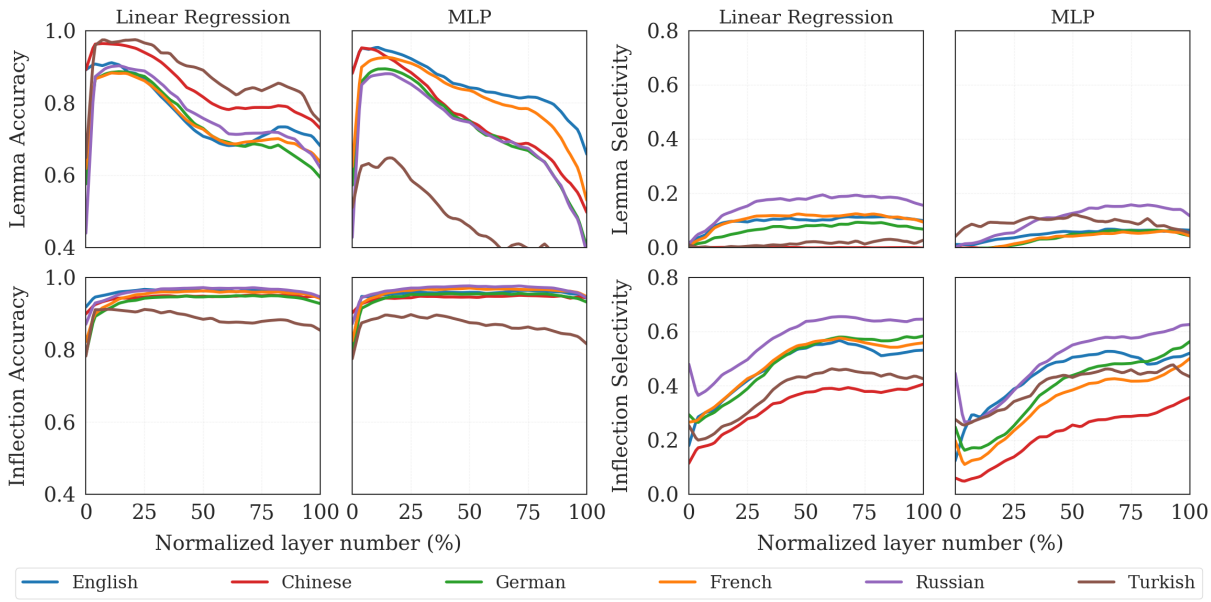



Figure 3: Cross-linguistic probing results *averaged across all models within each language*. Columns show lexeme and inflection accuracy (Linear vs. MLP) followed by selectivity scores. Note that for readability, the y-axis for accuracy starts at 0.4. Full (non-averaged) results for individual models are provided in Appendix §C.

### 4.2 Analysis

Our results show that lexical identity is encoded strongly in early layers but becomes less accessible in later layers, whereas inflectional features remain robustly decodable throughout the model. We run a variety of experiments to further investigate this.

**Linear probes are more selective than MLP probes.** We quantify the relationship between probe type and selectivity via the linear separability gap (defined in equation (4)); detailed plots appear in Appendix §F.3. As shown in Figure 4, the gap is negative for both tasks across layer depths, indicating that linear probes achieve higher selectivity than MLP probes. For lexeme, the gap is consistently negative, even though MLP probes achieve higher accuracy. This suggests that while non-linear probes can extract more lexeme information, some of that additional signal reflects memorization rather than genuine linguistic structure. For inflection, the gap is also negative on average but exhibits substantially higher variance. This suggests that in some cases both probe types extract similar amounts of selective linguistic signal (gap near zero), while in others linear probes are considerably more selective.

**Some models show extreme mid-layer dimensionality compression; others maintain stable dimensionality.** To characterize representation geometry, we measure the linear effective dimensionality across layers, building on prior work showing

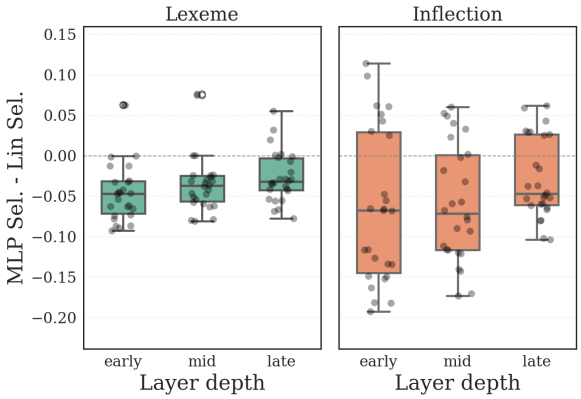



Figure 4: Linear separability gap (difference in probe selectivity) for lexeme and inflection, aggregated across models and languages and grouped by layer depth (early, mid, late). Each point represents a single model-language pair.

that neural network loss landscapes and language model fine-tuning subspaces often have low intrinsic dimensionality (Li et al., 2018; Aghajanyan et al., 2021). Following Subramani et al. (2019), who referred to this as effective dimension and approximated it linearly via PCA, we estimate how many PCA components (as a fraction of the full basis) are needed to explain fixed variance thresholds on our dataset of collected activations. As noted by Subramani et al. (2019), this PCA-based linear approximation serves as an upper bound for the unconstrained effective dimension.

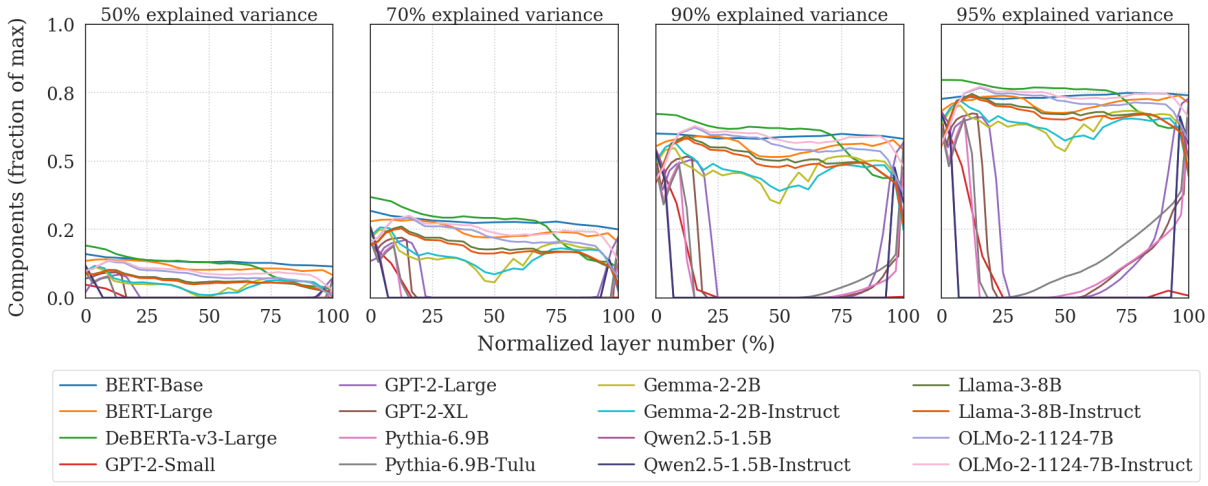



Figure 5: Linear effective dimensionality across layers. Lines show fraction of PCA components needed to reach variance thresholds (50–95%). Full results appear in §F.1.

Figure 5 summarizes these trajectories across thresholds (50–95%), and full results are provided in Appendix §F.1. In models with gradual compression, the curves decline smoothly with layer index, indicating steady consolidation without a single dominant bottleneck; this behavior appears in encoder-only models (`BERT`, `DeBERTa`) and several contemporary decoders (`Gemma`, `Llama`, `OLMo-2`). However, we also observe a sharp mid-layer collapse for `GPT-2`, `Qwen2.5`, and `Pythia`: even very high variance thresholds (90%, 95%) become explainable by very few components before re-expanding in the very last layers. When inspecting activation statistics (Appendix §F.2), we find that this collapse coincides with outlier dimensions of unusually large magnitudes: `Qwen2.5-1.5B` reaches maximum absolute activations of ∼8000 in its middle layers, whereas `Llama-3-8B` remains around ∼30–40 (Rudman et al., 2023; Sun et al., 2024).

**Residual streams retain more linguistic information than attention outputs.** We conduct a targeted probing experiment that compares *attention-head outputs* to *residual-stream activations* in `BERT` and contemporary decoders. Figure 6 summarizes the effect averaged across models and probe types (the corresponding full, per-model curves are shown in Figures 27 and 28 in Appendix §G). Across both lexeme and inflection, probes trained on attention outputs have lower accuracy than probes trained on the residual stream at nearly all depths. For lexeme, attention-based accuracy drops to roughly 0.2–0.4 in middle layers, while residual-stream accuracy remains closer to 0.6–0.9. For inflection, both components stay highly decodable (0.7–1.0), but residual streams still consistently outperform attention outputs, especially in middle layers. Selectivity mirrors these trends: lexeme selectivity stays near zero on attention outputs but is higher on residuals, while inflection selectivity reaches 0.4–0.5 in both streams with residuals slightly higher. Overall, these experiments support an interpretation in which attention emphasizes contextual aggregation, while the residual stream (including MLP mixing) more directly preserves token-level lexical and morphological information used by the probes, hinting that these could be better overall representations.

**Inflection representations are highly steerable.** We connect these representational measurements to causal control through inflection steering experiments (*e.g.*, singular vs. plural). For each category pair, we compute a difference vector between mean hidden states and apply scaled interventions at each layer. [3] To evaluate steering effectiveness, we apply these vectors to a set of test examples and use our trained linear classifier to assess whether the intervention successfully changes the predicted inflectional category. For example, when steering from past tense ('-ed') to gerund ('-ing'), we add the difference vector $\mathbf{v}_{\text{ing}-\text{ed}}$ to the hidden state of a word like "jumped" and measure whether the classifier now assigns higher probability to '-ing' than to '-ed'. Figures 29 and 30 show that across most architectures, even moderate intervention strengths ($\lambda = 5$) produce large probability shifts and high

[3] This is identical to difference of means interventions with steering vectors (Subramani et al., 2022).

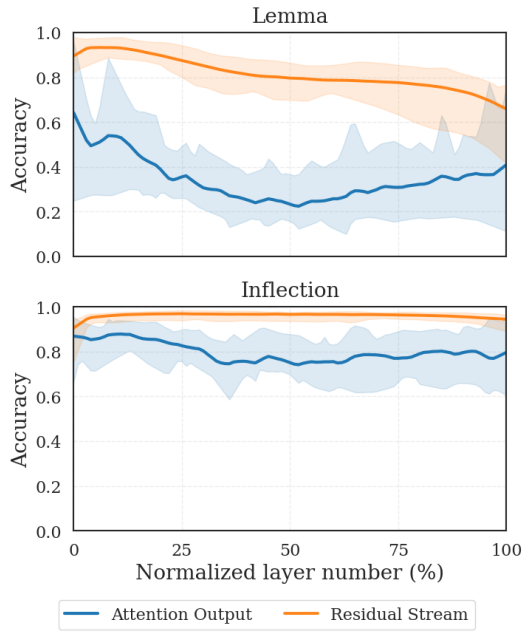



Figure 6: Probing accuracy for lexeme (top) and inflection (bottom) as a function of normalized layer depth, comparing attention-head outputs to residual-stream activations. Curves are averaged across models and probe type, with min-max shaded regions.

flip rates, consistent with inflection being controllable along a small number of directions. One exception is `DeBERTa-v3-Large`, which shows a sudden drop in steering effectiveness around 75% of model depth. The consistency of this pattern across all tested $\lambda$ values indicates that it reflects properties of the model's representational structure rather than the intervention strength $\lambda$.

**Inflection stabilizes early in training; lexeme continues to change.** Finally, we analyze pretraining dynamics by probing intermediate checkpoints for `OLMo-2-7B` and `Pythia-6.9B` (Figures 25 and 26 in Appendix §F.4). In both model families, inflection accuracy is already high at the earliest checkpoints and improves only slightly with further updates; inflection selectivity rises rapidly in the first few checkpoints and then stays near its final value. Lexical identity exhibits a different tradeoff: lemma/lexeme accuracy is highest in earlier checkpoints and tends to decline with additional training (and with depth), even as lexeme selectivity increases gradually over training, particularly in mid-to-late layers. Figure 7 visualizes these dynamics jointly across training and depth: inflection reaches a stable regime early, while lexical identity remains more variable and continues to be reshaped throughout training, particularly in later layers of decoder-only models. This is similar to findings from cognitive science that grammatical knowledge requires far less information storage than lexical semantics during human language acquisition (Mollica and Piantadosi, 2019).

### 4.3 Discussion

We find a consistent trend across models and languages. Inflection stays highly decodable across depth and shows strongly positive selectivity, while lexeme accuracy starts high and drops in later layers and lexeme selectivity stays near zero. This suggests a simple story in which models choose to retain morphosyntactic features, since they constrain surface realization, while moving away from token identity as representations become more contextual. However, there could be other explanations: lexeme prediction has far higher cardinality than inflection, and frequency drives many of the remaining errors, so task difficulty could explain part of the gap.

We also see substantial variation in representation geometry across model families. Some models show a sharp mid-layer collapse in PCA-based effective dimensionality, and this collapse coincides with unusually large outlier activations (Rudman et al., 2023; Sun et al., 2024). Other models compress more gradually without a clear bottleneck.

Finally, we observe, based on the linear separability gap, that additional probe capacity tends to capture spurious correlations rather than linguistic structure. This aligns with Hewitt and Liang (2019), who report a similar trend on `ELMo`, highlighting the need to report selectivity alongside accuracy when interpreting probing results.

## 5 Related Work

**Probing for linguistic information.** Probing studies typically use supervised classifiers to predict linguistic properties from model representations (Alain and Bengio, 2017; Adi et al., 2017). Extensive work has established that early transformer models (`BERT`, `GPT-2`) learn hierarchical linguistic structures, with different layers specializing in different information types: lower layers capture surface features and morphology, middle layers encode syntax, and upper layers represent semantics and context (Jawahar et al., 2019; Tenney et al., 2019; Rogers et al., 2020). More relevant to our work, Vulić et al. (2020) found that lexical information concentrates in lower layers, while Ethayarajh

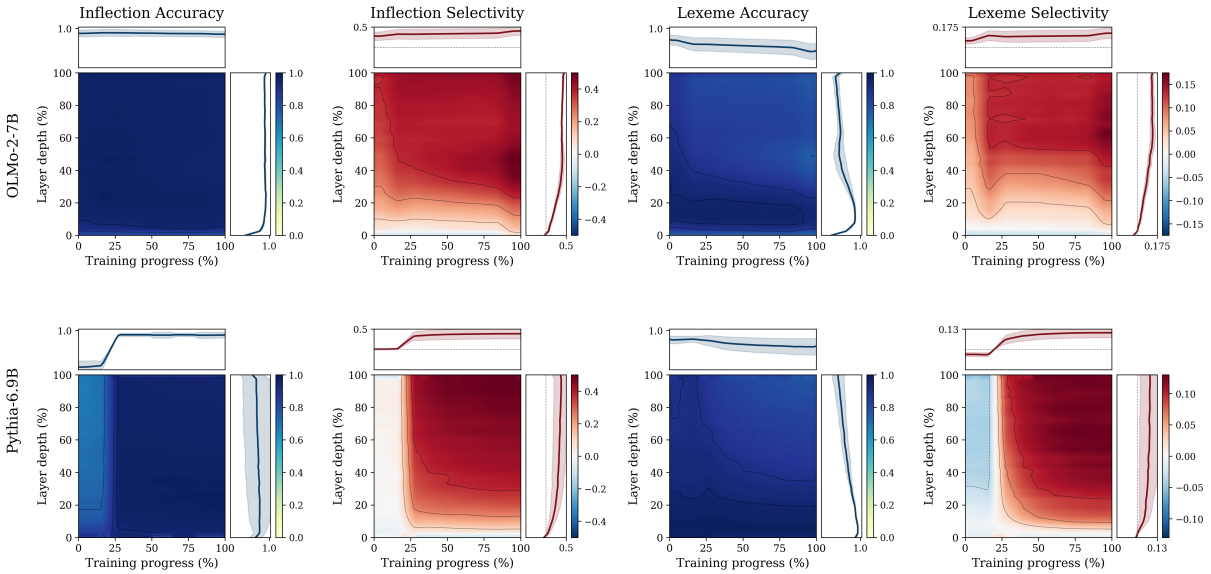



Figure 7: Checkpoint-by-layer heatmaps for `OLMo-2-7B` (top row) and `Pythia-6.9B` (bottom row). Columns show inflection accuracy, inflection selectivity, lexeme accuracy, and lexeme selectivity. The x-axis is checkpoint training progress and the y-axis is model layer depth.

(2019) showed that representations become increasingly context-specific in higher layers.

For morphology specifically, Acs et al. (2024) introduced an extensive multilingual probing dataset (247 tasks across 42 languages), finding that `mBERT` and `XLM-RoBERTa` encode morphosyntactic features strongly, with preceding context more informative than following context for morphological prediction. Lasri et al. (2022) distinguished between spurious and *functional* encodings—those actually used by the model—demonstrating that BERT relies on linear representations of grammatical number for agreement tasks, with nouns and verbs encoded in disjoint subspaces.

**Representation dynamics in modern LLMs.** Recent research has extended these analyses to modern, larger-scale generative models. Cheng et al. (2025) identify a distinct high-dimensional abstraction phase in the early-to-middle layers of models like Llama and OLMo, suggesting that the transition from surface-level to abstract linguistic features occurs earlier than in previous architectures. Similarly, Skean et al. (2025) demonstrate that intermediate layers in modern LLMs often encode richer task-transferable representations than final layers. These findings align with the layerwise dynamics we observe in Section §3.

**Activation steering and mechanistic interpretability.** Beyond probing, recent work has explored manipulating model behavior by intervening on internal representations, including steering vectors (Subramani et al., 2022), inference-time interventions (Li et al., 2023), and representation editing (Meng et al., 2022). Mechanistic interpretability approaches aim to reverse-engineer learned algorithms (Elhage et al., 2021), with recent work using sparse autoencoders to decompose representations into interpretable latent features (Cunningham et al., 2023; Bricken et al., 2023), providing clearer targets for interpretation than raw activations. See Appendix §A for detailed discussion.

## 6 Conclusion

In this work, we analyzed 25 transformer models to understand how they encode two token-level linguistic properties: lexical identity and inflectional features. We find that these properties follow distinct representational trajectories: lexical identity is most linearly accessible in early layers but becomes increasingly entangled deeper in the network, while inflectional information remains robustly and linearly decodable across layers and languages. Additional analyses of residual streams, attention outputs, activation steering, and pretraining dynamics further reveal that inflection occupies compact, steerable subspaces that stabilize early in training, and that linear probes are more selective than MLP probes for capturing these properties. Collectively, these findings suggest that despite rapid advances in model scale and training, transformers converge on robust internal representations of core morphological properties.

## 7 Limitations

**Representation Extraction for Decoder Models** Our current approach for extracting word representations from decoder-only models uses the final subword token. This assumption is an intuitive and natural choice, but may not be optimal for all architectures and models. Future work could develop better extraction methods that account for subword tokenization effects and leverage attention patterns to create more accurate word-level representations.

**Form and Function in Inflection** Some languages contain cases where different grammatical functions share the same surface form (*e.g.*, infinitive and non-past verb forms in English). We do not explicitly examine these cases in our classification experiments, but these ambiguities create opportunities to better examine how models separate form from function across languages.

**Indirect Nature of Classifiers** While our classifier methodology follows established best practices (Hewitt and Liang, 2019; Liu et al., 2019), we only detect correlations in hidden activations, not causal mechanisms.

**Scope of Steering Experiments** Our steering vector experiments measure changes in classifier performance rather than downstream model outputs. Evaluating effects on actual model generation would require more complex experimental designs to control for confounding factors and ensure that observed changes result from the intended representational modifications rather than other influences.

## References


Judit Acs, Endre Hamerlik, Roy Schwartz, Noah A. Smith, and Andras Kornai. 2024. Morphosyntactic probing of multilingual bert models. *Natural Language Engineering*, 30(4):753–792.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *5th International Conference on Learning Representations (Conference Track)*.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations (Workshop Track)*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, and 6 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *Preprint*, arXiv:2408.10441.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. 2025. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

Thao Anh Dang, Limor Raviv, and Lukas Galke. 2025. Tokenization and morphology in multilingual language models: A comparative analysis of mT5 and ByT5. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 242–257, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian ud treebanks. In *Proceedings*

*of the 17th international workshop on treebanks and linguistic theories (tlt 2018)*, volume 155, pages 53–66. Linköping University Electronic Press Linköping, Sweden.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Sheridan Feucht, Eric Todd, Byron C Wallace, and David Bau. 2025. The dual-route model of induction. In *Second Conference on Language Modeling*.

Atticus Geiger, Hanson Lu, Thomas F Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies [conversion and improvement of Universal Dependencies French corpora]. *Traitement Automatique des Langues*, 60(2):71–95.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. 2024. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497, Torino, Italia. ELRA and ICCL.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. 2025. From tokens to words: On the inner lexicon of LLMs. In *The Thirteenth International Conference on Learning Representations*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jiarui Liu, Jivitesh Jain, Mona Diab, and Nishant Subramani. 2025. Llm microscope: What model internals reveal about answer correctness and context utilization. *arXiv preprint arXiv:2510.04013*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Team Llama. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Francis Mollica and Steven T. Piantadosi. 2019. Humans store about 1.5 megabytes of information during language acquisition. *Royal Society Open Science*, 6(3):181393.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does BERT rediscover a classical NLP pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

nostalgebraist. 2020. Interpreting GPT: The logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`. LessWrong blog post.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. Steering llama 2 via contrastive activation addition. *Preprint*, arXiv:2312.06681.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

William Rudman, Catherine Chen, and Carsten Eickhoff. 2023. Outlier dimensions encode task specific knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14596–14605, Singapore. Association for Computational Linguistics.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*.

Nishant Subramani, Samuel Bowman, and Kyunghyun Cho. 2019. Can unconditional language models recover arbitrary sentences? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nishant Subramani, Jason Eisner, Justin Svegliato, Benjamin Van Durme, Yu Su, and Sam Thomson. 2025. MICE for CATs: Model-internal confidence estimation for calibrating agents with tools. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12362–12375, Albuquerque, New Mexico. Association for Computational Linguistics.

Nishant Subramani and Nivedita Suresh. 2020. Discovering useful sentence representations from large pretrained language models. *CoRR*, abs/2008.09049.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. In *First Conference on Language Modeling*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

# A Additional Related Work

## A.1 Advanced Probing Methodologies

Beyond standard linear probes, sophisticated approaches have emerged to understand model representations. Amnesic probing (Elazar et al., 2021) removes specific information from representations to test whether it's necessary for downstream tasks. Information-theoretic probing frameworks formalize probing as estimating how much information representations contain about linguistic structure (Pimentel et al., 2020), while minimum description length probes (Voita and Titov, 2020) balance probe complexity with performance to avoid overfitting. Causal abstraction (Geiger et al., 2021) aims to establish causal rather than merely correlational relationships between representations and linguistic properties.

Complementary work has revisited what layerwise probing results imply: Niu et al. (2022) re-examine prior evidence for pipeline-like separation of linguistic knowledge across BERT layers, arguing that the pattern is more nuanced than depth alone explains. Recently, Subramani et al. (2025) and Liu et al. (2025) find that decoding from or probing intermediate activations can yield reliable confidence and correctness estimators for LLMs (nostalgebraist, 2020).

For morphology specifically, Acs et al. (2024) introduced an extensive multilingual probing dataset (247 tasks across 42 languages), finding that `mBERT` and `XLM-RoBERTa` encode morphosyntactic features strongly, with preceding context more informative than following context for morphological prediction. Using minimal pairs, He et al. (2024) found that GPT-2 captures syntactic structure in its first third of layers, with morphological and semantics-syntax interface features proving harder to decode than pure syntax. Dang et al. (2025) compared mT5 and ByT5 on morphological probing, finding that linear classifiers match MLP performance, suggesting morphological features are encoded in linearly separable subspaces; tokenization strategies significantly impact morphological representation quality, particularly for morphologically rich languages.

## A.2 Model Manipulation and Steering

Steering vectors demonstrate that specific directions in activation space correspond to high-level behavioral changes (Subramani et al., 2019; Subramani and Suresh, 2020; Subramani et al., 2022). Building on this, Panickssery et al. (2024) achieves behavioral control by adding activation differences between contrasting examples. Li et al. (2023) introduce inference-time intervention, shifting model activations during inference across limited attention heads to control behavior. While these methods operate in activation space, task vectors enable arithmetic operations on model capabilities by manipulating weight space (Ilharco et al., 2023). Sparse autoencoders provide another avenue for feature discovery (Bricken et al., 2023), while causal mediation analysis (Vig et al., 2020) helps identify which components mediate specific behaviors.

# B Probe Training Details

We stratify each dataset into train, validation, and test splits. Probes are trained on the training split, hyperparameters are selected using the validation split, and we report accuracy and macro F1 on the held-out test split. For the linear regression probe we apply ridge regularization with $\lambda = 0.01$ and solve equation (1) in closed form. For the MLP probe we use a hidden dimension of 64, a learning rate of 0.001, weight decay of 0.01, and train for up to 100 epochs with early stopping based on validation loss, optimizing cross-entropy with AdamW. Both probes share the same data splits to enable fair comparison.

# C Full Lemma and Inflection Probe Results

We provide the full, non-averaged results for the linguistic probing tasks (lemma identity and inflectional features) for every individual model. Figure 8 shows the detailed breakdown for English models, and Figure 9 presents the results for all six languages.

# D Layer-wise Tables for Lemma and Inflection Results

This section contains detailed tables for layer-wise accuracy and selectivity across all models and languages.

For English, we provide separate tables for each probe type and metric:

- Table 2: Accuracy using linear probes
- Table 3: Selectivity using linear probes
- Table 4: Accuracy using MLP probes
- Table 5: Selectivity using MLP probes

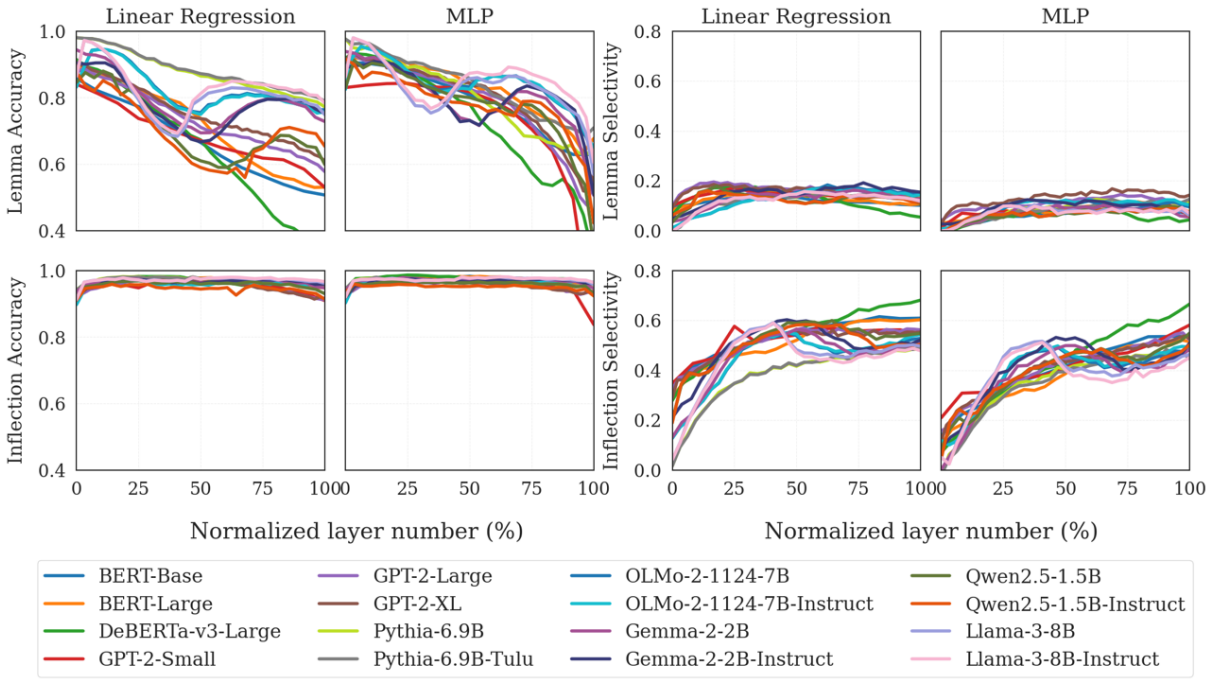



Figure 8: Full lemma and inflection probing results for English, showing individual curves for every model. Columns show prediction accuracy (Linear vs. MLP probes) and selectivity scores (linguistic minus control accuracy).

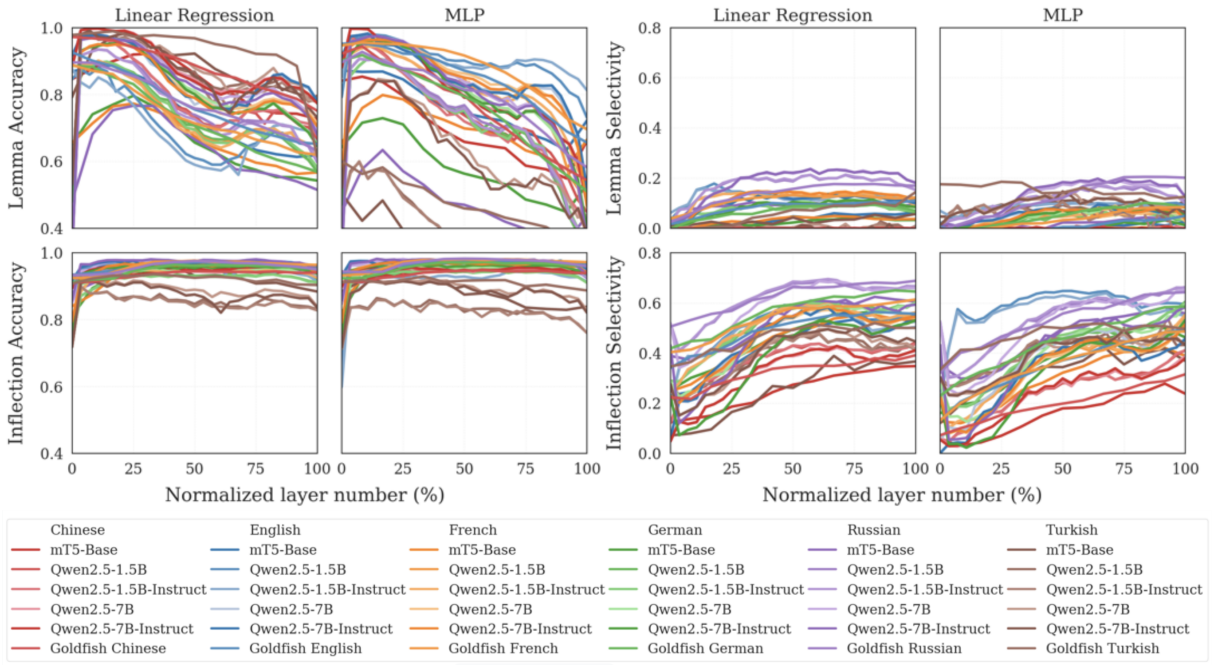



Figure 9: Full cross-linguistic probing results showing individual curves for every model within each language. Columns show lemma and inflection accuracy (Linear vs. MLP) followed by selectivity scores.

For the cross-linguistic experiments, we provide combined tables showing both linear and MLP results for accuracy and selectivity:

- Table 6: Probing results for Chinese
- Table 7: Probing results for French
- Table 8: Probing results for German
- Table 9: Probing results for Russian
- Table 10: Probing results for Turkish

# E Dataset Statistics

This section provides statistics and visualizations for the datasets and models used in our experiments across all six languages. Only words containing alphabetic characters and apostrophes were considered.

## E.1 English Dataset Details

For the English GUM corpus specifically, the data covers three main syntactic categories: nouns (49.5%), verbs (31.2%), and adjectives (19.4%).

Table 12a shows the distribution of word categories in the English dataset, and Table 12b presents the distribution of inflection categories.

## E.2 Tokenization Statistics

An important consideration for our analysis is how different models tokenize the words in our dataset. Table 14 shows tokenization statistics across the models we analyze. Encoder-only models like `BERT` and `DeBERTa` tend to split words into more tokens than decoder-only models like `GPT-2` and `Qwen2`, which may affect how information is encoded across layers.

## E.3 Effects of Tokenization

Tokenization is an essential component of language modeling. To test how tokenization influences our findings, we examine whether tokenizing versus not affects the encoding of linguistic information. We measure this with analogy completion tasks (*e.g.*, *man:king::woman:?*) using the embedding layer. When a word is split into subtokens, we compare two approaches to building representations - averaging versus summing those subtoken embeddings.

For each approach, we perform vector arithmetic on word representations (*e.g.*, *king - man + woman*). We measure performance by ranking all vocabulary words by cosine similarity to the resulting representation, and observe how highly the expected word (*e.g.*, *queen*) ranks, with a lower rank indicating better performance.

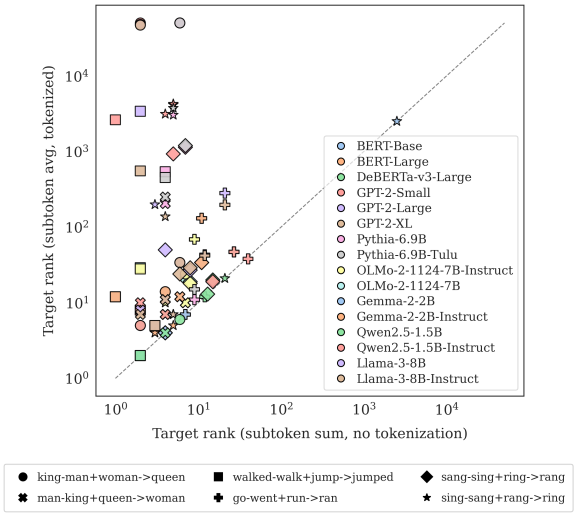



Figure 10: Effect of tokenization strategy on analogy completion rank. Each point corresponds to a model (color) and analogy (shape). The x-axis is the rank using whole-word representations. The y-axis is the rank using tokenized representations. Here, rank means the position of the expected word when all vocabulary words are sorted by similarity to the resulting embedding from vector arithmetic; lower is better. Points above the gray $y{=}x$ line mean tokenization hurts performance.

Figure 10 shows that whole-word representations yield markedly better analogy performance than averaged subtokens across all models. This implies that linguistic regularities are primarily stored in whole-word embeddings rather than compositionally across subtokens. Despite these tokenization effects, our classifier results show consistent patterns across models using different tokenizers (see Table 6), indicating that the separation of lexical and morphological information is robust.

# F Additional Analysis

## F.1 Linear Effective Dimensionality Results

This section presents detailed linear effective dimensionality analyses showing how representation compression varies across layers and between models.

Table 16 provides a numerical summary, reporting the number of principal components required to reach 50%, 70%, and 90% explained variance at the first, middle, and final layers of each model.

Figure 11 visualizes these relationships as curves for all models, with each subplot showing how

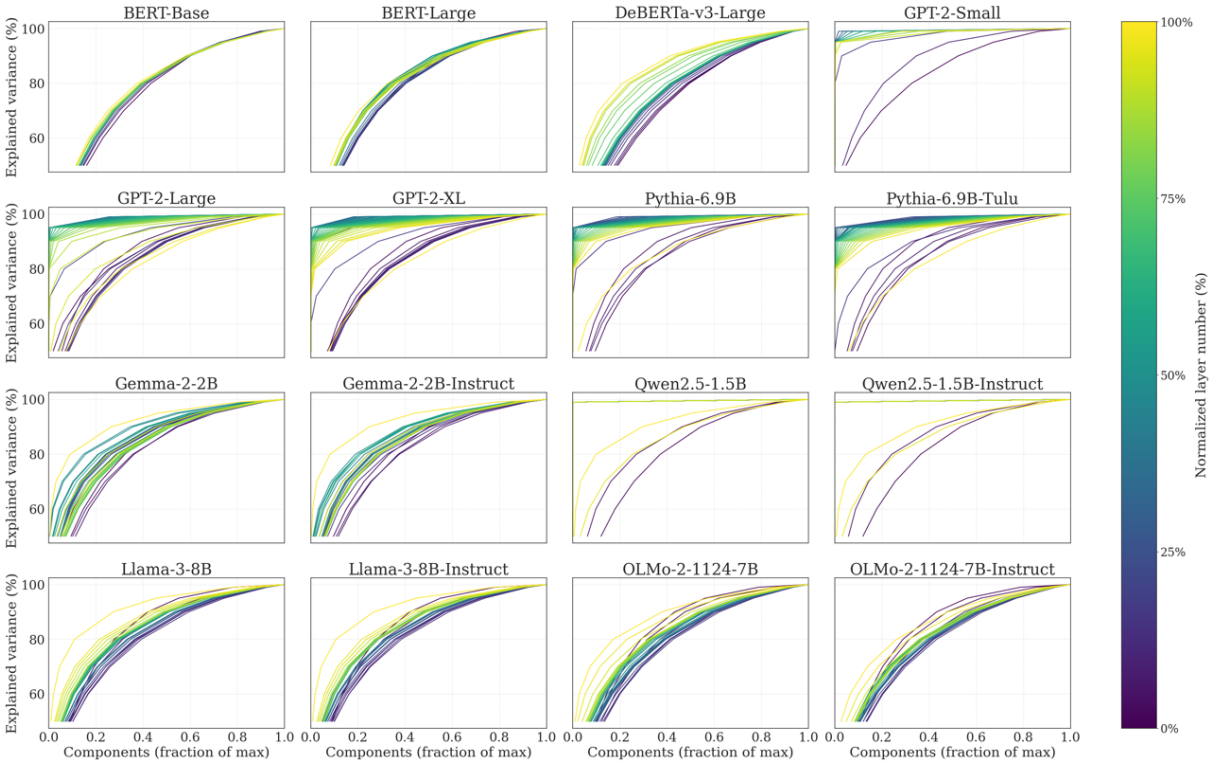



Figure 11: Linear effective dimensionality curves for all models for English. Each subplot shows the relationship between the percentage of maximum PCA components (x-axis) and the percentage of explained variance (y-axis) across different layers. The color gradient from purple (early layers, 0%) to yellow (late layers, 100%) indicates the relative layer depth within each model. Models like `BERT`, `Gemma`, and `Llama` show similar compression patterns, while `GPT-2` variants, `Qwen` and `Pythia` exhibit opposite trends in their middle layers.

explained variance accumulates as a function of the percentage of maximum PCA components, color-coded by relative layer depth.

### F.2 Massive Activations and Outlier Dimensions

We computed the maximum absolute activation, maximum mean (absolute value) per dimension, and maximum standard deviation per dimension across all layers for representative models to understand the low linear effective dimensionality observed in Table Table 16.

Figures Figures 12–18 show the results. Models like `Qwen2.5-1.5B` and `GPT-2` variants show large maximum activation values. For example, `Qwen2.5-1.5B` reaches maximum absolute activations around 8000, while models like `Llama-3-8B` and `OLMo-1124-7B` show gradual increases across layers, with maximum values only reaching 30-40 in final layers.

This corresponds with the linear effective dimensionality measurements in Table Table 16. Models with large activations in middle layers correspond to those requiring only 1-2 components to reach 50-90% explained variance at those depths. Models with gradual activation increases correspond to those requiring hundreds of components at all depths. The presence of outlier dimensions with large activations makes the representation anisotropic, with variance concentrated along a small number of directions.

### F.3 Linear Separability Gap

To determine whether the extracted linguistic signal benefits from non-linear decision boundaries, we compute the linear separability gap (defined in equation (4)) as the difference in selectivity between the MLP and linear probes. A positive gap means the MLP extracts more *selective* signal than the linear probe, while a negative gap indicates that additional probe capacity decreases selectivity, consistent with the MLP exploiting spurious correlations rather than improving extraction of genuine linguistic structure.

Figure 19 shows these results for English. For both inflection (§F.3) and lemma (§F.3), gaps are typically near zero or negative across much of the depth range, indicating that linear probes are at

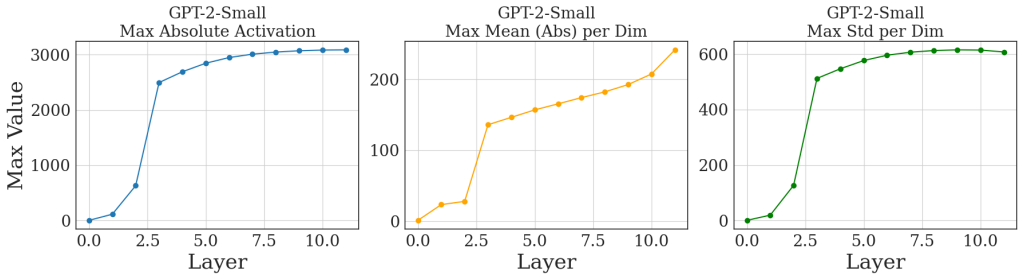



Figure 12: Activation statistics across layers for `GPT-2-Small`.

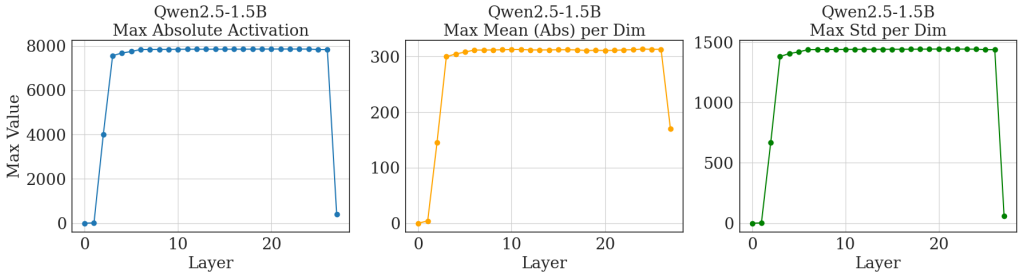



Figure 13: Activation statistics across layers for `Qwen2.5-1.5B`.

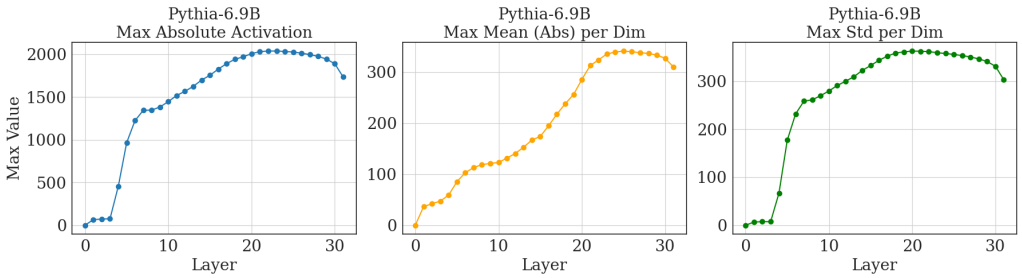



Figure 14: Activation statistics across layers for `Pythia-6.9B`.

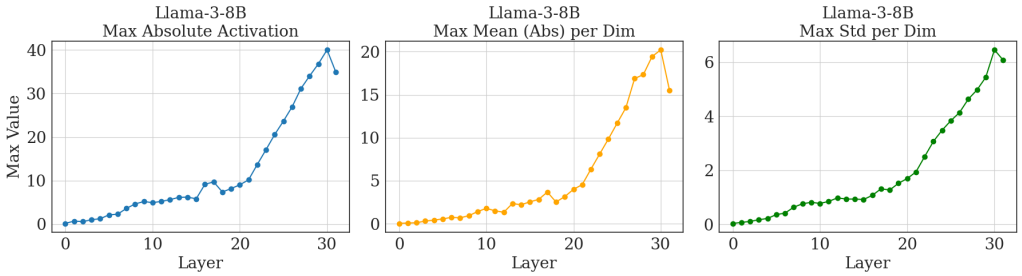



Figure 15: Activation statistics across layers for `Llama-3-8B`.

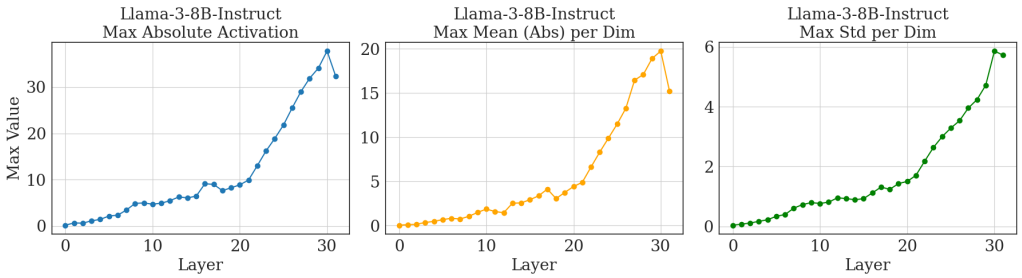



Figure 16: Activation statistics across layers for `Llama-3-8B-Instruct`.

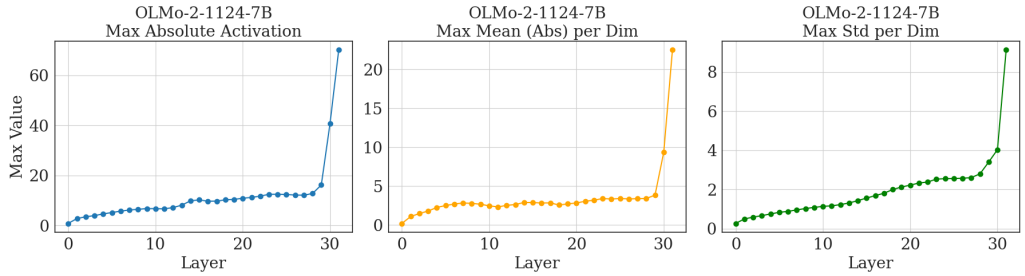



Figure 17: Activation statistics across layers for `OLMo-2-1124-7B`.

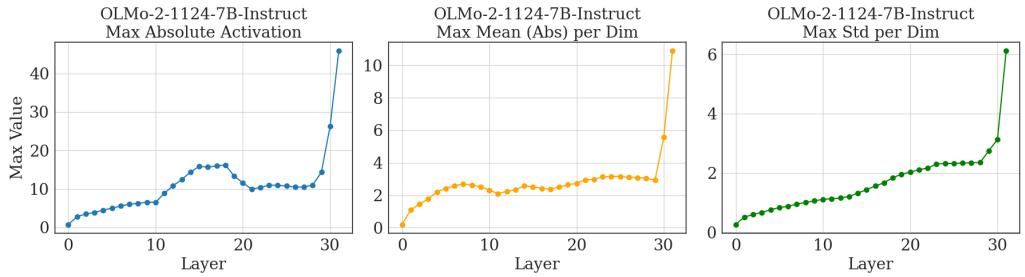



Figure 18: Activation statistics across layers for `OLMo-2-1124-7B-Instruct`.

least as selective as MLP probes in most models and layers. This is particularly salient for lemma: despite occasional accuracy improvements from the MLP, the selectivity gap often remains negative, suggesting that the additional accuracy is frequently driven by non-linguistic memorization effects rather than more structured lexical encoding.

Figures 20–24 extend this analysis cross-linguistically. Across Chinese, German, French, and Russian, the gap is predominantly negative for inflection and generally slightly negative (or near zero) for lemma, again implying that increased non-linear capacity rarely improves selectivity. Turkish is a notable exception: for several model families the gap is positive (especially for lemma), indicating cases where an MLP can extract additional selective signal beyond a linear readout. Overall, these trends suggest that by the selectivity criterion, most of the recoverable lexical and morphological signal is already accessible to a regularized linear probe, and extra probe capacity more often harms than helps.

### F.4 Training Dynamics

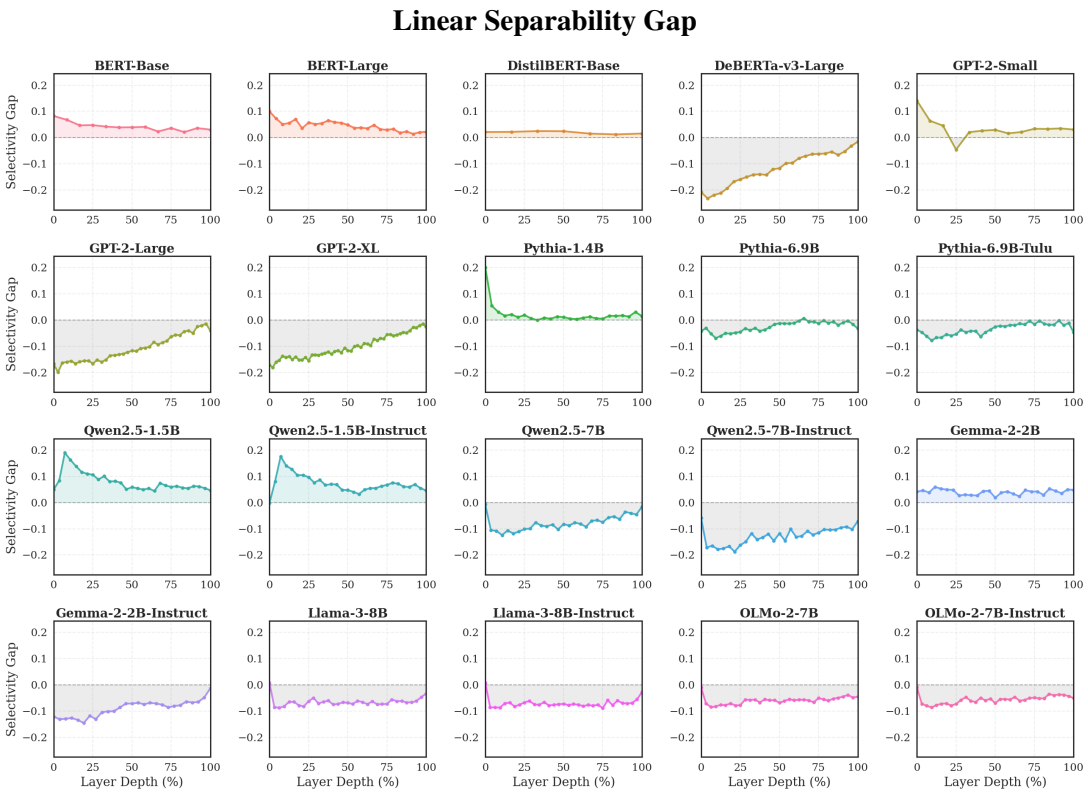
We analyze how linguistic information encoding evolves during the pretraining process. Figure 25 and Figure 26 visualize the probing performance across training checkpoints for `OLMo-2-7B` and `Pythia-6.9B`, respectively.

For `OLMo-2-7B` (Figure 25), we observe a counter-intuitive trend in lemma accuracy: early checkpoints (lighter curves) actually exhibit higher accuracy than the fully trained model (darker curves). This suggests that as the model matures, it abstracts away from rigid surface-level lemma identities, making them harder to probe linearly. However, the selectivity scores (right) consistently increase with training steps, indicating that while raw accuracy drops, the model becomes better at distinguishing linguistic features from control tasks.

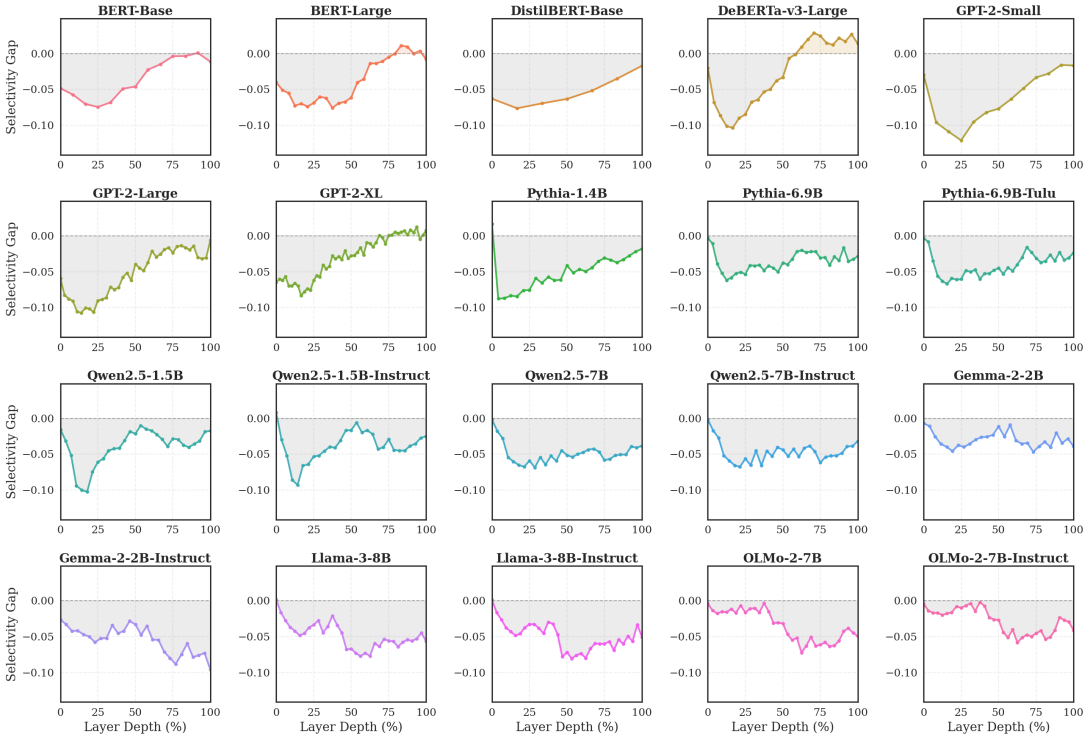
A similar pattern is observed for `Pythia-6.9B` (Figure 26). Lemma accuracy significantly declines both with increasing layer depth and with more training steps. Conversely, inflectional accuracy remains robust and high throughout training. The selectivity plots demonstrate a clear separation between early and late checkpoints, particularly for inflection tasks. This confirms that pretraining helps the model refine its internal representations, suppressing irrelevant correlations (the control task) while maintaining necessary information.

## G Attention Head Analysis

We conducted additional experiments analyzing attention head outputs alongside residual stream representations to understand how different components of transformer models contribute to linguistic encoding.

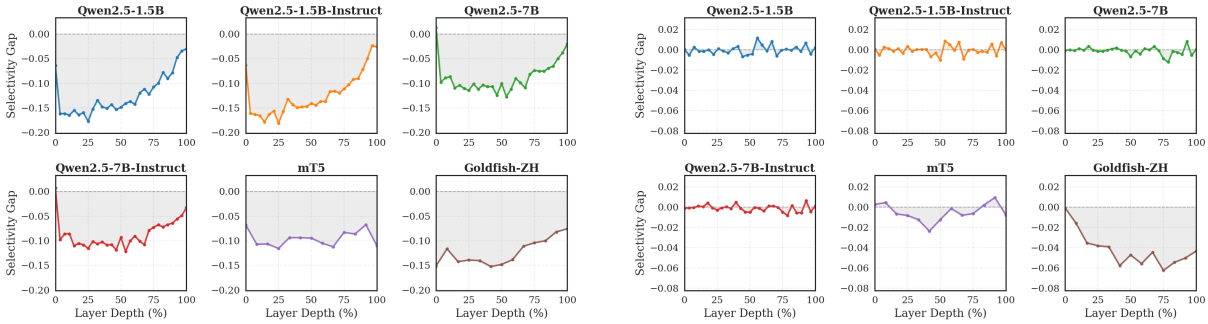



**(a)** Linear separability gap for inflection prediction



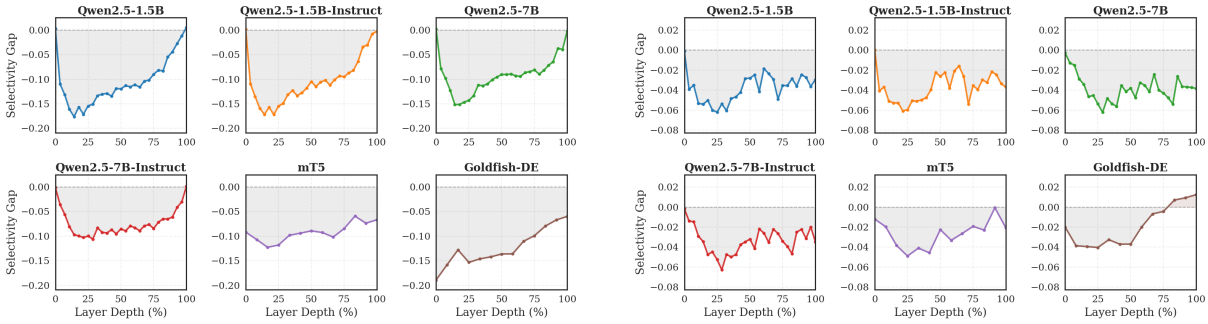

**(b)** Linear separability gap for lemma prediction

Figure 19: Linear separability gap (difference in selectivity between MLP and linear probes) across model layers for English. The gap measures how much a non-linear transformation improves the extraction of genuine linguistic signal compared to a simple linear mapping.



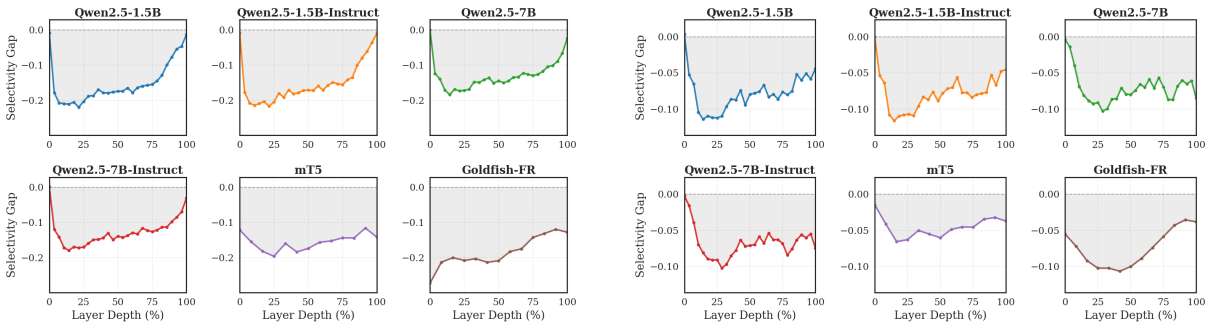

**(a)** Inflection **(b)** Lemma

Figure 20: Linear separability gap for Chinese.



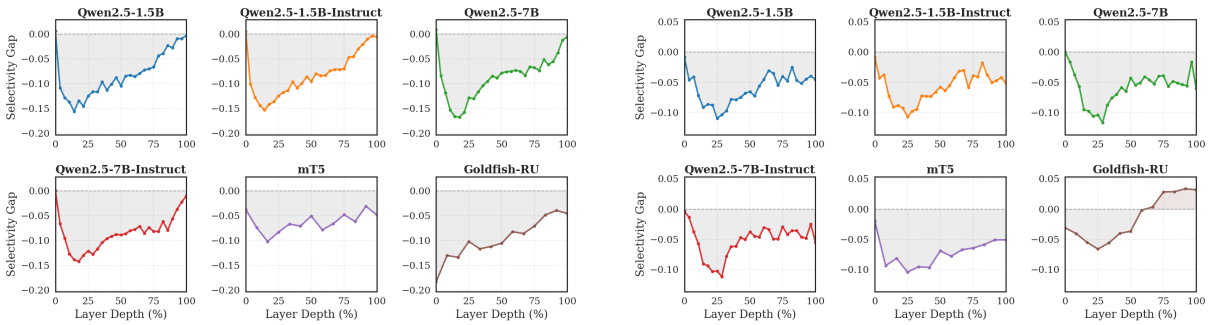

**(a)** Inflection **(b)** Lemma

Figure 21: Linear separability gap for German.



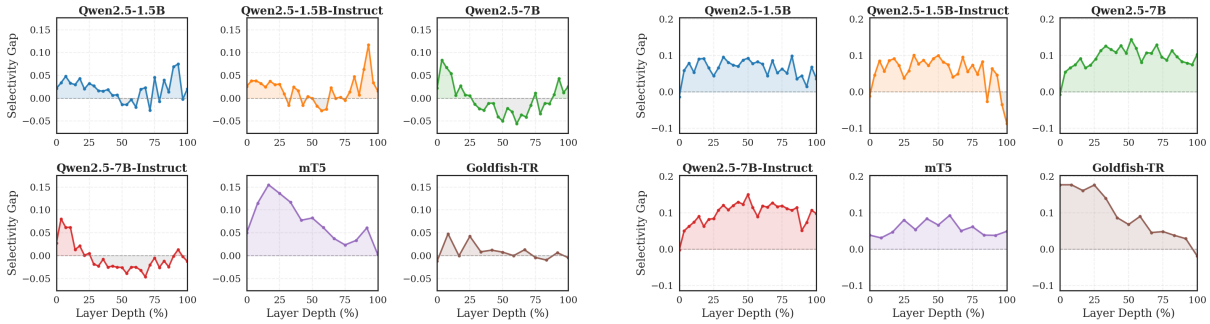

**(a)** Inflection **(b)** Lemma

Figure 22: Linear separability gap for French.




**(a)** Inflection **(b)** Lemma

Figure 23: Linear separability gap for Russian.




**(a)** Inflection **(b)** Lemma

Figure 24: Linear separability gap for Turkish.

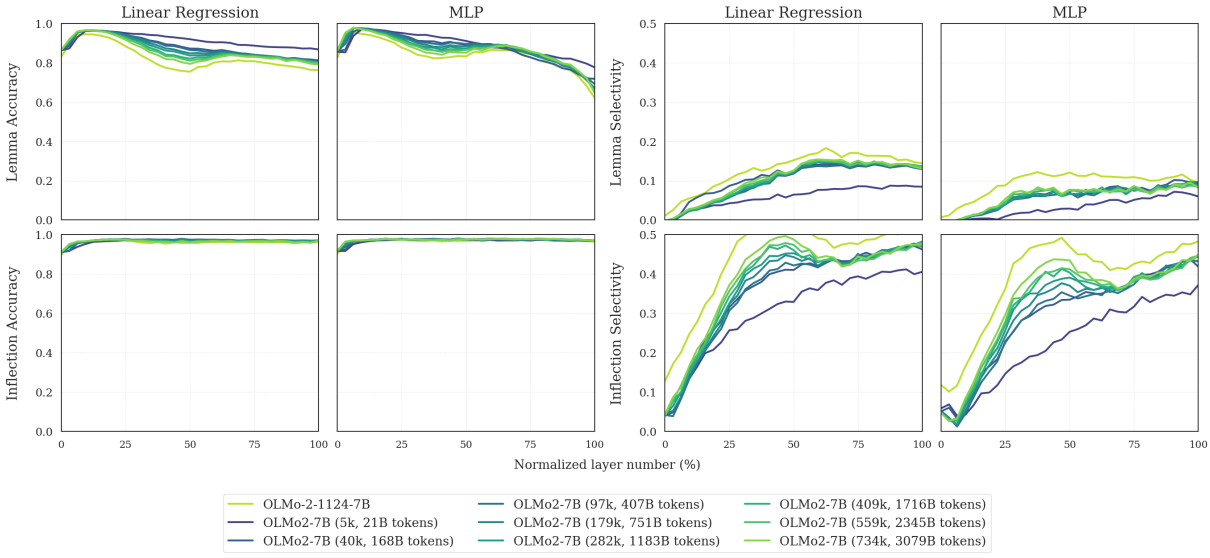



Figure 25: `OLMo-2-7B` **Training Dynamics.** Performance across pretraining checkpoints (5k–734k steps) for English. The full model is 928k steps. Checkpoints are colored from brightest (earliest) to darkest (latest). **Left:** Prediction accuracy for Lemma (top) and Inflection (bottom). Early checkpoints exhibit higher lemma accuracy than later ones, while inflectional accuracy remains flat. **Right:** Selectivity scores for the same tasks. Selectivity generally increases with model depth and training steps, particularly for inflection.

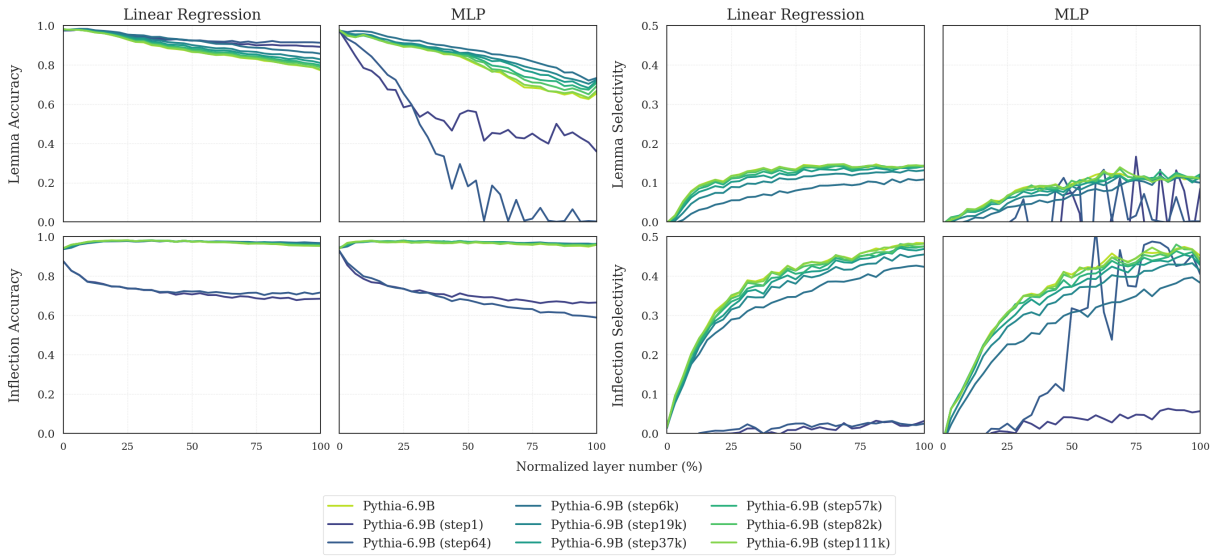



Figure 26: `Pythia-6.9B` **Training Dynamics.** Performance across pretraining checkpoints (step 1–111k) for English. The full model is 143k steps. Checkpoints are colored from brightest (earliest) to darkest (latest). **Left:** Prediction accuracy for Lemma (top) and Inflection (bottom). Lemma accuracy declines both with deeper layers and with more training, whereas inflectional accuracy stays uniformly high. **Right:** Selectivity scores for the same tasks, showing distinct separation between early and late checkpoints in the inflection task.

### G.1 Methodology

We averaged activations across all attention heads at each layer for `Qwen2.5-1.5B` and `Qwen2.5-1.5B-Instruct` models using the English dataset. We then trained linear regression and MLP classifiers on both attention head outputs and residual stream representations to compare their encoding patterns.

### G.2 Results

Figures 27 and 28 compare probe accuracy and selectivity when trained on attention head outputs versus residual stream representations for lemma and inflection tasks across all English model families. Across both encoder-only and decoder-only architectures, residual stream representations consistently yield higher accuracy.

## H Steering Vector Analysis

We conducted steering vector experiments to test whether inflectional representations can be functionally manipulated and to understand model sensitivity to activation interventions.

### H.1 Methodology

For each inflectional category, we computed steering vectors as:

$$\mathbf{s}_i = \mu_i - \lambda \cdot \frac{1}{|C| - 1} \sum_{j \in C, j \neq i} \mu_j \qquad (5)$$

We tested multiple values of $\lambda$ (1, 5, 10, 20, 100) and measured the impact on linear classifier performance when adding these steering vectors to existing activations for 1000 test words. We evaluated both mean probability change and prediction flip rate across all models.

### H.2 Results

Figure 29 shows the mean probability change for inflection prediction when applying steering vectors across different $\lambda$ values (1, 5, 10, 20, and 100). Each panel demonstrates that steering effectiveness is high across all models, with most maintaining stable performance throughout network depth. A notable exception is `DeBERTa-v3-Large`, which shows a sudden drop in steering effectiveness around 75% of model depth.

Figure 30 presents the corresponding prediction flip rates, which measure how often steering vectors successfully change the classifier's prediction. The patterns mirror the probability change results, with most models maintaining high flip rates (0.98–1.00) throughout all layers, except for `DeBERTa-v3-Large`, which exhibits a similar drop around 75% of model depth.

## I Classifier Error Analysis

We conducted a detailed error analysis of our classifiers to better understand their performance across different morphological features and languages.

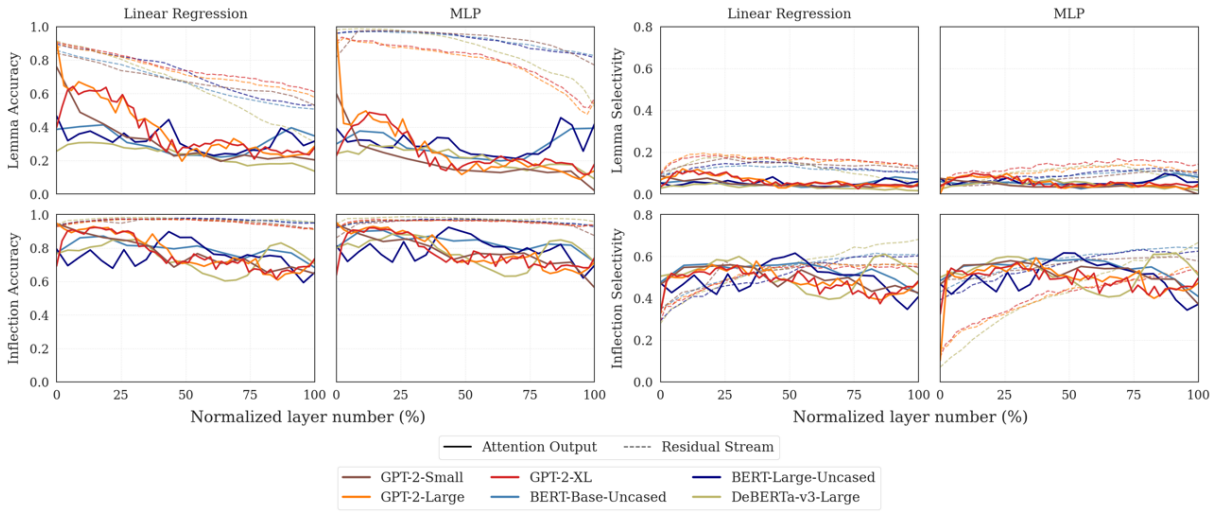



Figure 27: Linguistic task accuracy (left two columns) and classifier selectivity (right two columns) for attention head outputs (solid lines) versus residual stream representations (dashed lines) across BERT and GPT-2 model families. The top row corresponds to Lemma tasks, and the bottom row to Inflection tasks.

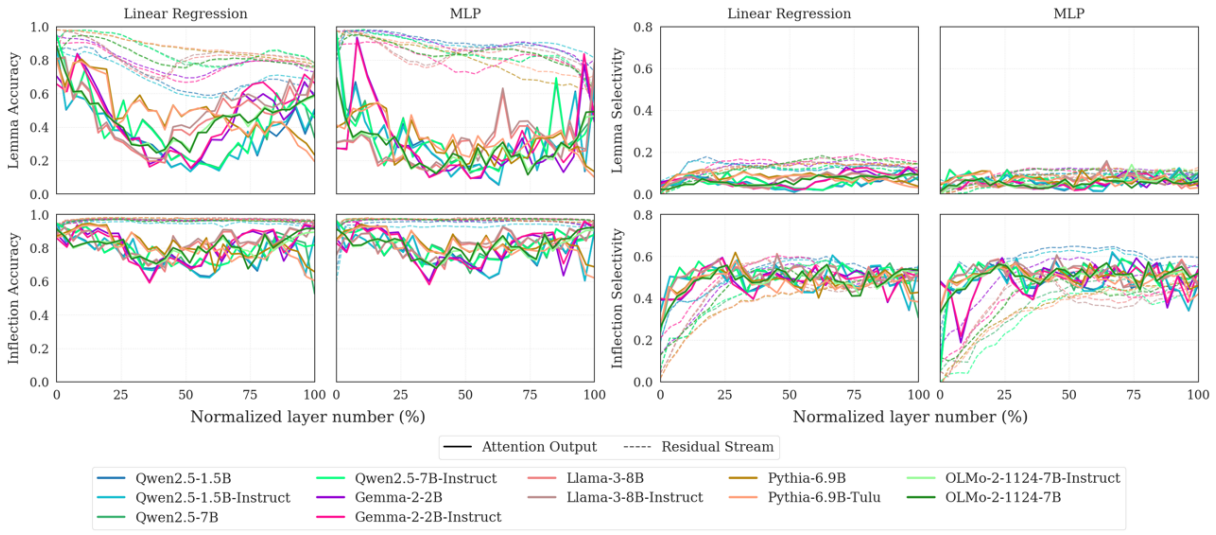



Figure 28: Linguistic task accuracy (left two columns) and classifier selectivity (right two columns) for attention head outputs (solid lines) versus residual stream representations (dashed lines) across contemporary model families. The top row corresponds to Lemma tasks, and the bottom row to Inflection tasks.

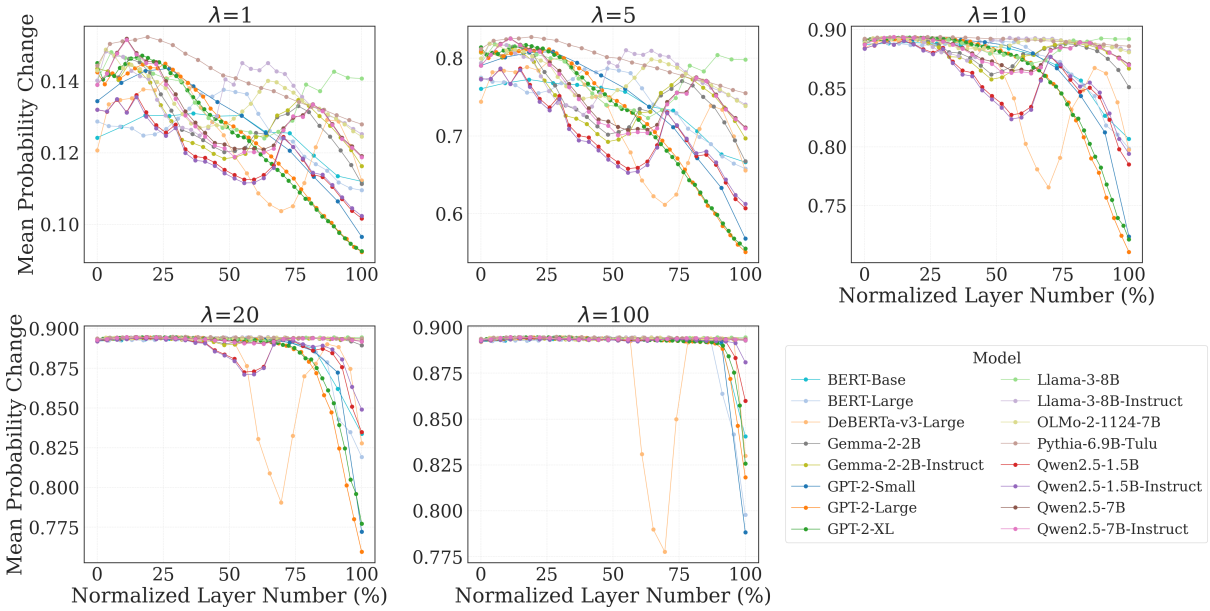



Figure 29: Mean probability change for inflection prediction when applying steering vectors across different $\lambda$ values. Five panels show results for $\lambda \in \{1, 5, 10, 20, 100\}$. All models start with high effectiveness (≈0.9-1.0) at layer 0. Most models maintain stable performance across depth, with `DeBERTa-v3-Large` a notable exception, showing a sudden drop around 75% of model depth. Higher $\lambda$ values increase steering effectiveness while preserving the overall pattern.

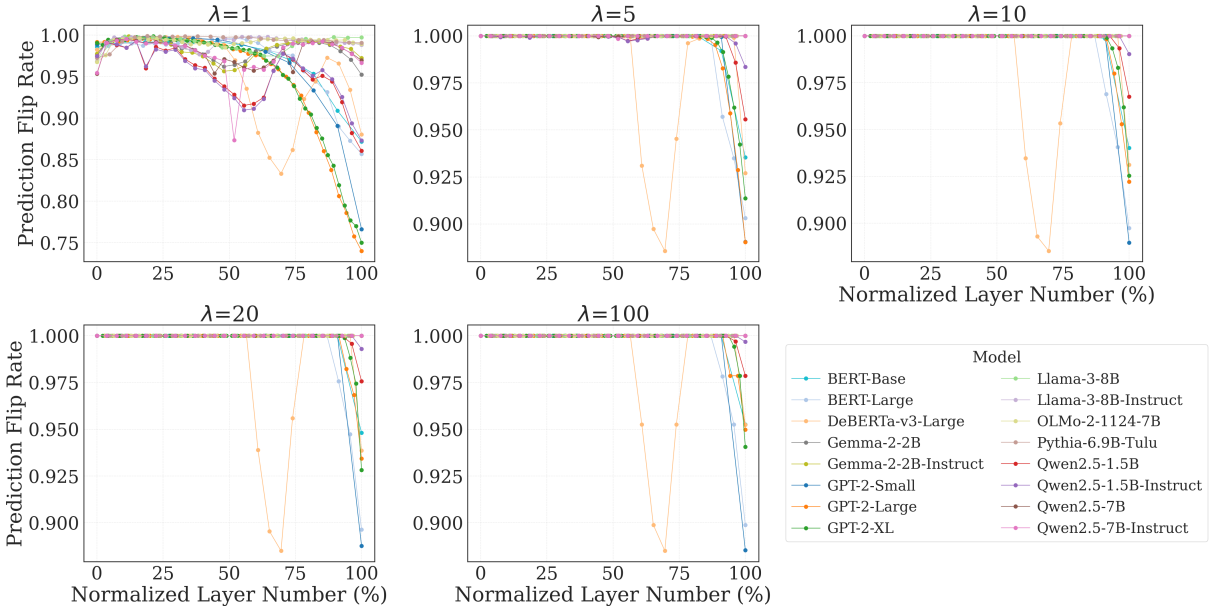



Figure 30: Prediction flip rate when applying steering vectors across different $\lambda$ values. The flip rate patterns mirror the probability change results, with most models maintaining high rates (0.98-1.00) throughout all layers. `DeBERTa-v3-Large` is an exception, showing a sudden drop around 75% of model depth. The consistency across $\lambda$ values suggests that steering effectiveness depends more on model architecture than intervention strength.

**English Results**

- Table 17: Inflection accuracy by morphological feature (linear probes)
- Table 18: Inflection accuracy by morphological feature (MLP probes)
- Table 19: Lemma accuracy by part of speech (linear probes)
- Table 20: Lemma accuracy by part of speech (MLP probes)

**Chinese Results**

- Table 21: Inflection accuracy by feature (linear and MLP)
- Table 22: Lemma accuracy by POS (linear probes)
- Table 23: Lemma accuracy by POS (MLP probes)

**German Results**

- Table 24: Inflection accuracy by feature (linear probes)
- Table 25: Inflection accuracy by feature (MLP probes)
- Table 26: Lemma accuracy by POS (linear and MLP)

**French Results**

- Table 27: Inflection accuracy by feature (linear probes)
- Table 28: Inflection accuracy by feature (MLP probes)
- Table 29: Lemma accuracy by POS (linear and MLP)

**Russian Results**

- Table 30: Inflection accuracy by feature (linear probes)
- Table 31: Inflection accuracy by feature (MLP probes)
- Table 32: Lemma accuracy by POS (linear and MLP)

**Turkish Results**

- Table 33: Inflection accuracy by feature (linear probes)
- Table 34: Inflection accuracy by feature (MLP probes)
- Table 35: Lemma accuracy by POS (linear and MLP)

Table 2: English Accuracy (Linear Probes) Across Model Depth

| | | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| **Model** | **Task** | **0%** | **25%** | **50%** | **75%** | **100%** |
| BERT-Base | Inflection | 0.934 | 0.971 | 0.977 | 0.966 | 0.951 |
| | Lexeme | 0.858 | 0.773 | 0.665 | 0.564 | 0.507 |
| BERT-Large | Inflection | 0.938 | 0.972 | 0.978 | 0.966 | 0.949 |
| | Lexeme | 0.896 | 0.820 | 0.737 | 0.585 | 0.531 |
| DeBERTa-v3-Large | Inflection | 0.939 | 0.982 | 0.972 | 0.960 | 0.955 |
| | Lexeme | 0.914 | 0.805 | 0.673 | 0.482 | 0.309 |
| Gemma-2-2B | Inflection | 0.936 | 0.974 | 0.963 | 0.972 | 0.948 |
| | Lexeme | 0.944 | 0.817 | 0.694 | 0.792 | 0.728 |
| Gemma-2-2B-Instruct | Inflection | 0.917 | 0.971 | 0.960 | 0.972 | 0.960 |
| | Lexeme | 0.904 | 0.802 | 0.667 | 0.791 | 0.752 |
| Goldfish English | Inflection | 0.934 | 0.956 | 0.972 | 0.964 | 0.949 |
| | Lexeme | 0.926 | 0.859 | 0.755 | 0.690 | 0.652 |
| GPT-2-Large | Inflection | 0.927 | 0.970 | 0.971 | 0.946 | 0.912 |
| | Lexeme | 0.874 | 0.818 | 0.711 | 0.665 | 0.577 |
| GPT-2-Small | Inflection | 0.939 | 0.948 | 0.971 | 0.956 | 0.909 |
| | Lexeme | 0.840 | 0.737 | 0.671 | 0.618 | 0.530 |
| GPT-2-XL | Inflection | 0.929 | 0.974 | 0.973 | 0.946 | 0.915 |
| | Lexeme | 0.906 | 0.827 | 0.737 | 0.690 | 0.609 |
| Llama-3-8B | Inflection | 0.912 | 0.971 | 0.974 | 0.976 | 0.962 |
| | Lexeme | 0.864 | 0.794 | 0.796 | 0.816 | 0.749 |
| Llama-3-8B-Instruct | Inflection | 0.913 | 0.972 | 0.974 | 0.977 | 0.967 |
| | Lexeme | 0.864 | 0.803 | 0.812 | 0.839 | 0.783 |
| mT5-Base | Inflection | 0.920 | 0.966 | 0.969 | 0.966 | 0.958 |
| | Lexeme | 0.868 | 0.862 | 0.731 | 0.634 | 0.619 |
| OLMo-2-1124-7B | Inflection | 0.897 | 0.970 | 0.959 | 0.961 | 0.965 |
| | Lexeme | 0.832 | 0.883 | 0.755 | 0.808 | 0.763 |
| OLMo-2-1124-7B-Instruct | Inflection | 0.897 | 0.970 | 0.958 | 0.962 | 0.961 |
| | Lexeme | 0.832 | 0.881 | 0.749 | 0.806 | 0.757 |
| Pythia-6.9B | Inflection | 0.942 | 0.982 | 0.974 | 0.966 | 0.953 |
| | Lexeme | 0.980 | 0.928 | 0.865 | 0.829 | 0.772 |
| Pythia-6.9B-Tulu | Inflection | 0.941 | 0.980 | 0.975 | 0.968 | 0.954 |
| | Lexeme | 0.980 | 0.928 | 0.872 | 0.841 | 0.789 |
| Qwen2.5-1.5B | Inflection | 0.913 | 0.969 | 0.959 | 0.961 | 0.930 |
| | Lexeme | 0.845 | 0.799 | 0.610 | 0.654 | 0.599 |
| Qwen2.5-1.5B-Instruct | Inflection | 0.910 | 0.957 | 0.944 | 0.949 | 0.914 |
| | Lexeme | 0.876 | 0.768 | 0.590 | 0.647 | 0.654 |
| Qwen2.5-7B | Inflection | 0.915 | 0.977 | 0.966 | 0.973 | 0.958 |
| | Lexeme | 0.916 | 0.952 | 0.769 | 0.808 | 0.781 |
| Qwen2.5-7B-Instruct | Inflection | 0.915 | 0.977 | 0.964 | 0.973 | 0.957 |
| | Lexeme | 0.916 | 0.950 | 0.791 | 0.810 | 0.781 |

Table 3: English Selectivity (Linear Probes) Across Model Depth

| | | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| **Model** | **Task** | **0%** | **25%** | **50%** | **75%** | **100%** |
| BERT-Base | Inflection | 0.348 | 0.486 | 0.556 | 0.596 | 0.609 |
| | Lexeme | 0.094 | 0.132 | 0.134 | 0.114 | 0.101 |
| BERT-Large | Inflection | 0.292 | 0.459 | 0.520 | 0.598 | 0.603 |
| | Lexeme | 0.083 | 0.145 | 0.148 | 0.119 | 0.107 |
| DeBERTa-v3-Large | Inflection | 0.279 | 0.471 | 0.545 | 0.644 | 0.681 |
| | Lexeme | 0.062 | 0.183 | 0.158 | 0.097 | 0.054 |
| Gemma-2-2B | Inflection | 0.131 | 0.466 | 0.565 | 0.464 | 0.512 |
| | Lexeme | 0.036 | 0.132 | 0.136 | 0.161 | 0.128 |
| Gemma-2-2B-Instruct | Inflection | 0.210 | 0.517 | 0.594 | 0.487 | 0.522 |
| | Lexeme | 0.052 | 0.157 | 0.144 | 0.184 | 0.154 |
| Goldfish English | Inflection | 0.341 | 0.437 | 0.522 | 0.549 | 0.556 |
| | Lexeme | 0.017 | 0.073 | 0.096 | 0.100 | 0.110 |
| GPT-2-Large | Inflection | 0.324 | 0.489 | 0.563 | 0.550 | 0.562 |
| | Lexeme | 0.099 | 0.183 | 0.162 | 0.161 | 0.121 |
| GPT-2-Small | Inflection | 0.347 | 0.576 | 0.541 | 0.558 | 0.547 |
| | Lexeme | 0.100 | 0.171 | 0.152 | 0.134 | 0.121 |
| GPT-2-XL | Inflection | 0.289 | 0.488 | 0.549 | 0.547 | 0.553 |
| | Lexeme | 0.092 | 0.172 | 0.172 | 0.162 | 0.135 |
| Llama-3-8B | Inflection | 0.041 | 0.516 | 0.482 | 0.474 | 0.497 |
| | Lexeme | -0.003 | 0.132 | 0.146 | 0.140 | 0.117 |
| Llama-3-8B-Instruct | Inflection | 0.042 | 0.508 | 0.473 | 0.453 | 0.478 |
| | Lexeme | -0.003 | 0.131 | 0.146 | 0.144 | 0.119 |
| mT5-Base | Inflection | 0.229 | 0.348 | 0.462 | 0.533 | 0.530 |
| | Lexeme | 0.003 | 0.005 | 0.047 | 0.053 | 0.063 |
| OLMo-2-1124-7B | Inflection | 0.128 | 0.440 | 0.538 | 0.476 | 0.528 |
| | Lexeme | 0.011 | 0.104 | 0.152 | 0.170 | 0.144 |
| OLMo-2-1124-7B-Instruct | Inflection | 0.127 | 0.443 | 0.545 | 0.483 | 0.543 |
| | Lexeme | 0.011 | 0.104 | 0.149 | 0.164 | 0.144 |
| Pythia-6.9B | Inflection | 0.017 | 0.347 | 0.416 | 0.457 | 0.484 |
| | Lexeme | -0.002 | 0.117 | 0.134 | 0.140 | 0.141 |
| Pythia-6.9B-Tulu | Inflection | 0.016 | 0.348 | 0.419 | 0.454 | 0.492 |
| | Lexeme | -0.002 | 0.118 | 0.137 | 0.143 | 0.150 |
| Qwen2.5-1.5B | Inflection | 0.199 | 0.483 | 0.589 | 0.566 | 0.548 |
| | Lexeme | 0.034 | 0.146 | 0.117 | 0.129 | 0.108 |
| Qwen2.5-1.5B-Instruct | Inflection | 0.189 | 0.474 | 0.581 | 0.541 | 0.503 |
| | Lexeme | 0.061 | 0.143 | 0.113 | 0.129 | 0.112 |
| Qwen2.5-7B | Inflection | 0.059 | 0.383 | 0.538 | 0.533 | 0.523 |
| | Lexeme | -0.006 | 0.098 | 0.121 | 0.131 | 0.099 |
| Qwen2.5-7B-Instruct | Inflection | 0.059 | 0.392 | 0.542 | 0.531 | 0.528 |
| | Lexeme | -0.006 | 0.098 | 0.116 | 0.129 | 0.100 |

Table 4: English Accuracy (MLP Probes) Across Model Depth

| | | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| **Model** | **Task** | **0%** | **25%** | **50%** | **75%** | **100%** |
| BERT-Base | Inflection | 0.941 | 0.978 | 0.983 | 0.975 | 0.960 |
| | Lexeme | 0.906 | 0.888 | 0.808 | 0.719 | 0.629 |
| BERT-Large | Inflection | 0.943 | 0.977 | 0.981 | 0.973 | 0.958 |
| | Lexeme | 0.920 | 0.904 | 0.843 | 0.762 | 0.677 |
| DeBERTa-v3-Large | Inflection | 0.944 | 0.986 | 0.976 | 0.971 | 0.957 |
| | Lexeme | 0.920 | 0.885 | 0.781 | 0.574 | 0.318 |
| Gemma-2-2B | Inflection | 0.940 | 0.977 | 0.970 | 0.975 | 0.946 |
| | Lexeme | 0.939 | 0.868 | 0.732 | 0.812 | 0.506 |
| Gemma-2-2B-Instruct | Inflection | 0.919 | 0.973 | 0.965 | 0.973 | 0.960 |
| | Lexeme | 0.890 | 0.874 | 0.731 | 0.831 | 0.508 |
| Goldfish English | Inflection | 0.937 | 0.962 | 0.976 | 0.971 | 0.964 |
| | Lexeme | 0.952 | 0.922 | 0.871 | 0.790 | 0.656 |
| GPT-2-Large | Inflection | 0.930 | 0.964 | 0.964 | 0.952 | 0.937 |
| | Lexeme | 0.892 | 0.878 | 0.825 | 0.719 | 0.552 |
| GPT-2-Small | Inflection | 0.943 | 0.963 | 0.965 | 0.957 | 0.837 |
| | Lexeme | 0.830 | 0.843 | 0.817 | 0.705 | 0.075 |
| GPT-2-XL | Inflection | 0.929 | 0.965 | 0.963 | 0.948 | 0.937 |
| | Lexeme | 0.906 | 0.884 | 0.844 | 0.734 | 0.571 |
| Llama-3-8B | Inflection | 0.920 | 0.972 | 0.977 | 0.976 | 0.958 |
| | Lexeme | 0.863 | 0.808 | 0.857 | 0.840 | 0.568 |
| Llama-3-8B-Instruct | Inflection | 0.920 | 0.972 | 0.977 | 0.977 | 0.964 |
| | Lexeme | 0.863 | 0.821 | 0.873 | 0.870 | 0.605 |
| mT5-Base | Inflection | NaN | NaN | NaN | NaN | NaN |
| | Lexeme | 0.871 | 0.845 | 0.744 | 0.686 | 0.722 |
| OLMo-2-1124-7B | Inflection | 0.903 | 0.974 | 0.970 | 0.975 | 0.964 |
| | Lexeme | 0.825 | 0.877 | 0.833 | 0.845 | 0.621 |
| OLMo-2-1124-7B-Instruct | Inflection | 0.903 | 0.975 | 0.968 | 0.973 | 0.964 |
| | Lexeme | 0.825 | 0.880 | 0.825 | 0.847 | 0.650 |
| Pythia-6.9B | Inflection | 0.940 | 0.973 | 0.971 | 0.959 | 0.959 |
| | Lexeme | 0.976 | 0.891 | 0.823 | 0.683 | 0.655 |
| Pythia-6.9B-Tulu | Inflection | 0.944 | 0.973 | 0.970 | 0.963 | 0.961 |
| | Lexeme | 0.976 | 0.904 | 0.858 | 0.752 | 0.709 |
| Qwen2.5-1.5B | Inflection | 0.666 | 0.956 | 0.956 | 0.961 | 0.929 |
| | Lexeme | 0.792 | 0.959 | 0.901 | 0.886 | 0.731 |
| Qwen2.5-1.5B-Instruct | Inflection | 0.598 | 0.922 | 0.928 | 0.942 | 0.913 |
| | Lexeme | 0.852 | 0.939 | 0.880 | 0.900 | 0.812 |
| Qwen2.5-7B | Inflection | 0.919 | 0.970 | 0.963 | 0.970 | 0.953 |
| | Lexeme | 0.913 | 0.935 | 0.831 | 0.818 | 0.506 |
| Qwen2.5-7B-Instruct | Inflection | 0.930 | 0.976 | 0.970 | 0.976 | 0.951 |
| | Lexeme | 0.913 | 0.933 | 0.824 | 0.818 | 0.521 |

Table 5: English Selectivity (MLP Probes) Across Model Depth

| | | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| **Model** | **Task** | **0%** | **25%** | **50%** | **75%** | **100%** |
| BERT-Base | Inflection | 0.139 | 0.307 | 0.416 | 0.505 | 0.536 |
| | Lexeme | 0.025 | 0.057 | 0.069 | 0.084 | 0.068 |
| BERT-Large | Inflection | 0.104 | 0.294 | 0.386 | 0.496 | 0.514 |
| | Lexeme | 0.028 | 0.084 | 0.088 | 0.106 | 0.074 |
| DeBERTa-v3-Large | Inflection | 0.071 | 0.313 | 0.428 | 0.582 | 0.666 |
| | Lexeme | 0.005 | 0.061 | 0.096 | 0.065 | 0.043 |
| Gemma-2-2B | Inflection | 0.002 | 0.360 | 0.499 | 0.395 | 0.483 |
| | Lexeme | -0.007 | 0.067 | 0.090 | 0.088 | 0.054 |
| Gemma-2-2B-Instruct | Inflection | 0.089 | 0.393 | 0.523 | 0.404 | 0.509 |
| | Lexeme | 0.026 | 0.102 | 0.111 | 0.100 | 0.058 |
| Goldfish English | Inflection | 0.131 | 0.272 | 0.371 | 0.438 | 0.493 |
| | Lexeme | 0.006 | -0.021 | 0.005 | 0.031 | 0.050 |
| GPT-2-Large | Inflection | 0.157 | 0.323 | 0.446 | 0.487 | 0.521 |
| | Lexeme | 0.040 | 0.093 | 0.123 | 0.138 | 0.115 |
| GPT-2-Small | Inflection | 0.210 | 0.342 | 0.410 | 0.470 | 0.581 |
| | Lexeme | 0.012 | 0.077 | 0.094 | 0.107 | 0.061 |
| GPT-2-XL | Inflection | 0.118 | 0.334 | 0.433 | 0.491 | 0.527 |
| | Lexeme | 0.027 | 0.110 | 0.145 | 0.163 | 0.142 |
| Llama-3-8B | Inflection | 0.050 | 0.454 | 0.414 | 0.402 | 0.463 |
| | Lexeme | -0.001 | 0.098 | 0.079 | 0.084 | 0.060 |
| Llama-3-8B-Instruct | Inflection | 0.050 | 0.442 | 0.400 | 0.364 | 0.451 |
| | Lexeme | -0.002 | 0.098 | 0.074 | 0.084 | 0.069 |
| mT5-Base | Inflection | NaN | NaN | NaN | NaN | NaN |
| | Lexeme | -0.012 | -0.023 | 0.003 | 0.011 | 0.026 |
| OLMo-2-1124-7B | Inflection | 0.118 | 0.363 | 0.470 | 0.426 | 0.483 |
| | Lexeme | 0.007 | 0.088 | 0.121 | 0.109 | 0.095 |
| OLMo-2-1124-7B-Instruct | Inflection | 0.118 | 0.371 | 0.476 | 0.434 | 0.493 |
| | Lexeme | 0.007 | 0.094 | 0.122 | 0.119 | 0.103 |
| Pythia-6.9B | Inflection | -0.025 | 0.301 | 0.404 | 0.444 | 0.451 |
| | Lexeme | -0.005 | 0.064 | 0.096 | 0.110 | 0.113 |
| Pythia-6.9B-Tulu | Inflection | -0.021 | 0.294 | 0.394 | 0.438 | 0.445 |
| | Lexeme | -0.005 | 0.057 | 0.092 | 0.112 | 0.126 |
| Qwen2.5-1.5B | Inflection | 0.248 | 0.588 | 0.647 | 0.625 | 0.595 |
| | Lexeme | 0.019 | 0.085 | 0.096 | 0.101 | 0.090 |
| Qwen2.5-1.5B-Instruct | Inflection | 0.187 | 0.571 | 0.629 | 0.608 | 0.549 |
| | Lexeme | 0.069 | 0.090 | 0.097 | 0.100 | 0.087 |
| Qwen2.5-7B | Inflection | 0.051 | 0.282 | 0.456 | 0.457 | 0.507 |
| | Lexeme | -0.008 | 0.039 | 0.070 | 0.073 | 0.061 |
| Qwen2.5-7B-Instruct | Inflection | 0.001 | 0.229 | 0.424 | 0.416 | 0.457 |
| | Lexeme | -0.008 | 0.042 | 0.072 | 0.067 | 0.068 |

Table 6: Probing Results for Chinese

(a) Accuracy (Linear Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish Chinese (Chinese) | Inflection | 0.911 | 0.928 | 0.944 | 0.942 | 0.941 |
| | Lexeme | 0.972 | 0.941 | 0.887 | 0.824 | 0.751 |
| Qwen2.5-1.5B (Chinese) | Inflection | 0.898 | 0.948 | 0.949 | 0.950 | 0.946 |
| | Lexeme | 0.883 | 0.905 | 0.735 | 0.743 | 0.667 |
| Qwen2.5-1.5B-Instruct (Chinese) | Inflection | 0.897 | 0.948 | 0.949 | 0.950 | 0.948 |
| | Lexeme | 0.883 | 0.907 | 0.729 | 0.748 | 0.678 |
| Qwen2.5-7B (Chinese) | Inflection | 0.893 | 0.957 | 0.951 | 0.956 | 0.950 |
| | Lexeme | 0.883 | 0.983 | 0.844 | 0.828 | 0.776 |
| Qwen2.5-7B-Instruct (Chinese) | Inflection | 0.893 | 0.957 | 0.950 | 0.956 | 0.949 |
| | Lexeme | 0.883 | 0.981 | 0.839 | 0.823 | 0.773 |
| mT5-Base (Chinese) | Inflection | 0.901 | 0.933 | 0.945 | 0.941 | 0.943 |
| | Lexeme | 0.846 | 0.919 | 0.863 | 0.757 | 0.727 |

(b) Selectivity (Linear Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish Chinese (Chinese) | Inflection | 0.223 | 0.292 | 0.346 | 0.356 | 0.391 |
| | Lexeme | -0.000 | -0.001 | -0.002 | -0.001 | -0.003 |
| Qwen2.5-1.5B (Chinese) | Inflection | 0.122 | 0.345 | 0.429 | 0.412 | 0.441 |
| | Lexeme | -0.000 | -0.000 | -0.002 | -0.002 | -0.003 |
| Qwen2.5-1.5B-Instruct (Chinese) | Inflection | 0.122 | 0.345 | 0.430 | 0.412 | 0.437 |
| | Lexeme | -0.000 | -0.001 | -0.003 | -0.001 | -0.002 |
| Qwen2.5-7B (Chinese) | Inflection | 0.047 | 0.250 | 0.387 | 0.386 | 0.408 |
| | Lexeme | -0.000 | 0.000 | -0.001 | 0.000 | -0.000 |
| Qwen2.5-7B-Instruct (Chinese) | Inflection | 0.048 | 0.250 | 0.392 | 0.387 | 0.411 |
| | Lexeme | -0.000 | -0.000 | -0.001 | 0.000 | -0.001 |
| mT5-Base (Chinese) | Inflection | 0.123 | 0.186 | 0.274 | 0.321 | 0.348 |
| | Lexeme | 0.001 | 0.000 | -0.001 | 0.000 | -0.003 |

(c) Accuracy (MLP Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish Chinese (Chinese) | Inflection | 0.913 | 0.930 | 0.946 | 0.944 | 0.939 |
| | Lexeme | 0.922 | 0.874 | 0.797 | 0.652 | 0.543 |
| Qwen2.5-1.5B (Chinese) | Inflection | 0.896 | 0.947 | 0.943 | 0.950 | 0.942 |
| | Lexeme | 0.882 | 0.869 | 0.738 | 0.695 | 0.449 |
| Qwen2.5-1.5B-Instruct (Chinese) | Inflection | 0.896 | 0.941 | 0.942 | 0.950 | 0.943 |
| | Lexeme | 0.883 | 0.864 | 0.719 | 0.691 | 0.383 |
| Qwen2.5-7B (Chinese) | Inflection | 0.899 | 0.952 | 0.947 | 0.955 | 0.946 |
| | Lexeme | 0.881 | 0.951 | 0.795 | 0.749 | 0.471 |
| Qwen2.5-7B-Instruct (Chinese) | Inflection | 0.900 | 0.952 | 0.948 | 0.955 | 0.943 |
| | Lexeme | 0.881 | 0.950 | 0.791 | 0.750 | 0.475 |
| mT5-Base (Chinese) | Inflection | 0.907 | 0.938 | 0.947 | 0.942 | 0.948 |
| | Lexeme | 0.841 | 0.796 | 0.658 | 0.587 | 0.661 |

(d) Selectivity (MLP Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish Chinese (Chinese) | Inflection | 0.072 | 0.153 | 0.198 | 0.252 | 0.315 |
| | Lexeme | -0.001 | -0.039 | -0.049 | -0.064 | -0.047 |
| Qwen2.5-1.5B (Chinese) | Inflection | 0.058 | 0.168 | 0.280 | 0.305 | 0.411 |
| | Lexeme | 0.000 | -0.001 | -0.006 | -0.001 | 0.002 |
| Qwen2.5-1.5B-Instruct (Chinese) | Inflection | 0.059 | 0.164 | 0.290 | 0.301 | 0.411 |
| | Lexeme | 0.000 | 0.002 | -0.012 | -0.000 | -0.002 |
| Qwen2.5-7B (Chinese) | Inflection | 0.061 | 0.136 | 0.287 | 0.312 | 0.388 |
| | Lexeme | -0.001 | -0.001 | -0.007 | -0.008 | 0.000 |
| Qwen2.5-7B-Instruct (Chinese) | Inflection | 0.055 | 0.135 | 0.298 | 0.314 | 0.379 |
| | Lexeme | -0.001 | -0.003 | -0.006 | -0.004 | -0.000 |
| mT5-Base (Chinese) | Inflection | 0.055 | 0.070 | 0.179 | 0.238 | 0.238 |
| | Lexeme | 0.003 | -0.008 | -0.013 | -0.006 | -0.011 |

Table 7: Probing Results for French

(a) Accuracy (Linear Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish French (French) | Inflection | 0.924 | 0.959 | 0.976 | 0.970 | 0.963 |
| | Lexeme | 0.888 | 0.813 | 0.714 | 0.665 | 0.619 |
| Qwen2.5-1.5B (French) | Inflection | 0.792 | 0.947 | 0.954 | 0.952 | 0.928 |
| | Lexeme | 0.541 | 0.850 | 0.696 | 0.696 | 0.602 |
| Qwen2.5-1.5B-Instruct (French) | Inflection | 0.792 | 0.945 | 0.951 | 0.949 | 0.925 |
| | Lexeme | 0.541 | 0.845 | 0.687 | 0.690 | 0.611 |
| Qwen2.5-7B (French) | Inflection | 0.793 | 0.966 | 0.965 | 0.964 | 0.945 |
| | Lexeme | 0.541 | 0.943 | 0.801 | 0.769 | 0.714 |
| Qwen2.5-7B-Instruct (French) | Inflection | 0.793 | 0.963 | 0.962 | 0.961 | 0.941 |
| | Lexeme | 0.541 | 0.942 | 0.790 | 0.760 | 0.706 |
| mT5-Base (French) | Inflection | 0.840 | 0.943 | 0.967 | 0.961 | 0.944 |
| | Lexeme | 0.656 | 0.773 | 0.674 | 0.596 | 0.567 |

(b) Selectivity (Linear Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish French (French) | Inflection | 0.403 | 0.497 | 0.581 | 0.585 | 0.613 |
| | Lexeme | 0.039 | 0.109 | 0.132 | 0.131 | 0.122 |
| Qwen2.5-1.5B (French) | Inflection | 0.244 | 0.463 | 0.576 | 0.559 | 0.561 |
| | Lexeme | 0.002 | 0.137 | 0.126 | 0.132 | 0.088 |
| Qwen2.5-1.5B-Instruct (French) | Inflection | 0.244 | 0.467 | 0.582 | 0.565 | 0.562 |
| | Lexeme | 0.002 | 0.137 | 0.126 | 0.132 | 0.092 |
| Qwen2.5-7B (French) | Inflection | 0.228 | 0.370 | 0.538 | 0.545 | 0.540 |
| | Lexeme | 0.003 | 0.116 | 0.145 | 0.139 | 0.112 |
| Qwen2.5-7B-Instruct (French) | Inflection | 0.228 | 0.375 | 0.545 | 0.552 | 0.544 |
| | Lexeme | 0.002 | 0.118 | 0.143 | 0.138 | 0.112 |
| mT5-Base (French) | Inflection | 0.248 | 0.386 | 0.495 | 0.548 | 0.530 |
| | Lexeme | -0.006 | 0.027 | 0.045 | 0.046 | 0.037 |

(c) Accuracy (MLP Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish French (French) | Inflection | 0.932 | 0.972 | 0.980 | 0.979 | 0.971 |
| | Lexeme | 0.947 | 0.949 | 0.916 | 0.829 | 0.696 |
| Qwen2.5-1.5B (French) | Inflection | 0.789 | 0.956 | 0.965 | 0.960 | 0.929 |
| | Lexeme | 0.536 | 0.910 | 0.845 | 0.788 | 0.360 |
| Qwen2.5-1.5B-Instruct (French) | Inflection | 0.791 | 0.954 | 0.962 | 0.959 | 0.929 |
| | Lexeme | 0.537 | 0.911 | 0.835 | 0.798 | 0.535 |
| Qwen2.5-7B (French) | Inflection | 0.791 | 0.971 | 0.971 | 0.968 | 0.943 |
| | Lexeme | 0.533 | 0.953 | 0.862 | 0.830 | 0.461 |
| Qwen2.5-7B-Instruct (French) | Inflection | 0.791 | 0.968 | 0.969 | 0.965 | 0.939 |
| | Lexeme | 0.534 | 0.949 | 0.851 | 0.823 | 0.467 |
| mT5-Base (French) | Inflection | 0.851 | 0.962 | 0.975 | 0.967 | 0.969 |
| | Lexeme | 0.654 | 0.785 | 0.698 | 0.633 | 0.665 |

(d) Selectivity (MLP Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish French (French) | Inflection | 0.131 | 0.289 | 0.372 | 0.443 | 0.486 |
| | Lexeme | -0.016 | 0.006 | 0.032 | 0.072 | 0.084 |
| Qwen2.5-1.5B (French) | Inflection | 0.236 | 0.260 | 0.402 | 0.405 | 0.548 |
| | Lexeme | 0.006 | 0.025 | 0.046 | 0.056 | 0.043 |
| Qwen2.5-1.5B-Instruct (French) | Inflection | 0.235 | 0.263 | 0.411 | 0.410 | 0.552 |
| | Lexeme | 0.002 | 0.030 | 0.048 | 0.055 | 0.064 |
| Qwen2.5-7B (French) | Inflection | 0.227 | 0.198 | 0.394 | 0.415 | 0.516 |
| | Lexeme | -0.001 | 0.025 | 0.065 | 0.070 | 0.027 |
| Qwen2.5-7B-Instruct (French) | Inflection | 0.229 | 0.205 | 0.407 | 0.426 | 0.514 |
| | Lexeme | 0.000 | 0.028 | 0.072 | 0.070 | 0.037 |
| mT5-Base (French) | Inflection | 0.127 | 0.190 | 0.321 | 0.404 | 0.388 |
| | Lexeme | -0.020 | -0.036 | -0.015 | -0.000 | -0.000 |

## Table 8: Probing Results for German

(a) Accuracy (Linear Probes)

| Model | Task | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| | | 0% | 25% | 50% | 75% | 100% |
| Goldfish German (German) | Inflection | 0.911 | 0.946 | 0.961 | 0.961 | 0.952 |
| | Lexeme | 0.886 | 0.831 | 0.707 | 0.627 | 0.569 |
| Qwen2.5-1.5B (German) | Inflection | 0.744 | 0.929 | 0.931 | 0.930 | 0.911 |
| | Lexeme | 0.479 | 0.865 | 0.707 | 0.690 | 0.569 |
| Qwen2.5-1.5B-Instruct (German) | Inflection | 0.745 | 0.928 | 0.929 | 0.929 | 0.912 |
| | Lexeme | 0.479 | 0.862 | 0.694 | 0.686 | 0.582 |
| Qwen2.5-7B (German) | Inflection | 0.744 | 0.949 | 0.946 | 0.950 | 0.935 |
| | Lexeme | 0.480 | 0.942 | 0.811 | 0.764 | 0.651 |
| Qwen2.5-7B-Instruct (German) | Inflection | 0.760 | 0.958 | 0.954 | 0.958 | 0.938 |
| | Lexeme | 0.480 | 0.943 | 0.801 | 0.757 | 0.646 |
| mT5-Base (German) | Inflection | 0.811 | 0.942 | 0.954 | 0.956 | 0.916 |
| | Lexeme | 0.650 | 0.796 | 0.656 | 0.574 | 0.543 |

(b) Selectivity (Linear Probes)

| Model | Task | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| | | 0% | 25% | 50% | 75% | 100% |
| Goldfish German (German) | Inflection | 0.418 | 0.503 | 0.593 | 0.622 | 0.645 |
| | Lexeme | -0.001 | 0.057 | 0.088 | 0.088 | 0.069 |
| Qwen2.5-1.5B (German) | Inflection | 0.289 | 0.446 | 0.577 | 0.582 | 0.599 |
| | Lexeme | 0.009 | 0.084 | 0.082 | 0.096 | 0.065 |
| Qwen2.5-1.5B-Instruct (German) | Inflection | 0.289 | 0.450 | 0.577 | 0.583 | 0.602 |
| | Lexeme | 0.009 | 0.082 | 0.082 | 0.098 | 0.069 |
| Qwen2.5-7B (German) | Inflection | 0.286 | 0.357 | 0.546 | 0.574 | 0.597 |
| | Lexeme | 0.009 | 0.065 | 0.104 | 0.113 | 0.085 |
| Qwen2.5-7B-Instruct (German) | Inflection | 0.228 | 0.210 | 0.468 | 0.509 | 0.530 |
| | Lexeme | 0.009 | 0.067 | 0.105 | 0.111 | 0.084 |
| mT5-Base (German) | Inflection | 0.251 | 0.371 | 0.494 | 0.560 | 0.529 |
| | Lexeme | 0.000 | 0.012 | 0.023 | 0.043 | 0.033 |

(c) Accuracy (MLP Probes)

| Model | Task | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| | | 0% | 25% | 50% | 75% | 100% |
| Goldfish German (German) | Inflection | 0.923 | 0.955 | 0.969 | 0.969 | 0.960 |
| | Lexeme | 0.902 | 0.876 | 0.794 | 0.657 | 0.511 |
| Qwen2.5-1.5B (German) | Inflection | 0.741 | 0.943 | 0.944 | 0.942 | 0.910 |
| | Lexeme | 0.473 | 0.869 | 0.763 | 0.682 | 0.292 |
| Qwen2.5-1.5B-Instruct (German) | Inflection | 0.741 | 0.943 | 0.943 | 0.941 | 0.910 |
| | Lexeme | 0.474 | 0.870 | 0.753 | 0.688 | 0.300 |
| Qwen2.5-7B (German) | Inflection | 0.740 | 0.956 | 0.954 | 0.956 | 0.935 |
| | Lexeme | 0.471 | 0.943 | 0.820 | 0.746 | 0.383 |
| Qwen2.5-7B-Instruct (German) | Inflection | 0.758 | 0.962 | 0.959 | 0.961 | 0.935 |
| | Lexeme | 0.471 | 0.943 | 0.810 | 0.749 | 0.397 |
| mT5-Base (German) | Inflection | 0.820 | 0.956 | 0.954 | 0.952 | 0.939 |
| | Lexeme | 0.641 | 0.710 | 0.563 | 0.486 | 0.530 |

(d) Selectivity (MLP Probes)

| Model | Task | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| | | 0% | 25% | 50% | 75% | 100% |
| Goldfish German (German) | Inflection | 0.229 | 0.350 | 0.457 | 0.523 | 0.585 |
| | Lexeme | -0.021 | 0.017 | 0.051 | 0.084 | 0.082 |
| Qwen2.5-1.5B (German) | Inflection | 0.292 | 0.292 | 0.457 | 0.492 | 0.604 |
| | Lexeme | 0.008 | 0.022 | 0.054 | 0.060 | 0.035 |
| Qwen2.5-1.5B-Instruct (German) | Inflection | 0.291 | 0.295 | 0.472 | 0.488 | 0.600 |
| | Lexeme | 0.009 | 0.023 | 0.056 | 0.063 | 0.032 |
| Qwen2.5-7B (German) | Inflection | 0.288 | 0.214 | 0.455 | 0.493 | 0.596 |
| | Lexeme | 0.007 | 0.011 | 0.066 | 0.070 | 0.046 |
| Qwen2.5-7B-Instruct (German) | Inflection | 0.227 | 0.111 | 0.382 | 0.424 | 0.532 |
| | Lexeme | 0.008 | 0.015 | 0.073 | 0.077 | 0.049 |
| mT5-Base (German) | Inflection | 0.160 | 0.253 | 0.405 | 0.475 | 0.462 |
| | Lexeme | -0.012 | -0.037 | 0.000 | 0.023 | 0.012 |

## Table 9: Probing Results for Russian

(a) Accuracy (Linear Probes)

| Model | Task | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| | | 0% | 25% | 50% | 75% | 100% |
| Goldfish Russian (Russian) | Inflection | 0.932 | 0.952 | 0.975 | 0.968 | 0.950 |
| | Lexeme | 0.896 | 0.854 | 0.758 | 0.710 | 0.631 |
| Qwen2.5-1.5B (Russian) | Inflection | 0.850 | 0.966 | 0.966 | 0.962 | 0.933 |
| | Lexeme | 0.315 | 0.896 | 0.739 | 0.720 | 0.598 |
| Qwen2.5-1.5B-Instruct (Russian) | Inflection | 0.850 | 0.965 | 0.964 | 0.960 | 0.932 |
| | Lexeme | 0.315 | 0.893 | 0.725 | 0.714 | 0.600 |
| Qwen2.5-7B (Russian) | Inflection | 0.850 | 0.977 | 0.976 | 0.974 | 0.954 |
| | Lexeme | 0.315 | 0.960 | 0.834 | 0.798 | 0.696 |
| Qwen2.5-7B-Instruct (Russian) | Inflection | 0.858 | 0.977 | 0.974 | 0.980 | 0.953 |
| | Lexeme | 0.315 | 0.959 | 0.821 | 0.785 | 0.680 |
| mT5-Base (Russian) | Inflection | 0.882 | 0.944 | 0.974 | 0.971 | 0.952 |
| | Lexeme | 0.480 | 0.766 | 0.666 | 0.570 | 0.515 |

(b) Selectivity (Linear Probes)

| Model | Task | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| | | 0% | 25% | 50% | 75% | 100% |
| Goldfish Russian (Russian) | Inflection | 0.505 | 0.568 | 0.663 | 0.675 | 0.688 |
| | Lexeme | 0.024 | 0.115 | 0.157 | 0.175 | 0.170 |
| Qwen2.5-1.5B (Russian) | Inflection | 0.518 | 0.576 | 0.679 | 0.661 | 0.665 |
| | Lexeme | 0.012 | 0.190 | 0.187 | 0.202 | 0.153 |
| Qwen2.5-1.5B-Instruct (Russian) | Inflection | 0.518 | 0.582 | 0.685 | 0.664 | 0.666 |
| | Lexeme | 0.012 | 0.191 | 0.190 | 0.205 | 0.152 |
| Qwen2.5-7B (Russian) | Inflection | 0.517 | 0.504 | 0.670 | 0.671 | 0.658 |
| | Lexeme | 0.011 | 0.165 | 0.218 | 0.222 | 0.183 |
| Qwen2.5-7B-Instruct (Russian) | Inflection | 0.431 | 0.332 | 0.581 | 0.594 | 0.593 |
| | Lexeme | 0.011 | 0.167 | 0.221 | 0.222 | 0.181 |
| mT5-Base (Russian) | Inflection | 0.388 | 0.418 | 0.548 | 0.595 | 0.605 |
| | Lexeme | 0.004 | 0.088 | 0.092 | 0.099 | 0.092 |

(c) Accuracy (MLP Probes)

| Model | Task | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| | | 0% | 25% | 50% | 75% | 100% |
| Goldfish Russian (Russian) | Inflection | 0.944 | 0.966 | 0.982 | 0.977 | 0.959 |
| | Lexeme | 0.896 | 0.878 | 0.814 | 0.732 | 0.582 |
| Qwen2.5-1.5B (Russian) | Inflection | 0.848 | 0.972 | 0.971 | 0.969 | 0.924 |
| | Lexeme | 0.302 | 0.874 | 0.768 | 0.690 | 0.281 |
| Qwen2.5-1.5B-Instruct (Russian) | Inflection | 0.848 | 0.971 | 0.970 | 0.967 | 0.927 |
| | Lexeme | 0.302 | 0.870 | 0.760 | 0.692 | 0.282 |
| Qwen2.5-7B (Russian) | Inflection | 0.850 | 0.981 | 0.981 | 0.977 | 0.940 |
| | Lexeme | 0.312 | 0.960 | 0.838 | 0.770 | 0.361 |
| Qwen2.5-7B-Instruct (Russian) | Inflection | 0.857 | 0.976 | 0.976 | 0.978 | 0.935 |
| | Lexeme | 0.308 | 0.958 | 0.828 | 0.765 | 0.371 |
| mT5-Base (Russian) | Inflection | 0.883 | 0.956 | 0.978 | 0.974 | 0.971 |
| | Lexeme | 0.448 | 0.571 | 0.470 | 0.397 | 0.437 |

(d) Selectivity (MLP Probes)

| Model | Task | Relative Depth (%) | | | | |
|---|---|---|---|---|---|---|
| | | 0% | 25% | 50% | 75% | 100% |
| Goldfish Russian (Russian) | Inflection | 0.322 | 0.466 | 0.557 | 0.604 | 0.642 |
| | Lexeme | -0.007 | 0.049 | 0.120 | 0.203 | 0.201 |
| Qwen2.5-1.5B (Russian) | Inflection | 0.524 | 0.451 | 0.575 | 0.594 | 0.662 |
| | Lexeme | 0.004 | 0.080 | 0.122 | 0.162 | 0.108 |
| Qwen2.5-1.5B-Instruct (Russian) | Inflection | 0.523 | 0.457 | 0.589 | 0.594 | 0.660 |
| | Lexeme | 0.004 | 0.085 | 0.133 | 0.166 | 0.101 |
| Qwen2.5-7B (Russian) | Inflection | 0.526 | 0.376 | 0.591 | 0.604 | 0.652 |
| | Lexeme | 0.011 | 0.062 | 0.175 | 0.183 | 0.123 |
| Qwen2.5-7B-Instruct (Russian) | Inflection | 0.432 | 0.211 | 0.493 | 0.512 | 0.583 |
| | Lexeme | 0.008 | 0.065 | 0.184 | 0.193 | 0.125 |
| mT5-Base (Russian) | Inflection | 0.351 | 0.335 | 0.497 | 0.547 | 0.557 |
| | Lexeme | -0.016 | -0.016 | 0.023 | 0.035 | 0.042 |

Table 10: Probing Results for Turkish

(a) Accuracy (Linear Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish Turkish (Turkish) | Inflection | 0.907 | 0.930 | 0.925 | 0.913 | 0.903 |
| | Lexeme | 0.978 | 0.973 | 0.968 | 0.921 | 0.614 |
| Qwen2.5-1.5B (Turkish) | Inflection | 0.719 | 0.869 | 0.847 | 0.849 | 0.831 |
| | Lexeme | 0.530 | 0.959 | 0.868 | 0.815 | 0.796 |
| Qwen2.5-1.5B-Instruct (Turkish) | Inflection | 0.719 | 0.869 | 0.838 | 0.846 | 0.827 |
| | Lexeme | 0.530 | 0.961 | 0.852 | 0.804 | 0.786 |
| Qwen2.5-7B (Turkish) | Inflection | 0.718 | 0.917 | 0.889 | 0.879 | 0.839 |
| | Lexeme | 0.531 | 0.974 | 0.878 | 0.839 | 0.777 |
| Qwen2.5-7B-Instruct (Turkish) | Inflection | 0.718 | 0.911 | 0.875 | 0.874 | 0.836 |
| | Lexeme | 0.531 | 0.975 | 0.854 | 0.803 | 0.731 |
| mT5-Base (Turkish) | Inflection | 0.913 | 0.972 | 0.931 | 0.908 | 0.884 |
| | Lexeme | 0.792 | 0.952 | 0.922 | 0.819 | 0.785 |

(b) Selectivity (Linear Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish Turkish (Turkish) | Inflection | 0.345 | 0.407 | 0.491 | 0.492 | 0.500 |
| | Lexeme | -0.002 | 0.008 | 0.074 | 0.089 | 0.144 |
| Qwen2.5-1.5B (Turkish) | Inflection | 0.277 | 0.323 | 0.452 | 0.455 | 0.411 |
| | Lexeme | 0.013 | 0.001 | 0.000 | -0.014 | 0.008 |
| Qwen2.5-1.5B-Instruct (Turkish) | Inflection | 0.277 | 0.319 | 0.447 | 0.452 | 0.416 |
| | Lexeme | 0.013 | 0.003 | 0.004 | -0.009 | 0.009 |
| Qwen2.5-7B (Turkish) | Inflection | 0.275 | 0.293 | 0.472 | 0.462 | 0.417 |
| | Lexeme | 0.014 | -0.002 | 0.003 | -0.010 | -0.021 |
| Qwen2.5-7B-Instruct (Turkish) | Inflection | 0.276 | 0.291 | 0.463 | 0.497 | 0.445 |
| | Lexeme | 0.014 | 0.003 | -0.003 | -0.017 | -0.037 |
| mT5-Base (Turkish) | Inflection | 0.071 | 0.165 | 0.261 | 0.331 | 0.366 |
| | Lexeme | 0.008 | 0.013 | 0.042 | 0.035 | 0.057 |

(c) Accuracy (MLP Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish Turkish (Turkish) | Inflection | 0.911 | 0.918 | 0.916 | 0.899 | 0.887 |
| | Lexeme | 0.601 | 0.536 | 0.459 | 0.418 | 0.360 |
| Qwen2.5-1.5B (Turkish) | Inflection | 0.713 | 0.854 | 0.823 | 0.829 | 0.760 |
| | Lexeme | 0.459 | 0.500 | 0.369 | 0.325 | 0.232 |
| Qwen2.5-1.5B-Instruct (Turkish) | Inflection | 0.712 | 0.857 | 0.831 | 0.817 | 0.760 |
| | Lexeme | 0.462 | 0.491 | 0.367 | 0.333 | 0.233 |
| Qwen2.5-7B (Turkish) | Inflection | 0.713 | 0.923 | 0.902 | 0.900 | 0.828 |
| | Lexeme | 0.519 | 0.805 | 0.639 | 0.525 | 0.441 |
| Qwen2.5-7B-Instruct (Turkish) | Inflection | 0.717 | 0.914 | 0.900 | 0.895 | 0.820 |
| | Lexeme | 0.521 | 0.797 | 0.615 | 0.523 | 0.383 |
| mT5-Base (Turkish) | Inflection | 0.884 | 0.915 | 0.875 | 0.836 | 0.839 |
| | Lexeme | 0.503 | 0.395 | 0.300 | 0.257 | 0.312 |

(d) Selectivity (MLP Probes)

| Model | Task | Relative Depth (%) 0% | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Goldfish Turkish (Turkish) | Inflection | 0.333 | 0.449 | 0.499 | 0.488 | 0.496 |
| | Lexeme | 0.175 | 0.185 | 0.142 | 0.138 | 0.125 |
| Qwen2.5-1.5B (Turkish) | Inflection | 0.299 | 0.355 | 0.438 | 0.500 | 0.432 |
| | Lexeme | -0.001 | 0.045 | 0.093 | 0.049 | 0.045 |
| Qwen2.5-1.5B-Instruct (Turkish) | Inflection | 0.303 | 0.350 | 0.447 | 0.447 | 0.432 |
| | Lexeme | 0.001 | 0.042 | 0.104 | 0.065 | -0.078 |
| Qwen2.5-7B (Turkish) | Inflection | 0.297 | 0.299 | 0.422 | 0.473 | 0.442 |
| | Lexeme | 0.007 | 0.084 | 0.128 | 0.105 | 0.065 |
| Qwen2.5-7B-Instruct (Turkish) | Inflection | 0.303 | 0.297 | 0.437 | 0.492 | 0.432 |
| | Lexeme | 0.012 | 0.087 | 0.147 | 0.103 | 0.061 |
| mT5-Base (Turkish) | Inflection | 0.120 | 0.301 | 0.343 | 0.355 | 0.368 |
| | Lexeme | 0.047 | 0.094 | 0.109 | 0.097 | 0.106 |

| Language | Total Words | Unique Lemmas | Unique Forms | Inflection Types | Sentences | Avg. Length |
|---|---|---|---|---|---|---|
| English | 54,816 | 7,848 | 11,720 | 8 | 8,415 | 6.5 |
| Chinese | 44,166 | 11,184 | 11,237 | 4 | 7,892 | 5.8 |
| German | 84,710 | 24,140 | 31,890 | 9 | 9,234 | 7.3 |
| French | 115,847 | 13,804 | 24,485 | 6 | 8,765 | 6.6 |
| Russian | 193,320 | 20,943 | 59,830 | 8 | 10,234 | 7.1 |
| Turkish | 20,881 | 3,776 | 11,680 | 7 | 6,789 | 6.4 |

Table 11: Dataset statistics across all six languages. Russian has the largest dataset and the highest number of unique forms, reflecting its rich inflectional morphology. Turkish has the fewest total words and lemmas, while Chinese has the fewest inflection types.

| Category | Count | % |
|---|---|---|
| Noun | 27111 | 49.5 |
| Verb | 17093 | 31.2 |
| Adjective | 10612 | 19.4 |

(a) Word categories

| Inflection | Count | % |
|---|---|---|
| Singular | 19830 | 36.2 |
| Base | 10076 | 18.4 |
| Positive | 9926 | 18.1 |
| Plural | 7281 | 13.3 |
| Past | 5604 | 10.2 |
| 3rd Person | 1413 | 2.6 |
| Comparative | 403 | 0.7 |
| Superlative | 283 | 0.5 |

(b) Inflection categories

| Metric | Value |
|---|---|
| Avg. Words | 6.5 |
| Median Words | 5 |
| Min. Words | 1 |
| Max. Words | 40 |

(c) Sentence length stats

Table 12: Distribution statistics for the English dataset. Table (a) shows syntactic categories, (b) details inflection types, and (c) provides sentence length heuristics.

| Model | Tokenizer Type |
|---|---|
| BERT Base/Large | WordPiece |
| DeBERTa V3 Large | SentencePiece |
| GPT-2 variants | BPE |
| Pythia variants | BPE |
| OLMo 2 variants | BPE (tiktoken) |
| Gemma 2 variants | SentencePiece |
| Qwen 2.5 variants | Byte-level BPE |
| Llama 3.1 variants | BPE (tiktoken) |

Table 13: Tokenization strategies used by different model families. BPE means byte-pair encoding.

| Model | Avg. tokens per word | Med. tokens per word | Max tokens per word | Percent multitoken |
|---|---|---|---|---|
| BERT variants | 1.11 | 1.0 | 6.0 | 6.95 |
| DeBERTa-v3-large | 1.03 | 1.0 | 4.0 | 2.2 |
| GPT-2 variants | 1.52 | 1.0 | 5.0 | 42.25 |
| Pythia-6.9B variants | 1.48 | 1.0 | 5.0 | 39.1 |
| OLMo2-7B variants | 1.43 | 1.0 | 4.0 | 35.9 |
| Gemma2-2B variants | 1.19 | 1.0 | 4.0 | 16.55 |
| Qwen2.5-1.5B variants | 1.43 | 1.0 | 4.0 | 35.9 |
| Llama-3.1-8B variants | 1.43 | 1.0 | 4.0 | 35.85 |

Table 14: Tokenization statistics across different models (English only). Most models have an average of 1.0-1.5 tokens per word and a median of 1, indicating that most words are tokenized as a single unit. However, there is variation in the proportion of words split into multiple tokens. Decoder-only models (*e.g.*, GPT-2, Pythia, Qwen2, LLaMA) split 35-42% of words, while BERT and DeBERTa variants split fewer words (2-7%). Maximum tokens per word range from 4 to 6 across all models.

| Model | HuggingFace ID |
|---|---|
| BERT-Base | bert-base-uncased |
| BERT-Large | bert-large-uncased |
| DeBERTa-v3-Large | microsoft/deberta-v3-large |
| mT5-base | google/mt5-base |
| GPT-2-Small | openai-community/gpt2 |
| GPT-2-Large | openai-community/gpt2-large |
| GPT-2-XL | openai-community/gpt2-xl |
| Pythia-6.9B | EleutherAI/pythia-6.9b |
| Pythia-6.9B-Tulu | allenai/open-instruct-pythia-6.9b-tulu |
| OLMo-2-1124-7B | allenai/OLMo-2-1124-7B |
| OLMo-2-1124-7B-Instruct | allenai/OLMo-2-1124-7B-Instruct |
| Gemma-2-2B | google/gemma-2-2b |
| Gemma-2-2B-Instruct | google/gemma-2-2b-it |
| Qwen2.5-1.5B | Qwen/Qwen2.5-1.5B |
| Qwen2.5-1.5B-Instruct | Qwen/Qwen2.5-1.5B-Instruct |
| Qwen2.5-7B | Qwen/Qwen2.5-7B |
| Qwen2.5-7B-Instruct | Qwen/Qwen2.5-7B-Instruct |
| Llama-3.1-8B | meta-llama/Llama-3.1-8B |
| Llama-3.1-8B-Instruct | meta-llama/Llama-3.1-8B-Instruct |
| Goldfish English | goldfish-models/goldfish_eng_latn_1000mb |
| Goldfish Chinese | goldfish-models/goldfish_zho_hans_1000mb |
| Goldfish German | goldfish-models/goldfish_deu_latn_1000mb |
| Goldfish French | goldfish-models/goldfish_fra_latn_1000mb |
| Goldfish Russian | goldfish-models/goldfish_rus_cyrl_1000mb |
| Goldfish Turkish | goldfish-models/goldfish_tur_latn_1000mb |

Table 15: Canonical HuggingFace model IDs used to load models in our study.

| Model | $d_{\text{model}}$ | $\text{ID}_{50}$ | | | $\text{ID}_{70}$ | | | $\text{ID}_{90}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | First | Mid | Final | First | Mid | Final | First | Mid | Final |
| BERT-Base | 768 | 123 | 100 | 88 | 244 | 212 | 192 | 461 | 451 | 446 |
| BERT-Large | 1024 | 138 | 105 | 85 | 286 | 226 | 208 | 567 | 527 | 554 |
| DeBERTa-v3-Large | 1024 | 196 | 133 | 29 | 377 | 299 | 113 | 688 | 635 | 423 |
| GPT-2-Small | 768 | 37 | 1 | 1 | 152 | 1 | 1 | 402 | 1 | 3 |
| GPT-2-Large | 1280 | 24 | 1 | 95 | 172 | 1 | 284 | 583 | 1 | 726 |
| GPT-2-XL | 1600 | 113 | 1 | 118 | 340 | 1 | 356 | 838 | 1 | 914 |
| Pythia-6.9B | 4096 | 391 | 1 | 96 | 865 | 1 | 517 | 1952 | 1 | 1925 |
| Pythia-6.9B-Tulu | 4096 | 390 | 1 | 244 | 862 | 1 | 832 | 1949 | 1 | 2292 |
| OLMo-2-7B | 4096 | 404 | 310 | 41 | 833 | 896 | 299 | 1772 | 2279 | 1550 |
| OLMo-2-7B-Instruct | 4096 | 404 | 358 | 111 | 833 | 974 | 567 | 1772 | 2361 | 1964 |
| Gemma-2-2B | 2304 | 216 | 8 | 11 | 505 | 130 | 70 | 1129 | 794 | 611 |
| Gemma-2-2B-Instruct | 2304 | 222 | 22 | 8 | 520 | 198 | 57 | 1153 | 899 | 572 |
| Qwen-2.5-1.5B | 1536 | 184 | 1 | 9 | 399 | 1 | 50 | 835 | 1 | 452 |
| Qwen-2.5-1.5B-Instruct | 1536 | 184 | 1 | 11 | 394 | 1 | 70 | 820 | 1 | 533 |
| Llama-3.1-8B | 4096 | 373 | 240 | 35 | 789 | 727 | 187 | 1722 | 2051 | 1119 |
| Llama-3.1-8B-Instruct | 4096 | 372 | 215 | 31 | 788 | 664 | 181 | 1722 | 1957 | 1093 |

Table 16: Number of principal-component axes required to reach 50% ($\text{ID}_{50}$), 70% ($\text{ID}_{70}$) and 90% ($\text{ID}_{90}$) explained variance in the first, middle and last layers of each model.

| **Model** | 3rd person (n=249) | Base (n=1,833) | Comparative (n=76) | Past (n=1,003) | Plural (n=1,247) | Positive (n=1,785) | Singular (n=3,587) | Superlative (n=52) |
|---|---|---|---|---|---|---|---|---|
| BERT-Base | 0.960 | 0.965 | 0.817 | 0.967 | 0.983 | 0.946 | 0.971 | 0.759 |
| BERT-Large | 0.956 | 0.964 | 0.861 | 0.968 | 0.982 | 0.950 | 0.971 | 0.768 |
| DeBERTa-v3-Large | 0.938 | 0.974 | 0.831 | 0.961 | 0.986 | 0.954 | 0.977 | 0.706 |
| GPT-2-Small | 0.828 | 0.958 | 0.840 | 0.956 | 0.974 | 0.941 | 0.964 | 0.754 |
| GPT-2-Large | 0.812 | 0.958 | 0.826 | 0.951 | 0.975 | 0.936 | 0.967 | 0.792 |
| GPT-2-XL | 0.817 | 0.959 | 0.813 | 0.948 | 0.977 | 0.940 | 0.968 | 0.788 |
| Pythia-6.9B | 0.886 | 0.972 | 0.904 | 0.964 | 0.989 | 0.957 | 0.977 | 0.907 |
| Pythia-6.9B-Tulu | 0.899 | 0.973 | 0.909 | 0.967 | 0.989 | 0.956 | 0.976 | 0.910 |
| OLMo-2-1124-7B | 0.938 | 0.968 | 0.902 | 0.972 | 0.981 | 0.923 | 0.966 | 0.888 |
| OLMo-2-1124-7B-Instruct | 0.927 | 0.967 | 0.896 | 0.971 | 0.981 | 0.923 | 0.965 | 0.872 |
| Gemma-2-2B | 0.901 | 0.968 | 0.797 | 0.969 | 0.986 | 0.947 | 0.974 | 0.833 |
| Gemma-2-2B-Instruct | 0.913 | 0.966 | 0.863 | 0.973 | 0.988 | 0.938 | 0.972 | 0.872 |
| Qwen2.5-1.5B | 0.856 | 0.950 | 0.802 | 0.942 | 0.972 | 0.919 | 0.957 | 0.688 |
| Qwen2.5-1.5B-Instruct | 0.774 | 0.954 | 0.647 | 0.945 | 0.972 | 0.921 | 0.965 | 0.630 |

Table 17: Breakdown of inflection classification accuracy by morphological feature for each model using linear regression classifiers (English). Inflections are grouped by their morphological features (*e.g.*, Past, Plural, Comparative). For each group, the reported accuracy is the average of accuracies from classifiers trained at each model layer. All accuracy values are on a 0–1 scale. Comparative and superlative forms consistently show the lowest accuracy across all models, reflecting the challenges of these less frequent morphological categories.

| Model | 3rd person (n=249) | Base (n=1,833) | Comparative (n=76) | Past (n=1,003) | Plural (n=1,247) | Positive (n=1,785) | Singular (n=3,587) | Superlative (n=52) |
|---|---|---|---|---|---|---|---|---|
| BERT-Base | 0.973 | 0.969 | 0.910 | 0.972 | 0.989 | 0.959 | 0.974 | 0.939 |
| BERT-Large | 0.967 | 0.970 | 0.910 | 0.973 | 0.988 | 0.961 | 0.975 | 0.931 |
| DeBERTa-v3-Large | 0.954 | 0.976 | 0.925 | 0.966 | 0.989 | 0.962 | 0.979 | 0.867 |
| GPT-2-Small | 0.921 | 0.963 | 0.928 | 0.952 | 0.972 | 0.930 | 0.963 | 0.870 |
| GPT-2-Large | 0.857 | 0.962 | 0.872 | 0.955 | 0.976 | 0.942 | 0.967 | 0.854 |
| GPT-2-XL | 0.921 | 0.963 | 0.928 | 0.952 | 0.972 | 0.930 | 0.963 | 0.870 |
| Pythia-6.9B | 0.932 | 0.972 | 0.921 | 0.961 | 0.982 | 0.949 | 0.971 | 0.886 |
| Pythia-6.9B-Tulu | 0.948 | 0.974 | 0.932 | 0.964 | 0.983 | 0.949 | 0.971 | 0.897 |
| OLMo-2-1124-7B | 0.957 | 0.968 | 0.926 | 0.966 | 0.989 | 0.949 | 0.973 | 0.905 |
| OLMo-2-1124-7B-Instruct | 0.939 | 0.967 | 0.903 | 0.967 | 0.988 | 0.949 | 0.973 | 0.873 |
| Gemma-2-2B | 0.913 | 0.967 | 0.863 | 0.968 | 0.990 | 0.950 | 0.976 | 0.907 |
| Gemma-2-2B-Instruct | 0.930 | 0.970 | 0.878 | 0.975 | 0.989 | 0.946 | 0.974 | 0.906 |
| Qwen2.5-1.5B | 0.882 | 0.948 | 0.822 | 0.943 | 0.974 | 0.927 | 0.957 | 0.736 |
| Qwen2.5-1.5B-Instruct | 0.808 | 0.953 | 0.697 | 0.947 | 0.974 | 0.930 | 0.965 | 0.682 |

Table 18: Breakdown of inflection classification accuracy by morphological feature for each model using Multi-Layer Perceptron (MLP) classifiers (English). Inflections are grouped by their morphological features (*e.g.*, Past, Plural, Comparative). For each group, the reported accuracy is the average of accuracies from classifiers trained at each model layer. All accuracy values are on a 0–1 scale. MLP classifiers provide modest improvements over linear regression, particularly for comparative and superlative forms, though the relative ordering across morphological features remains consistent.

| Model | Noun (n=1,739) | Verb (n=641) | Adjective (n=641) | Adverb (n=23) | Pronoun (n=1) | Preposition (n=1) | Conjunction (n=1) | Interjection (n=1) | Other (n=9) |
|---|---|---|---|---|---|---|---|---|---|
| BERT-Base | 0.636 | 0.737 | 0.609 | 0.805 | 0.292 | 0.000 | 0.585 | 0.000 | 0.902 |
| BERT-Large | 0.684 | 0.777 | 0.653 | 0.826 | 0.580 | 0.154 | 0.662 | 0.065 | 0.897 |
| DeBERTa-v3-Large | 0.592 | 0.737 | 0.585 | 0.723 | 0.440 | 0.077 | 0.438 | 0.081 | 0.866 |
| GPT-2-Small | 0.631 | 0.789 | 0.612 | 0.813 | 0.542 | 0.000 | 0.415 | 0.033 | 0.896 |
| GPT-2-Large | 0.691 | 0.810 | 0.688 | 0.847 | 0.853 | 0.174 | 0.267 | 0.115 | 0.912 |
| GPT-2-XL | 0.713 | 0.827 | 0.708 | 0.847 | 0.724 | 0.222 | 0.311 | 0.241 | 0.899 |
| Pythia-6.9B | 0.856 | 0.926 | 0.836 | 0.926 | 0.938 | 0.443 | 0.566 | 0.488 | 0.934 |
| Pythia-6.9B-Tulu | 0.864 | 0.930 | 0.843 | 0.930 | 0.923 | 0.514 | 0.651 | 0.476 | 0.936 |
| OLMo-2-1124-7B | 0.798 | 0.875 | 0.794 | 0.913 | 0.697 | 0.339 | 0.363 | 0.495 | 0.913 |
| OLMo-2-1124-7B-Instruct | 0.798 | 0.868 | 0.792 | 0.902 | 0.606 | 0.339 | 0.331 | 0.495 | 0.910 |
| Gemma-2-2B | 0.757 | 0.869 | 0.736 | 0.876 | 0.667 | 0.179 | 0.205 | 0.288 | 0.891 |
| Gemma-2-2B-Instruct | 0.749 | 0.844 | 0.742 | 0.872 | 0.620 | 0.137 | 0.152 | 0.247 | 0.912 |
| Qwen2.5-1.5B | 0.652 | 0.801 | 0.650 | 0.828 | 0.526 | 0.082 | 0.223 | 0.068 | 0.867 |
| Qwen2.5-1.5B-Instruct | 0.642 | 0.800 | 0.632 | 0.831 | 0.544 | 0.082 | 0.245 | 0.068 | 0.877 |
| Llama-3.1-8B | 0.776 | 0.882 | 0.771 | 0.887 | 0.831 | 0.286 | 0.396 | 0.321 | 0.911 |
| Llama-3.1-8B-Instruct | 0.796 | 0.892 | 0.788 | 0.896 | 0.908 | 0.300 | 0.443 | 0.357 | 0.917 |

Table 19: Breakdown of lemma classification accuracy by Part of Speech (POS) for each model using linear regression classifiers (English). Lemmas are grouped by their POS tags (*e.g.*, Noun, Verb, Adjective). For each group, the reported accuracy is the average of accuracies from classifiers trained at each model layer. All accuracy values are on a 0–1 scale. Performance varies significantly with frequency: frequent categories like nouns and verbs achieve higher accuracy, while infrequent categories like pronouns and prepositions show lower performance due to limited training examples.

| Model | Noun (n=1,739) | Verb (n=641) | Adjective (n=641) | Adverb (n=23) | Pronoun (n=1) | Preposition (n=1) | Conjunction (n=1) | Interjection (n=1) | Other (n=9) |
|---|---|---|---|---|---|---|---|---|---|
| BERT-Base | 0.775 | 0.831 | 0.748 | 0.873 | 0.458 | 0.125 | 0.756 | 0.267 | 0.898 |
| BERT-Large | 0.813 | 0.863 | 0.785 | 0.884 | 0.540 | 0.231 | 0.725 | 0.323 | 0.897 |
| DeBERTa-v3-Large | 0.689 | 0.803 | 0.682 | 0.802 | 0.700 | 0.115 | 0.662 | 0.242 | 0.861 |
| GPT-2-Small | 0.678 | 0.792 | 0.665 | 0.765 | 0.042 | 0.000 | 0.610 | 0.000 | 0.830 |
| GPT-2-Large | 0.754 | 0.837 | 0.755 | 0.827 | 0.347 | 0.188 | 0.596 | 0.385 | 0.871 |
| GPT-2-XL | 0.774 | 0.844 | 0.771 | 0.827 | 0.561 | 0.232 | 0.561 | 0.431 | 0.860 |
| Pythia-6.9B | 0.774 | 0.856 | 0.768 | 0.862 | 0.554 | 0.229 | 0.528 | 0.310 | 0.868 |
| Pythia-6.9B-Tulu | 0.818 | 0.880 | 0.803 | 0.887 | 0.554 | 0.343 | 0.613 | 0.381 | 0.889 |
| OLMo-2-1124-7B | 0.818 | 0.877 | 0.828 | 0.896 | 0.727 | 0.290 | 0.734 | 0.505 | 0.885 |
| OLMo-2-1124-7B-Instruct | 0.822 | 0.874 | 0.829 | 0.897 | 0.667 | 0.306 | 0.750 | 0.473 | 0.886 |
| Gemma-2-2B | 0.763 | 0.860 | 0.763 | 0.881 | 0.574 | 0.125 | 0.443 | 0.182 | 0.880 |
| Gemma-2-2B-Instruct | 0.777 | 0.846 | 0.785 | 0.882 | 0.580 | 0.137 | 0.400 | 0.299 | 0.875 |
| Qwen2.5-1.5B | 0.747 | 0.838 | 0.742 | 0.811 | 0.228 | 0.131 | 0.628 | 0.164 | 0.857 |
| Qwen2.5-1.5B-Instruct | 0.749 | 0.840 | 0.738 | 0.818 | 0.211 | 0.098 | 0.564 | 0.123 | 0.860 |
| Llama-3.1-8B | 0.798 | 0.879 | 0.807 | 0.886 | 0.800 | 0.214 | 0.679 | 0.393 | 0.882 |
| Llama-3.1-8B-Instruct | 0.824 | 0.893 | 0.826 | 0.895 | 0.831 | 0.257 | 0.689 | 0.429 | 0.887 |

Table 20: Breakdown of lemma classification accuracy by Part of Speech (POS) for each model using Multi-Layer Perceptron (MLP) classifiers (English). Lemmas are grouped by their POS tags (*e.g.*, Noun, Verb, Adjective). For each group, the reported accuracy is the average of accuracies from classifiers trained at each model layer. All accuracy values are on a 0–1 scale. MLP classifiers provide consistent improvements over linear regression across all POS categories, though the frequency-dependent performance patterns persist.

| Model | Linear Regression | | | | MLP | | | |
|---|---|---|---|---|---|---|---|---|
| | Positive (n=300) | Base (n=2,074) | Plural (n=3) | Singular (n=3,947) | Positive (n=300) | Base (n=2,074) | Plural (n=3) | Singular (n=3,947) |
| mT5-Base | 0.739 | 0.913 | 0.436 | 0.962 | 0.783 | 0.919 | 0.231 | 0.961 |
| Qwen2.5-1.5B | 0.785 | 0.929 | 0.034 | 0.969 | 0.801 | 0.924 | 0.092 | 0.967 |
| Qwen2.5-1.5B-Instruct | 0.779 | 0.925 | 0.034 | 0.964 | 0.803 | 0.923 | 0.057 | 0.967 |
| Qwen2.5-7B | 0.824 | 0.937 | 0.310 | 0.970 | 0.828 | 0.929 | 0.310 | 0.969 |
| Qwen2.5-7B-Instruct | 0.819 | 0.936 | 0.299 | 0.970 | 0.823 | 0.928 | 0.276 | 0.969 |
| Goldfish Chinese | 0.793 | 0.912 | 0.000 | 0.958 | 0.816 | 0.915 | 0.000 | 0.957 |

Table 21: Breakdown of inflection classification accuracy for each model by inflection type using Linear Regression and Multi-Layer Perceptron (MLP) classifiers (Chinese). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Noun (n=1,179) | Verb (n=564) | Adjective (n=108) | Adverb (n=22) | Preposition (n=20) | Other (n=50) |
|---|---|---|---|---|---|---|
| mT5-Base | 0.838 | 0.828 | 0.786 | 0.762 | 0.920 | 0.726 |
| Qwen2.5-1.5B | 0.810 | 0.797 | 0.746 | 0.715 | 0.872 | 0.699 |
| Qwen2.5-1.5B-Instruct | 0.813 | 0.799 | 0.748 | 0.713 | 0.873 | 0.700 |
| Qwen2.5-7B | 0.887 | 0.882 | 0.846 | 0.847 | 0.915 | 0.817 |
| Qwen2.5-7B-Instruct | 0.886 | 0.877 | 0.843 | 0.835 | 0.913 | 0.811 |
| Goldfish Chinese | 0.883 | 0.878 | 0.845 | 0.875 | 0.954 | 0.858 |

Table 22: Breakdown of lemma classification accuracy by Part of Speech (POS) for each model, using Linear Regression classifiers (Chinese). Lemmas are grouped by their POS tags (*e.g.*, Noun, Verb, Adjective). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Noun (n=1,179) | Verb (n=564) | Adjective (n=108) | Adverb (n=22) | Preposition (n=20) | Other (n=50) |
|---|---|---|---|---|---|---|
| mT5-Base | 0.698 | 0.712 | 0.564 | 0.571 | 0.884 | 0.569 |
| Qwen2.5-1.5B | 0.748 | 0.761 | 0.658 | 0.668 | 0.826 | 0.669 |
| Qwen2.5-1.5B-Instruct | 0.735 | 0.745 | 0.643 | 0.643 | 0.814 | 0.655 |
| Qwen2.5-7B | 0.815 | 0.826 | 0.749 | 0.745 | 0.848 | 0.750 |
| Qwen2.5-7B-Instruct | 0.815 | 0.822 | 0.747 | 0.734 | 0.845 | 0.744 |
| Goldfish Chinese | 0.766 | 0.771 | 0.647 | 0.621 | 0.912 | 0.682 |

Table 23: Breakdown of lemma classification accuracy by Part of Speech (POS) for each model, using Multi-Layer Perceptron (MLP) classifiers (Chinese). Lemmas are grouped by their POS tags (*e.g.*, Noun, Verb, Adjective). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Base (n=417) | 3rd person (n=517) | Positive (n=1,720) | Past (n=839) | Plural (n=1,076) | Superlative (n=52) | Singular (n=3,197) | Comparative (n=141) |
|---|---|---|---|---|---|---|---|---|
| mT5-Base | 0.908 | 0.941 | 0.940 | 0.960 | 0.882 | 0.572 | 0.962 | 0.636 |
| Qwen2.5-1.5B | 0.849 | 0.889 | 0.922 | 0.914 | 0.888 | 0.657 | 0.953 | 0.796 |
| Qwen2.5-1.5B-Instruct | 0.844 | 0.887 | 0.922 | 0.910 | 0.889 | 0.659 | 0.952 | 0.795 |
| Qwen2.5-7B | 0.892 | 0.922 | 0.939 | 0.947 | 0.909 | 0.826 | 0.962 | 0.878 |
| Qwen2.5-7B-Instruct | 0.915 | 0.934 | 0.945 | 0.962 | 0.924 | 0.866 | 0.968 | 0.909 |
| Goldfish German | 0.938 | 0.941 | 0.955 | 0.979 | 0.916 | 0.542 | 0.968 | 0.708 |

Table 24: Breakdown of inflection classification accuracy for each model by inflection type using Linear Regression classifiers (German). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Base (n=417) | 3rd person (n=517) | Positive (n=1,720) | Past (n=839) | Plural (n=1,076) | Superlative (n=52) | Singular (n=3,197) | Comparative (n=141) |
|---|---|---|---|---|---|---|---|---|
| mT5-Base | 0.921 | 0.945 | 0.948 | 0.959 | 0.884 | 0.723 | 0.967 | 0.770 |
| Qwen2.5-1.5B | 0.890 | 0.915 | 0.930 | 0.940 | 0.897 | 0.831 | 0.958 | 0.892 |
| Qwen2.5-1.5B-Instruct | 0.888 | 0.914 | 0.930 | 0.938 | 0.898 | 0.825 | 0.957 | 0.897 |
| Qwen2.5-7B | 0.912 | 0.932 | 0.944 | 0.956 | 0.913 | 0.868 | 0.964 | 0.924 |
| Qwen2.5-7B-Instruct | 0.925 | 0.941 | 0.950 | 0.966 | 0.928 | 0.901 | 0.970 | 0.936 |
| Goldfish German | 0.947 | 0.957 | 0.964 | 0.978 | 0.923 | 0.817 | 0.970 | 0.896 |

Table 25: Breakdown of inflection classification accuracy for each model by inflection type using Multi-Layer Perceptron (MLP) classifiers (German). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Linear Regression | | | | MLP | | | |
|---|---|---|---|---|---|---|---|---|
| | Noun (n=1,262) | Verb (n=395) | Adjective (n=406) | Other (n=12) | Noun (n=1,262) | Verb (n=395) | Adjective (n=406) | Other (n=12) |
| mT5-Base | 0.685 | 0.662 | 0.568 | 0.750 | 0.611 | 0.602 | 0.486 | 0.723 |
| Qwen2.5-1.5B | 0.743 | 0.725 | 0.715 | 0.775 | 0.721 | 0.700 | 0.687 | 0.711 |
| Qwen2.5-1.5B-Instruct | 0.740 | 0.722 | 0.715 | 0.766 | 0.722 | 0.698 | 0.687 | 0.704 |
| Qwen2.5-7B | 0.821 | 0.809 | 0.808 | 0.829 | 0.795 | 0.786 | 0.783 | 0.814 |
| Qwen2.5-7B-Instruct | 0.815 | 0.803 | 0.803 | 0.821 | 0.795 | 0.785 | 0.782 | 0.813 |
| Goldfish German | 0.720 | 0.747 | 0.701 | 0.769 | 0.758 | 0.772 | 0.742 | 0.769 |

Table 26: Breakdown of lemma classification accuracy by Part of Speech (POS) for each model, using Linear Regression and Multi-Layer Perceptron (MLP) classifiers (German). Lemmas are grouped by their POS tags (*e.g.*, Noun, Verb, Adjective). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Base (n=688) | 3rd person (n=776) | Positive (n=1,833) | Past (n=857) | Plural (n=1,457) | Singular (n=5,169) |
|---|---|---|---|---|---|---|
| mT5-Base | 0.934 | 0.912 | 0.879 | 0.908 | 0.954 | 0.970 |
| Qwen2.5-1.5B | 0.933 | 0.858 | 0.896 | 0.903 | 0.958 | 0.967 |
| Qwen2.5-1.5B-Instruct | 0.930 | 0.852 | 0.893 | 0.898 | 0.958 | 0.966 |
| Qwen2.5-7B | 0.955 | 0.918 | 0.918 | 0.931 | 0.965 | 0.975 |
| Qwen2.5-7B-Instruct | 0.951 | 0.913 | 0.915 | 0.928 | 0.964 | 0.974 |
| Goldfish French | 0.942 | 0.955 | 0.937 | 0.930 | 0.968 | 0.976 |

Table 27: Breakdown of inflection classification accuracy for each model by inflection type using Linear Regression classifiers (French). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Base (n=688) | 3rd person (n=776) | Positive (n=1,833) | Past (n=857) | Plural (n=1,457) | Singular (n=5,169) |
|---|---|---|---|---|---|---|
| mT5-Base | 0.957 | 0.937 | 0.911 | 0.935 | 0.957 | 0.977 |
| Qwen2.5-1.5B | 0.954 | 0.905 | 0.914 | 0.925 | 0.965 | 0.968 |
| Qwen2.5-1.5B-Instruct | 0.954 | 0.902 | 0.911 | 0.924 | 0.965 | 0.968 |
| Qwen2.5-7B | 0.966 | 0.936 | 0.930 | 0.937 | 0.970 | 0.976 |
| Qwen2.5-7B-Instruct | 0.962 | 0.931 | 0.926 | 0.934 | 0.970 | 0.975 |
| Goldfish French | 0.974 | 0.967 | 0.945 | 0.942 | 0.973 | 0.979 |

Table 28: Breakdown of inflection classification accuracy for each model by inflection type using Multi-Layer Perceptron (MLP) classifiers (French). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Linear Regression | | | | MLP | | | |
|---|---|---|---|---|---|---|---|---|
| | Noun (n=1,496) | Verb (n=406) | Adjective (n=358) | Other (n=15) | Noun (n=1,496) | Verb (n=406) | Adjective (n=358) | Other (n=15) |
| mT5-Base | 0.708 | 0.577 | 0.605 | 0.799 | 0.755 | 0.560 | 0.636 | 0.820 |
| Qwen2.5-1.5B | 0.754 | 0.725 | 0.673 | 0.824 | 0.807 | 0.765 | 0.751 | 0.853 |
| Qwen2.5-1.5B-Instruct | 0.750 | 0.718 | 0.671 | 0.820 | 0.824 | 0.776 | 0.768 | 0.869 |
| Qwen2.5-7B | 0.840 | 0.814 | 0.764 | 0.869 | 0.856 | 0.825 | 0.794 | 0.884 |
| Qwen2.5-7B-Instruct | 0.833 | 0.805 | 0.758 | 0.860 | 0.851 | 0.818 | 0.792 | 0.883 |
| Goldfish French | 0.749 | 0.758 | 0.661 | 0.811 | 0.894 | 0.869 | 0.813 | 0.888 |

Table 29: Breakdown of lemma classification accuracy by Part of Speech (POS) for each model, using Linear Regression and Multi-Layer Perceptron (MLP) classifiers (French). Lemmas are grouped by their POS tags (*e.g.*, Noun, Verb, Adjective). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Base (n=690) | 3rd person (n=456) | Positive (n=1,192) | Past (n=455) | Plural (n=1,333) | Superlative (n=3) | Singular (n=3,316) | Comparative (n=23) |
|---|---|---|---|---|---|---|---|---|
| mT5-Base | 0.930 | 0.978 | 0.975 | 0.957 | 0.877 | 0.000 | 0.977 | 0.799 |
| Qwen2.5-1.5B | 0.925 | 0.946 | 0.974 | 0.938 | 0.923 | 0.015 | 0.966 | 0.835 |
| Qwen2.5-1.5B-Instruct | 0.924 | 0.943 | 0.974 | 0.934 | 0.921 | 0.015 | 0.966 | 0.817 |
| Qwen2.5-7B | 0.949 | 0.966 | 0.979 | 0.958 | 0.948 | 0.094 | 0.977 | 0.872 |
| Qwen2.5-7B-Instruct | 0.951 | 0.974 | 0.980 | 0.970 | 0.948 | 0.080 | 0.980 | 0.918 |
| Goldfish Russian | 0.940 | 0.950 | 0.976 | 0.931 | 0.921 | 0.000 | 0.976 | 0.867 |

Table 30: Breakdown of inflection classification accuracy for each model by inflection type using Linear Regression classifiers (Russian). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Base (n=690) | 3rd person (n=456) | Positive (n=1,192) | Past (n=455) | Plural (n=1,333) | Superlative (n=3) | Singular (n=3,316) | Comparative (n=23) |
|---|---|---|---|---|---|---|---|---|
| mT5-Base | 0.959 | 0.978 | 0.969 | 0.966 | 0.904 | 0.000 | 0.978 | 0.849 |
| Qwen2.5-1.5B | 0.952 | 0.955 | 0.972 | 0.948 | 0.933 | 0.089 | 0.970 | 0.899 |
| Qwen2.5-1.5B-Instruct | 0.950 | 0.954 | 0.973 | 0.947 | 0.933 | 0.089 | 0.969 | 0.911 |
| Qwen2.5-7B | 0.963 | 0.964 | 0.978 | 0.960 | 0.951 | 0.246 | 0.979 | 0.910 |
| Qwen2.5-7B-Instruct | 0.961 | 0.970 | 0.978 | 0.966 | 0.949 | 0.126 | 0.980 | 0.924 |
| Goldfish Russian | 0.965 | 0.972 | 0.978 | 0.948 | 0.943 | 0.000 | 0.977 | 0.934 |

Table 31: Breakdown of inflection classification accuracy for each model by inflection type using Multi-Layer Perceptron (MLP) classifiers (Russian). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Linear Regression | | | | MLP | | | |
|---|---|---|---|---|---|---|---|---|
| | Noun (n=982) | Verb (n=333) | Adjective (n=275) | Other (n=4) | Noun (n=982) | Verb (n=333) | Adjective (n=275) | Other (n=4) |
| mT5-Base | 0.660 | 0.614 | 0.542 | 0.648 | 0.492 | 0.484 | 0.387 | 0.426 |
| Qwen2.5-1.5B | 0.777 | 0.712 | 0.759 | 0.720 | 0.712 | 0.696 | 0.716 | 0.647 |
| Qwen2.5-1.5B-Instruct | 0.772 | 0.704 | 0.756 | 0.720 | 0.710 | 0.689 | 0.717 | 0.643 |
| Qwen2.5-7B | 0.854 | 0.790 | 0.843 | 0.812 | 0.798 | 0.794 | 0.813 | 0.749 |
| Qwen2.5-7B-Instruct | 0.845 | 0.778 | 0.835 | 0.807 | 0.794 | 0.785 | 0.809 | 0.744 |
| Goldfish Russian | 0.795 | 0.723 | 0.764 | 0.676 | 0.810 | 0.776 | 0.759 | 0.657 |

Table 32: Breakdown of lemma classification accuracy by Part of Speech (POS) for each model, using Linear Regression and Multi-Layer Perceptron (MLP) classifiers (Russian). Lemmas are grouped by their POS tags (*e.g.*, Noun, Verb, Adjective). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Base (n=154) | 3rd person (n=51) | Positive (n=401) | Past (n=168) | Plural (n=33) | Singular (n=632) |
|---|---|---|---|---|---|---|
| mT5-Base | 0.860 | 0.911 | 0.928 | 0.966 | 0.837 | 0.952 |
| Qwen2.5-1.5B | 0.808 | 0.802 | 0.721 | 0.928 | 0.861 | 0.892 |
| Qwen2.5-1.5B-Instruct | 0.809 | 0.817 | 0.720 | 0.941 | 0.878 | 0.899 |
| Qwen2.5-7B | 0.865 | 0.879 | 0.810 | 0.966 | 0.903 | 0.909 |
| Qwen2.5-7B-Instruct | 0.850 | 0.874 | 0.796 | 0.960 | 0.886 | 0.900 |
| Goldfish Turkish | 0.847 | 0.915 | 0.880 | 0.964 | 0.872 | 0.963 |

Table 33: Breakdown of inflection classification accuracy for each model by inflection type using Linear Regression classifiers (Turkish). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Base (n=154) | 3rd person (n=51) | Positive (n=401) | Past (n=168) | Plural (n=33) | Singular (n=632) |
|---|---|---|---|---|---|---|
| mT5-Base | 0.755 | 0.760 | 0.848 | 0.922 | 0.515 | 0.949 |
| Qwen2.5-1.5B | 0.770 | 0.767 | 0.667 | 0.919 | 0.765 | 0.914 |
| Qwen2.5-1.5B-Instruct | 0.762 | 0.757 | 0.662 | 0.917 | 0.766 | 0.913 |
| Qwen2.5-7B | 0.853 | 0.845 | 0.791 | 0.956 | 0.875 | 0.937 |
| Qwen2.5-7B-Instruct | 0.845 | 0.844 | 0.786 | 0.956 | 0.875 | 0.932 |
| Goldfish Turkish | 0.832 | 0.879 | 0.870 | 0.957 | 0.834 | 0.957 |

Table 34: Breakdown of inflection classification accuracy for each model by inflection type using Multi-Layer Perceptron (MLP) classifiers (Turkish). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.

| Model | Linear Regression | | | | MLP | | | |
|---|---|---|---|---|---|---|---|---|
| | Noun (n=221) | Verb (n=53) | Adjective (n=104) | Other (n=13) | Noun (n=221) | Verb (n=53) | Adjective (n=104) | Other (n=13) |
| mT5-Base | 0.866 | 0.823 | 0.921 | 0.955 | 0.215 | 0.421 | 0.374 | 0.637 |
| Qwen2.5-1.5B | 0.834 | 0.805 | 0.866 | 0.877 | 0.307 | 0.439 | 0.449 | 0.693 |
| Qwen2.5-1.5B-Instruct | 0.816 | 0.791 | 0.860 | 0.874 | 0.305 | 0.439 | 0.448 | 0.691 |
| Qwen2.5-7B | 0.871 | 0.850 | 0.900 | 0.904 | 0.595 | 0.625 | 0.695 | 0.809 |
| Qwen2.5-7B-Instruct | 0.850 | 0.823 | 0.883 | 0.885 | 0.579 | 0.613 | 0.678 | 0.800 |
| Goldfish Turkish | 0.929 | 0.904 | 0.940 | 0.969 | 0.386 | 0.550 | 0.477 | 0.808 |

Table 35: Breakdown of lemma classification accuracy by Part of Speech (POS) for each model, using Linear Regression and Multi-Layer Perceptron (MLP) classifiers (Turkish). Lemmas are grouped by their POS tags (*e.g.*, Noun, Verb, Adjective). Accuracies are calculated over all examples for a given group across all layers. Counts (n) are derived from a single representative layer for each group. All accuracy values are on a 0–1 scale.